IBM SPSS Modeler 18.2 — węzły modelowania

Uwaga

Przed skorzystaniem z niniejszych informacji oraz produktu, którego one dotyczą, należy zapoznać się z informacjami zamieszczonymi w sekcji “Uwagi” na stronie 385.

Informacje o produkcie

Niniejsze wydanie publikacji dotyczy wersji 18, wydania 2, modyfikacji 0 produktu IBM SPSS Modeler oraz wszystkich następnych wydań i modyfikacji do czasu, aż w kolejnym wydaniu publikacji zostanie zawarta informacja o stosownej zmianie.

Spis treści

Przedmowa

IBM® SPSS Modeler to oferowane przez IBM Corp. zaawansowane środowisko eksploracji danych. SPSS Modeler pomaga przedsiębiorstwom i instytucjom w rozwijaniu relacji z klientami i obywatelami w oparciu o pogłębioną interpretację dostępnych danych. Organizacje korzystają z wiedzy uzyskanej dzięki programowi SPSS Modeler w bardzo szerokim spektrum zastosowań, m.in. do zatrzymywania najbardziej wartościowych klientów, określania możliwości sprzedaży wiązanej, przyciągania nowych klientów, wykrywania oszustw, ograniczania ryzyka i podnoszenia jakości usług publicznych.

Interfejs graficzny produktu SPSS Modeler zachęca użytkowników, aby wykorzystywali specjalistyczną wiedzę, dzięki której możliwe będzie opracowanie bardziej wydajnych modeli predykcyjnych i skrócenie czasu potrzebnego do uzyskania rozwiązania. SPSS Modeler oferuje wiele technik modelowania, takich jak predykcja, klasyfikacja, segmentacja i algorytmy do wykrywania związków. Po utworzeniu modeli program IBM SPSS Modeler Solution Publisher umożliwia udostępnienie ich osobom podejmującym decyzje w całym przedsiębiorstwie lub zapisanie w bazie danych.

Informacje o programie IBM Business Analytics

Oprogramowanie IBM Business Analytics dostarcza kompletne, spójne i dokładne informacje, na których mogą polegać osoby decyzyjne chcące polepszyć wyniki biznesowe. Wszechstronne portfolio obejmujące moduły: analiza biznesowa, analiza prognostyczna, zarządzanie wynikami i strategiami finansowymi oraz aplikacje analityczne, zapewnia jasny, natychmiastowy i pozwalający na podjęcie działań wgląd w bieżące wyniki oraz daje możliwość przewidywania przyszłych wyników. W połączeniu z licznymi rozwiązaniami branżowymi, sprawdzonymi praktykami i profesjonalnymi usługami, organizacje o różnych rozmiarach mogą wspomagać najwyższą produktywność, w sposób pewny zautomatyzować decyzje i uzyskać lepsze wyniki.

Oprogramowanie IBM SPSS Predictive Analytics będące częścią tego portfolio wspomaga organizacje w zakresie przewidywania przyszłych zdarzeń oraz proaktywnie wpływać na na ten wgląd z korzyścią dla wyników finansowych. Klienci komercyjni, rządowi i uczelnie na całym świecie polegają na technologii IBM SPSS, zapewniającej przewagę konkurencyjną, dzięki której przyciągają, zatrzymują i pozyskują nowych klientów, walcząc z nieuczciwością i ograniczając ryzyko. Wdrażając oprogramowanie IBM SPSS do swojej codziennej działalności, organizacje stają się przewidującymi przedsiębiorstwami, zdolnymi do zarządzania i automatyzacji decyzji w celu realizacji celów biznesowych i osiągnięcia mierzalnej przewagi konkurencyjnej. W celu uzyskania dalszych informacji lub skontaktowania się z przedstawicielem, proszę wejść na stronę http://www.ibm.com/spss.

Wsparcie techniczne

Wsparcie techniczne jest dostępne w celu zapewnienia klientom obsługi technicznej. Klienci mogą się kontaktować z działem Wsparcia technicznego w celu uzyskania pomocy dotyczącej korzystania z produktów lub pomocy w instalacji IBM Corp. dla jednego z obsługiwanych środowisk sprzętowych. Aby skontaktować się z działem Wsparcia technicznego, wejdź na stronę internetową IBM Corp. pod adresem http://www.ibm.com/support. W przypadku prośby o pomoc, należy przygotować swoje dane identyfikacyjne, dane swojej organizacji, a także dane dotyczące usług wsparcia.

Rozdział 1. O programie IBM SPSS Modeler

IBM SPSS Modeler to zestaw narzędzi do eksploracji danych. Produkt umożliwia szybkie opracowywanie modeli predykcyjnych przy wykorzystaniu wiedzy specjalistycznej i stosowanie tych modeli w procesach biznesowych, jako wsparcia przy podejmowaniu decyzji. Rozwiązania zawarte w oprogramowaniu IBM SPSS Modeler zapewniają możliwość wykorzystywania branżowego modelu CRISP-DM i pozwalają na obsługę całego procesu eksploracji danych: od pozyskiwania danych do uzyskiwania lepszych wyników biznesowych.

Oprogramowanie IBM SPSS Modeler umożliwia korzystanie z wielu metod modelowania opartych na sztucznej inteligencji, uczeniu maszynowym i statystykach. Metody dostępne na palecie Modelowanie pozwalają na ekstrahowanie nowych informacji z danych i tworzenie modeli predykcyjnych. Każda metoda ma określone mocne strony i jest dostosowana do rozwiązywania określonych problemów.

Oprogramowanie SPSS Modeler można zakupić jako produkt samodzielny lub jako program kliencki używany wraz z oprogramowaniem SPSS Modeler Server. Dostępnych jest wiele opcji dodatkowych, które przedstawiono w kolejnych rozdziałach. Aby uzyskać więcej informacji, patrz https://www.ibm.com/analytics/us/en/technology/spss/.

Produkty IBM SPSS Modeler

Rodzina produktów IBM SPSS Modeler i towarzyszącego im oprogramowania składa się z elementów przedstawionych poniżej.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console (towarzyszącyIBM SPSS Deployment Manager)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server adapters for IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

Oprogramowanie SPSS Modeler to w pełni funkcjonalna wersja produktu instalowana i uruchamiana na komputerze osobistym. Oprogramowanie SPSS Modeler można uruchomić lokalnie jako produkt samodzielny lub korzystać z niego w trybie rozproszonym wraz z serwerem IBM SPSS Modeler Server. Tego typu rozwiązanie zapewnia zwiększenie wydajności obsługi dużych zbiorów danych.

Dzięki oprogramowaniu SPSS Modeler można szybko tworzyć dokładne modele predykcyjne, stosując intuicyjne metody niewymagające umiejętności programowania. Unikatowy interfejs graficzny pozwala na wizualizowanie procedur eksploracji danych. Zaawansowane metody opracowywania analiz dostępne w programie umożliwiają określanie wcześniej niezauważalnych wzorców i trendów zawartych w danych. Użytkownik może modelować wyniki i poznawać czynniki wpływające na ich wartości. W ten sposób można wykorzystywać nowe szanse biznesowe i obniżać ryzyko.

Dostępne są dwie edycje oprogramowania SPSS Modeler: SPSS Modeler Professional oraz SPSS Modeler Premium. Więcej informacji można znaleźć w temacie “Wydania programu IBM SPSS Modeler” na stronie 2.

IBM SPSS Modeler Server

Oprogramowanie SPSS Modeler działa w oparciu o architekturę klient-serwer, w której żądania wymagające zaangażowania dużych zasobów kierowane są do zaawansowanego oprogramowania serwerowego. Takie rozwiązanie umożliwia bardziej wydajną obsługę dużych zbiorów danych.

SPSS Modeler Server to produkt wymagający dodatkowej licencji, działający stale na serwerze w trybie analizy rozproszonej. Współpracuje on z co najmniej jedną instalacją oprogramowania IBM SPSS Modeler. W ten sposób oprogramowanie SPSS Modeler Server poprawia wydajność podczas obsługi dużych zbiorów danych, ponieważ operacje wymagające dużej mocy obliczeniowej można wykonywać na serwerze bez potrzeby pobierania danych na komputer kliencki. Oprogramowanie IBM SPSS Modeler Server optymalizuje również obsługę SQL i funkcje modelowania wewnątrz bazy danych, co dodatkowo zwiększa wydajność działania i sprzyja automatyzacji pracy.

IBM SPSS Modeler Administration Console

Oprogramowanie Modeler Administration Console to interfejs graficzny służący do obsługi wielu opcji konfiguracji SPSS Modeler Server, które można dostosować również za pomocą pliku opcji. W ramach IBM SPSS Deployment Manager dostępna jest konsola pozwalająca na monitorowanie i konfigurowanie instalacji SPSS Modeler Server. Interfejs jest dostępny bez dodatkowych opłat dla aktualnych użytkowników SPSS Modeler Server customers. Aplikację można zainstalować tylko na komputerach z systemem Windows, jednak administrować można serwerem zainstalowanym na dowolnej obsługiwanej platformie.

IBM SPSS Modeler Batch

Eksploracja danych jest zazwyczaj procesem interaktywnym, jednak oprogramowanie SPSS Modeler można też uruchomić z poziomu wiersza komend i zrezygnować z używania graficznego interfejsu użytkownika. Niekiedy użytkownik wykonuje długotrwałe lub powtarzalne zadania, które mogą być realizowane bez nadzoru. Oprogramowanie SPSS Modeler Batch to specjalna wersja produktu pozwalająca na wykonywanie wszystkich funkcji analitycznych SPSS Modeler bez potrzeby używania standardowego interfejsu użytkownika. Oprogramowanie SPSS Modeler Server jest wymagane do korzystania z aplikacji SPSS Modeler Batch.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher umożliwia tworzenie spakowanych wersji strumieni programu SPSS Modeler, które można uruchamiać za pomocą zewnętrznych środowisk wykonawczych lub osadzać w aplikacji zewnętrznej. W ten sposób można publikować i wdrażać pełne strumienie SPSS Modeler w celu używania ich w środowiskach, w których nie zainstalowano programu SPSS Modeler. SPSS Modeler Solution Publisher jest dystrybuowany jako część produktu IBM SPSS Collaboration and Deployment Services - Scoring, który do działania wymaga oddzielnej licencji. Wraz z licencją użytkownik otrzymuje oprogramowanie SPSS Modeler Solution Publisher Runtime umożliwiające uruchamianie opublikowanych strumieni.

Więcej informacji na temat programu SPSS Modeler Solution Publisher znajduje się w dokumentacji produktu IBM SPSS Collaboration and Deployment Services. W Centrum wiedzy IBM SPSS Collaboration and Deployment Services dostępne są sekcje "IBM SPSS Modeler Solution Publisher" oraz "IBM SPSS Analytics Toolkit".

IBM SPSS Modeler Server Adapters for IBM SPSS Collaboration and Deployment Services

Dostępnych jest wiele adapterów dla IBM SPSS Collaboration and Deployment Services, które umożliwiają współpracę programów SPSS Modeler i SPSS Modeler Server z repozytorium IBM SPSS Collaboration and Deployment Services. Dzięki temu strumień SPSS Modeler wdrożony w repozytorium można udostępnić wielu użytkownikom lub uzyskać do niego dostęp z poziomu uproszczonej aplikacji klienckiej IBM SPSS Modeler Advantage. Adapter należy zainstalować na systemie hostującym repozytorium.

Wydania programu IBM SPSS Modeler

Dostępne są następujące wydania oprogramowania SPSS Modeler.

SPSS Modeler Professional

Oprogramowanie SPSS Modeler Professional zapewnia wszystkie narzędzia wymagane do obsługi większości typów danych ustrukturyzowanych, takich jak np. zachowania i interakcje śledzone w systemach CRM, dane demograficzne, zachowania zakupowe i dane sprzedażowe.

SPSS Modeler Premium

Oprogramowanie SPSS Modeler Premium wymaga oddzielnej licencji. Dzięki temu rozwiązaniu oprogramowanie SPSS Modeler Professional może obsługiwać wyspecjalizowane dane oraz nieustrukturyzowane dane tekstowe. SPSS Modeler Premium zawiera IBM SPSS Modeler Text Analytics:

Program **IBM SPSS Modeler Text Analytics** korzysta z zaawansowanych rozwiązań lingwistycznych oraz przetwarzania języka naturalnego w celu szybkiego przetwarzania różnego rodzaju nieustrukturyzowanych danych tekstowych, wyodrębniania i porządkowania kluczowych pojęć oraz grupowania tych pojęć w kategorie. Wyodrębnione pojęcia i kategorie można łączyć z istniejącymi danymi ustrukturyzowanymi, takimi jak dane demograficzne, a następnie stosować w celu modelowania, korzystając z produktu IBM SPSS Modeler i zawartego w nim pełnego pakietu narzędzi do eksploracji danych, aby w rezultacie takiego połączenia podejmować lepsze decyzje przy zmniejszonej ilości zakłóceń.

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription oferuje te same funkcje analiz predykcyjnych, co tradycyjny klient IBM SPSS Modeler. Użytkownicy edycji Subscription mogą regularnie pobierać aktualizacje produktu.

Dokumentacja

Dokumentacja jest dostępna w programie SPSS Modeler z poziomu menu Pomoc. Spowoduje to otwarcie Centrum Wiedzy, które jest powszechnie dostępne poza produktem.

Pełna dokumentacja dla każdego produktu (obejmująca instrukcje instalacji) jest dostępna również w formacie PDF, w osobnym, skompresowanym folderze, w ramach materiałów dotyczących produktu do pobrania. Dokumenty dotyczące instalacji można także pobrać z Internetu pod adresem http://www.ibm.com/support/docview.wss?uid=swg27046871.

Dokumentacja SPSS Modeler Professional

Pakiet dokumentacji produktu SPSS Modeler Professional (bez instrukcji instalacyjnych) zawiera następujące publikacje.

- **IBM SPSS Modeler — podręcznik użytkownika.** Ogólne wprowadzenie do obsługi oprogramowania SPSS Modeler, w tym opisy procedur tworzenia strumieni danych, obsługi braków danych, tworzenia wyrażeń CLEM pracy z projektami i raportami, a także przygotowywania strumieni do wdrożenia w IBM SPSS Collaboration and Deployment Services lub IBM SPSS Modeler Advantage.
- **IBM SPSS Modeler — węzły źródłowe, procesowe i wyników.** Opisy wszystkich węzłów używanych do odczytywania, przetwarzania i tworzenia wynikowych postaci danych w różnych formatach. Czyli wszystkich węzłów poza węzłami modelowania.
- **IBM SPSS Modeler — węzły modelowania.** Opisy wszystkich węzłów używanych do tworzenia modeli eksploracji danych. Oprogramowanie IBM SPSS Modeler umożliwia korzystanie z wielu metod modelowania opartych na sztucznej inteligencji, uczeniu maszynowym i statystykach.
- **IBM SPSS Modeler — podręcznik zastosowań.** Przykłady zawarte w niniejszym przewodniku stanowią skrócone informacje związane z konkretnymi metodami i technikami modelowania. Wersja elektroniczna tego podręcznika jest również dostępna z poziomu menu Pomoc. Więcej informacji można znaleźć w temacie “Przykłady zastosowań” na stronie 4.
- **IBM SPSS Modeler — podręcznik tworzenia skryptów w języku Python i automatyzacji.** Informacje na temat automatyzacji działania systemu za pomocą skryptów Python wraz z właściwościami służącymi do obsługi węzłów i strumieni.
- **IBM SPSS Modeler — podręcznik wdrażania.** Informacje na temat uruchamiania strumieni IBM SPSS Modeler przedstawione w postaci krokowych operacji wykonywanych podczas przetwarzania zadań w oprogramowaniu IBM SPSS Deployment Manager.
- **IBM SPSS Modeler CLEF Developer's Guide.** Z oprogramowaniem CLEF można zintegrować inne programy pozwalające na przetwarzanie danych lub obsługę algorytmów modelujących w postaci węzłów w IBM SPSS Modeler.

- **IBM SPSS Modeler — podręcznik eksploracji w bazie danych.** Informacje na temat wydajnego wykorzystywania bazy danych w celu zwiększenia wydajności i zakresu funkcji analitycznych za pomocą algorytmów innych firm.
- **IBM SPSS Modeler Server — podręcznik administracji i wydajności.** Informacje na temat konfiguracji i funkcji administracyjnych w oprogramowaniu IBM SPSS Modeler Server.
- **IBM SPSS Deployment Manager Podręcznik użytkownika.** Informacje dotyczące obsługi interfejsu Administration Console w oprogramowaniu Deployment Manager do monitorowania i konfigurowania produktu IBM SPSS Modeler Server.
- **IBM SPSS Modeler — podręcznik CRISP-DM.** Szczegółowy podręcznik metodologii CRISP-DM w kontekście eksploracji danych za pomocą oprogramowania SPSS Modeler.
- **IBM SPSS Modeler Batch — podręcznik użytkownika.** Pełny podręcznik obsługi oprogramowania IBM SPSS Modeler w trybie wsadowym obejmujący szczegółowe informacje na temat pracy w trybie wsadowym i korzystania z argumentów z poziomu wiersza komend. Ten podręcznik jest dostępny tylko w formacie PDF.

Dokumentacja SPSS Modeler Premium

Pakiet dokumentacji produktu SPSS Modeler Premium (bez instrukcji instalacyjnych) zawiera następujące publikacje.

- **SPSS Modeler Text Analytics — podręcznik użytkownika.** Informacje na temat używania analiz tekstu za pomocą oprogramowania SPSS Modeler, obejmują procedury dotyczące węzłów eksploracji tekstu, interaktywnego pulpitu roboczego, szablonów oraz innych zasobów.

Przykłady zastosowań

Podczas gdy narzędzia do eksploracji danych w programie SPSS Modeler mogą pomóc w rozwiązaniu szeregu problemów biznesowych i organizacyjnych, przykłady aplikacji udostępniają krótkie, ukierunkowane wprowadzenia do konkretnych metod i technik modelowania. Używane tutaj zestawy danych są znacznie mniejsze niż ogromne składnice danych zarządzane przez programy do eksploracji danych, lecz używane koncepcje i metody są skalowalne odpowiednio do potrzeb rzeczywistych aplikacji.

Dostęp do przykładów można uzyskać, klikając opcję **Przykłady aplikacji** w menu Pomoc programu SPSS Modeler.

Pliki danych i przykładowe strumienie są instalowane w folderze Dema, w katalogu instalacyjnym produktu. Aby uzyskać więcej informacji, patrz “Folder Demos”.

Przykłady modelowania w bazach danych. Przykłady zamieszczono w publikacji *IBM SPSS Modeler — podręcznik eksploracji w bazie danych*.

Przykłady skryptów. Przykłady zamieszczono w publikacji *IBM SPSS Modeler — podręcznik tworzenia skryptów w języku Python i automatyzacji*.

Folder Demos

Pliki danych i przykładowe strumienie używane z przykładami do aplikacji są instalowane w folderze Demos wewnątrz katalogu instalacyjnego produktu (na przykład: C:\Program Files\IBM\SPSS\Modeler\<version>\Demos). Dostęp do tego folderu można także uzyskać z grupy programów IBM SPSS Modeler w menu Start systemu Windows lub klikając opcję Demos na liście ostatnich katalogów w oknie dialogowym **Plik > Otwórz strumień**.

Monitorowanie wykorzystania licencji

Podczas pracy z produktem SPSS Modeler wykorzystanie licencji jest monitorowanie i regularnie rejestrowane. Metryka wykorzystania licencji nosi nazwę *AUTHORIZED_USER* (użytkownik autoryzowany) lub *CONCURRENT_USER* (użytkownik pracujący jednocześnie), a typ rejestrowanej metryki zależy od typu licencji na produktSPSS Modeler, którą posiada użytkownik.

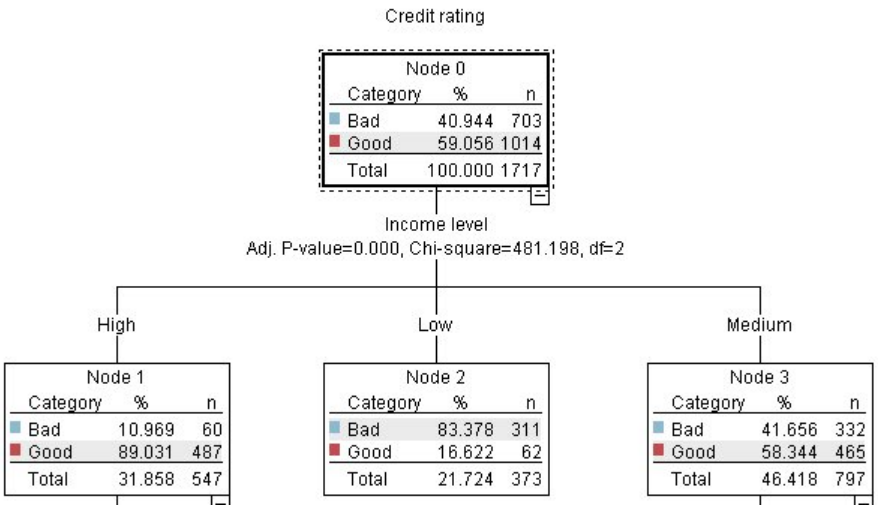
Generowane pliki dzienników mogą być przetwarzane przez program IBM License Metric Tool, z którego uzyskać można raporty o wykorzystaniu licencji.

Pliki dzienników wykorzystania licencji są tworzone w tym samym katalogu, w którym zapisywane są dzienniki klienta SPSS Modeler (domyślnie %ALLUSERSPROFILE%/IBM/SPSS/Modeler/<wersja>/log).

Rozdział 2. Wstęp do modelowania

Model to zestaw reguł, formuł lub równań, które mogą być używane do przewidywania danych wynikowych w oparciu o zestaw zmiennych wejściowych. Na przykład instytucja finansowa może używać modelu do przewidywania, czy osoby ubiegające się o kredyt są obarczone wysokim czy niskim poziomem ryzyka, w oparciu o uzyskane informacje na temat poprzednich wnioskujących.

Zdolność do przewidywania danych wynikowych jest podstawowym celem analizy predykcyjnej, a zrozumienie procesu modelowania ma kluczowe znaczenie dla korzystania z produktu IBM SPSS Modeler.



Rysunek 1. Prosty model drzewa decyzyjnego

W tym przykładzie zastosowano model **drzewa decyzyjnego**, który klasyfikuje rekordy (i przewiduje odpowiedź) używając szeregu reguł decyzyjnych, na przykład:

```
IF income = Medium
AND cards <5
THEN -> 'Good'
```

Ponieważ w tym przykładzie użyto modelu CHAID (automatyczna detekcja interakcji chi-kwadrat), stanowi on ogólne wprowadzenie, a większość koncepcji ma zasadniczo zastosowanie do innych typów modelowania w programie IBM SPSS Modeler.

Aby zrozumieć dowolny model, najpierw należy zapoznać się z danymi, które są w nim uwzględniane. Dane w tym przykładzie obejmują informacje na temat klientów banku. Używane są następujące zmienne:

Nazwa zmiennej	**Opis**
Credit_rating	Ocena kredytowa: 0=pozytywna, 1=negatywna, 9=brak wartości
Age	Wiek w latach
Income	Poziom przychodu: 1=niski, 2=średni, 3=wysoki
Credit_cards	Liczba posiadanych kart kredytowych: 1=mniej niż pięć, 2=pięć lub więcej
Education	Poziom wykształcenia: 1=wyższe, 2=średnie
Car_loans	Liczba zaciągniętych kredytów samochodowych: 1=brak lub jeden, 2=więcej niż dwa

Bank opracowuje bazę danych zawierającą historyczne informacje o klientach, którym bank udzielił kredytu, z uwzględnieniem faktu, czy kredyty te spłacili (Ocena kredytowa = pozytywna) czy nie (Ocena kredytowa = negatywna). Korzystając z istniejących danych, bank zamierza utworzyć model, który umożliwi przewidywanie prawdopodobieństwa, że przyszli wnioskujący o kredyt nie będą spłacać zobowiązań.

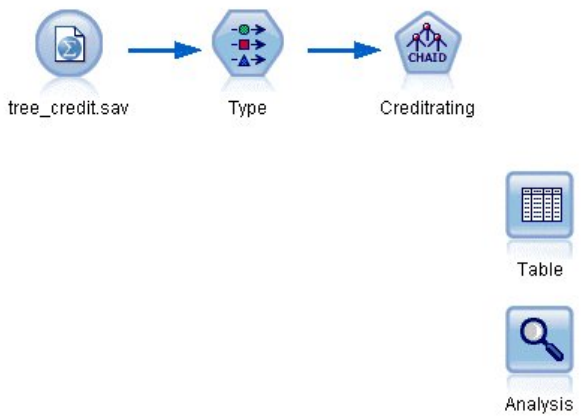
Używając modelu drzewa decyzyjnego, można przeprowadzić analizę cech dwóch grup klientów i przewidzieć prawdopodobieństwo niespłacania kredytu.

W tym przykładzie zastosowano strumień o nazwie *modelingintro.str*, który jest dostępny w folderze *Demos*, podfolder *streams*. Plik danych to *tree_credit.sav*. Więcej informacji można znaleźć w temacie “Folder Demos” na stronie 4.

Przyjrzyjmy się strumieniowi.

1. Wybierz następujące opcje z menu głównego:
   Plik > Otwórz strumień
2. Kliknij złotą ikonę modelu użytkowego na pasku narzędzi w oknie dialogowym Otwórz i wybierz folder Demos.
3. Kliknij dwukrotnie folder *streams*.
4. Kliknij dwukrotnie plik o nazwie *modelingintro.str*.

Tworzenie strumienia



Rysunek 2. Strumień modelowania

Aby utworzyć strumień, który utworzy model, potrzebne są co najmniej trzy elementy:

- Węzeł źródłowy, który odczytuje dane z jakiegoś zewnętrznego źródła, w tym przypadku jest to plik danych IBM SPSS Statistics.
- Węzeł źródła lub typu, który określa właściwości zmiennych, takie jak poziom pomiaru (typ danych, jakie zawiera zmienna) oraz role poszczególnych zmiennych w modelowaniu, takie jak zmienne przewidywane lub wejściowe.
- Węzeł modelowania, który generuje model użytkowy w czasie wykonywania strumienia.

W tym przykładzie korzystamy z węzła modelowania CHAID. CHAID lub automatyczna detekcja interakcji chi-kwadrat to metoda klasyfikacji, która umożliwia tworzenie drzew decyzyjnych na podstawie określonego typu statystyk znanych jako statystyki chi-kwadrat w celu określenia najlepszych miejsc podziału w drzewie decyzyjnym.

Jeśli w węźle źródłowym określone są poziomy pomiaru, można wyeliminować osobny węzeł typu. Funkcjonalnie wynik będzie taki sam.

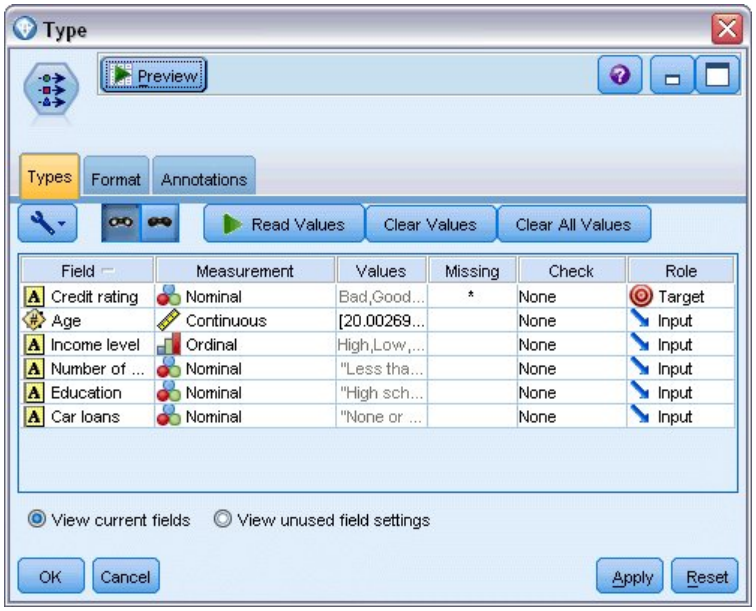
Ten strumień zawiera również węzły Tabela i Analiza, które będą używane do wyświetlania wyników oceniania po utworzeniu modelu użytkowego i dodaniu go do strumienia.

Węzeł źródłowy Statistics odczytuje dane w formacie IBM SPSS Statistics z pliku danych *tree_credit.sav*, który jest zainstalowany w folderze *Demos*. (Specjalna zmienna o nazwie *$CLEO_DEMOS* stanowi odniesienie do tego folderu w bieżącej instalacji produktu IBM SPSS Modeler. Dzięki temu ścieżka będzie poprawna niezależnie od folderu lub wersji bieżącej instalacji).

Rysunek 3. Odczyt danych z użyciem węzła źródłowego Plik Statistics

Węzeł typu określa **poziom pomiaru** dla każdej zmiennej. Poziom pomiaru to kategoria wskazująca typ danych w zmiennej. Nasz plik danych źródłowych korzysta z trzech różnych poziomów pomiaru.

Zmienna **Ilościowa** (np. zmienna *Age*) zawiera ilościowe wartości liczbowe, a zmienna **Nominalna** (np. zmienna *Credit rating*) zawiera co najmniej dwie wartości wyróżniające się, np. *Bad*, *Good* lub *No credit history*. Zmienna **Porządkowa** (np. zmienna *Income level*) opisuje dane z wieloma wartościami wyróżniającymi się, które mają dziedziczną kolejność — w tym przypadku *Low*, *Medium* i *High*.



Rysunek 4. Ustawienie zmiennych przewidywanych i wejściowych w węźle Typ

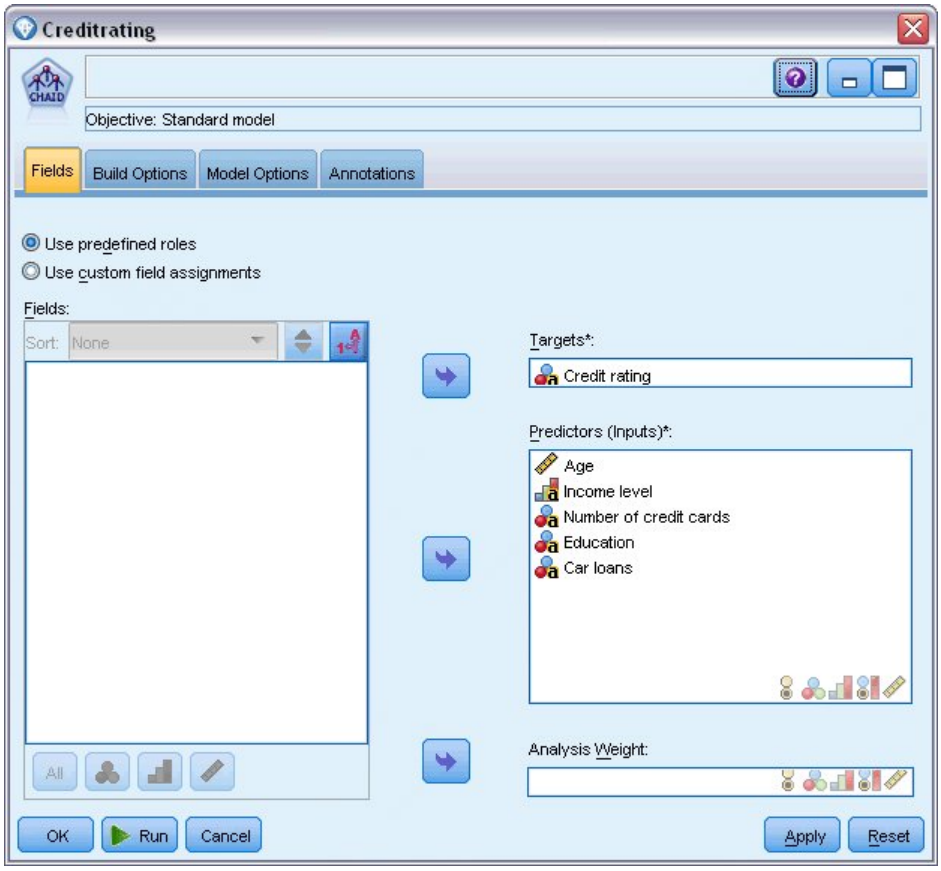
Dla każdej zmiennej węzeł Typ również określa **rolę** wskazującą udział poszczególnych zmiennych w modelowaniu. Rola jest ustawiana na wartość *Przewidywana* dla zmiennej *Credit rating*, która wskazuje, czy dany klient nie spłaca kredytu. Jest to **zmienna przewidywana** lub zmienna, dla której zamierzamy przewidzieć wartość.

Dla pozostałych zmiennych rola jest ustawiona jako *Wejście*. Zmienne wejściowe są niekiedy znane jako **predyktory** lub zmienne, których wartości są używane przez algorytm modelowania do przewidywania wartości zmiennej przewidywanej.

Węzeł modelowania CHAID generuje model.

Na karcie Zmienne w węźle modelowania zaznaczona jest opcja **Użyj wstępnie zdefiniowanych ról**, co oznacza, że użyte zostaną zmienne przewidywane i wejściowe określone w węźle Typ. W tym miejscu można zmienić role zmiennych, jednak na potrzeby przykładu pozostawimy je bez zmian.

1. Kliknij zakładkę Opcje budowania.



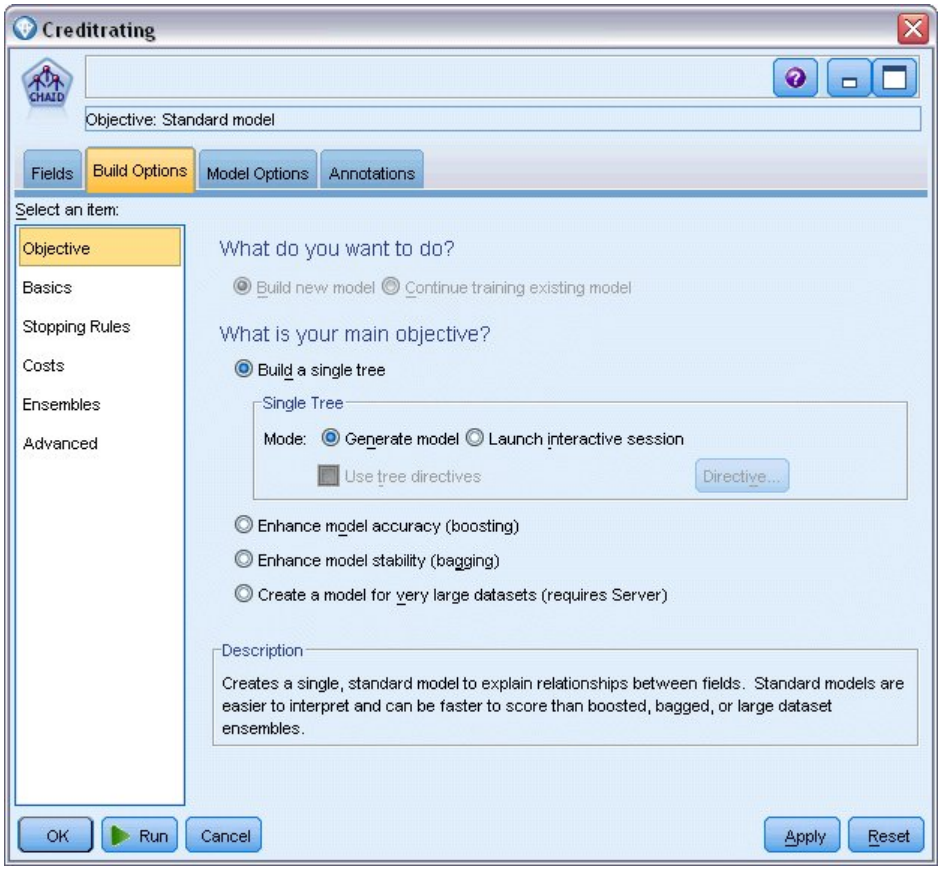
Rysunek 5. Węzeł modelowania CHAID, zakładka Zmienne

Dostępnych jest tutaj kilka opcji, które umożliwiają określenie rodzaju modelu, jaki ma zostać utworzony.

Zamierzamy utworzyć nowy model, dlatego użyjemy opcji domyślnej **Zbuduj nowy model**.

Ma to być pojedynczy, standardowy model drzewa decyzyjnego bez rozszerzeń, dlatego pozostawiamy domyślną opcję celu **Zbudować pojedyncze drzewo**.

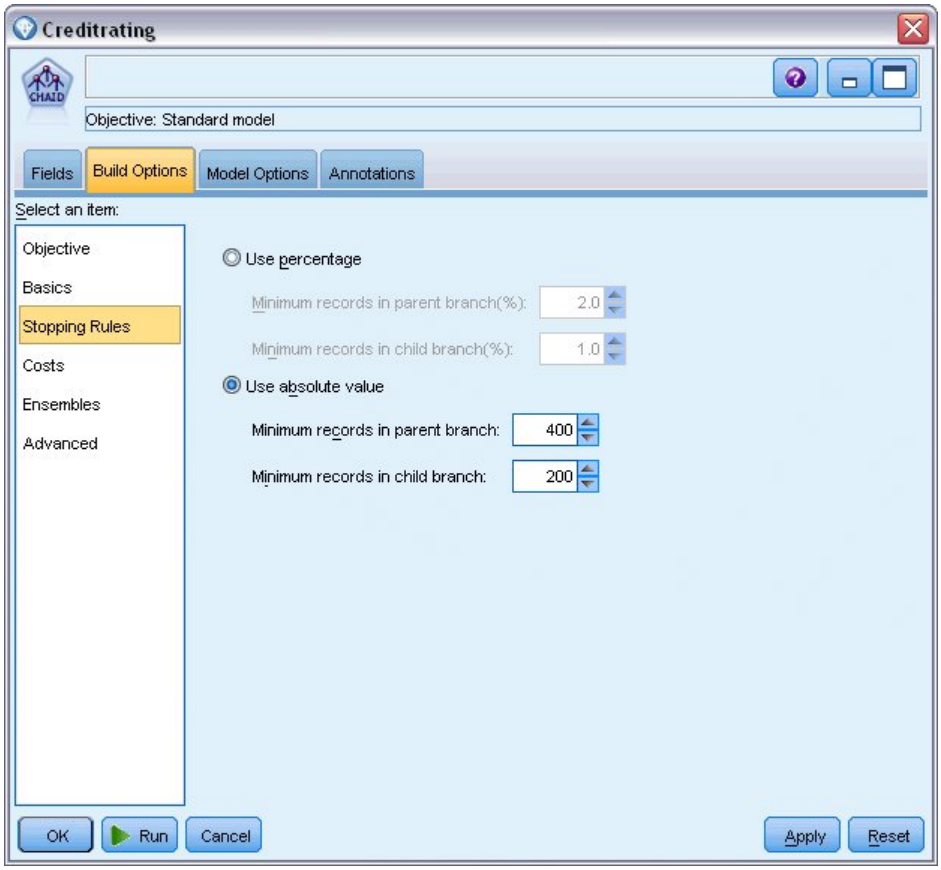
Teraz można opcjonalnie uruchomić interaktywną sesję modelowania, która pozwoli na dostosowanie modelu, jednak w tym przykładzie zostanie po prostu wygenerowany model z zastosowaniem domyślnego ustawienia trybu: **Generuj model**.



Rysunek 6. Węzeł modelowania CHAID, karta Opcje budowania

Na potrzeby przykładu zachowamy drzewo zupełnie proste, aby ograniczyć rozbudowę drzewa do minimalnej liczby obserwacji dla węzłów nadrzędnych i podrzędnych.

2. Na karcie Opcje budowania wybierz opcję **Reguły zatrzymujące** z panelu nawigacji po lewej stronie.
3. Wybierz opcję **Wartość bezwzględna**.
4. Ustaw wartość **Minimum rekordów w gałęzi nadrzędnej** na 400.
5. Ustaw wartość **Minimum rekordów w gałęzi podrzędnej** na 200.

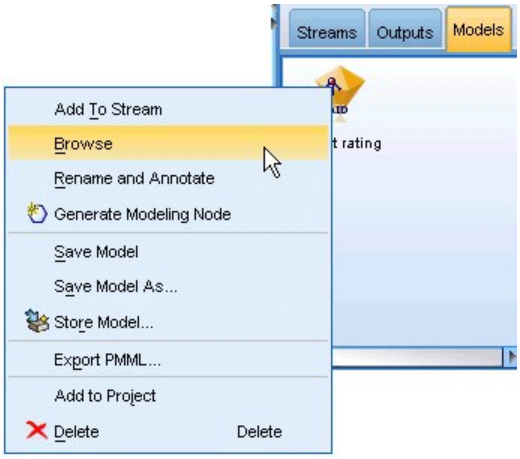


Rysunek 7. Ustawianie kryteriów zatrzymywania dla budowania drzewa decyzyjnego

W tym przykładzie można użyć wszystkich pozostałych opcji domyślnych, dlatego kliknij przycisk **Wykonaj**, aby utworzyć model. (Możesz też kliknąć prawym przyciskiem myszy węzeł i wybrać opcję **Wykonaj** z menu kontekstowego lub zaznaczyć węzeł i wybrać **Wykonaj** z menu Narzędzia).

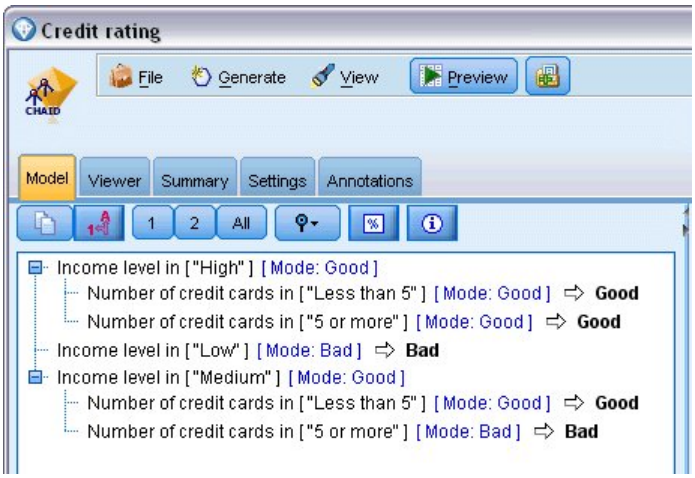
Przeglądanie modelu

Po zakończeniu wykonywania model użytkowy jest dodawany do palety modeli w prawym górnym rogu okna aplikacji, a także umieszczany w obszarze roboczym strumienia z odsyłaczem do węzła modelowania, z którego został utworzony. Aby wyświetlić szczegóły modelu, kliknij prawym przyciskiem myszy model użytkowy i wybierz opcję **Przeglądaj** (z palety modeli) lub **Edycja** (z obszaru roboczego).



Rysunek 8. Paleta modeli

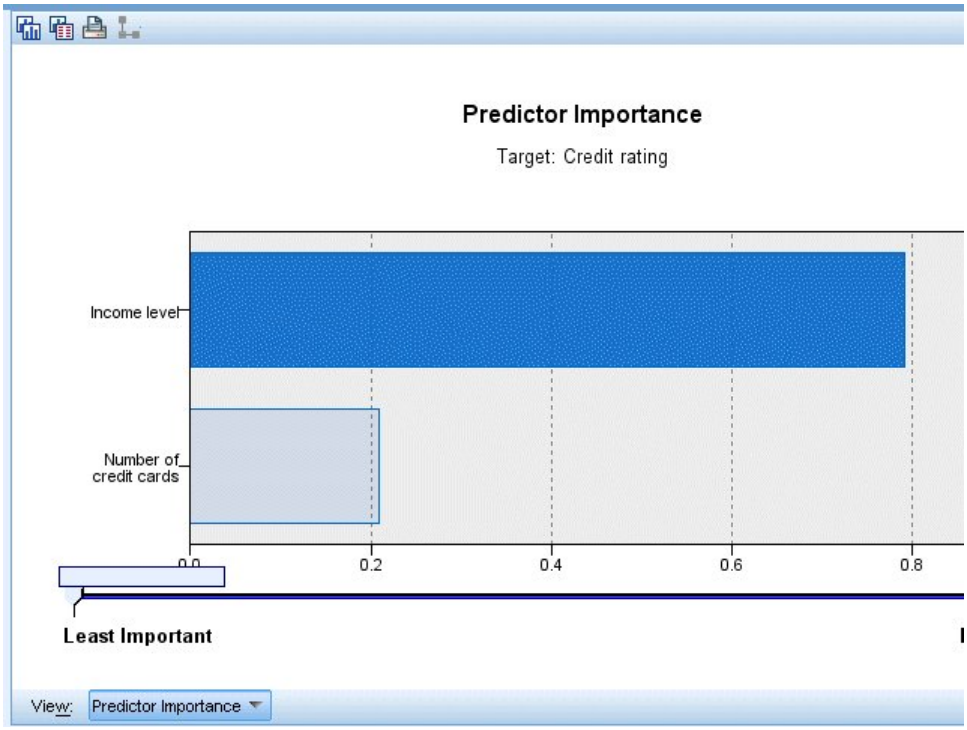
W przypadku modelu użytkowego CHAID na karcie Model szczegóły są wyświetlane w postaci zestawu reguł — zwykle jest to szereg reguł, jakie można zastosować w celu przypisania poszczególnych rekordów do węzłów podrzędnych w oparciu o wartości różnych zmiennych wejściowych.



Rysunek 9. Model użytkowy CHAID, zestaw reguł

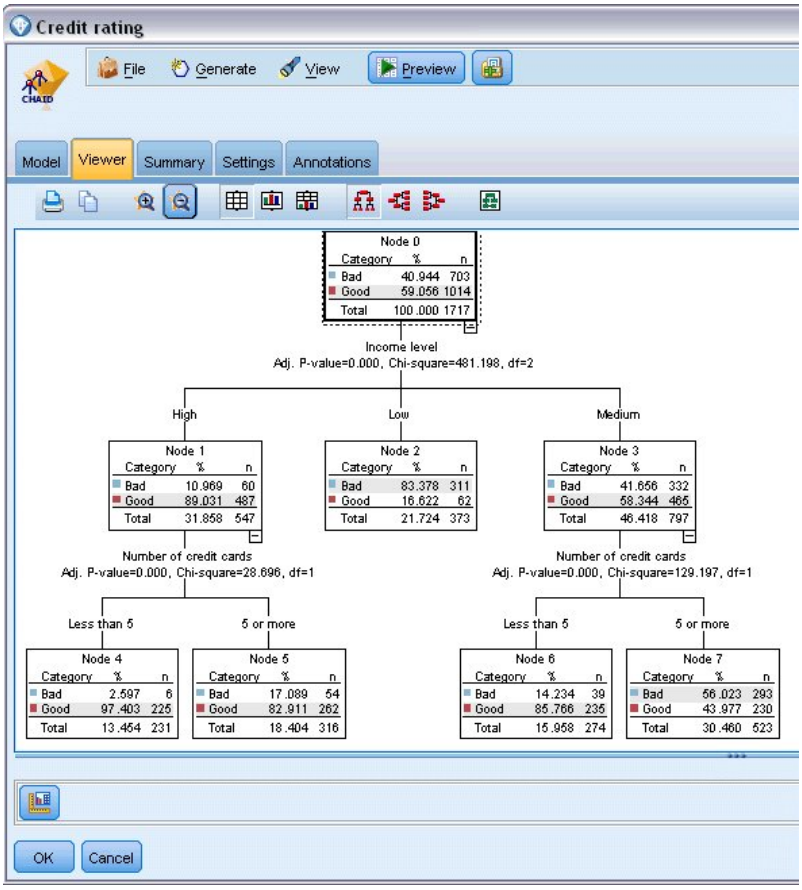
Dla każdego węzła końcowego drzewa decyzyjnego — to znaczy dla tych węzłów, które nie są dalej podzielone — zwracana jest predykcja *Good* lub *Bad*. W każdym przypadku predykcja jest określana według **dominanty** lub najczęściej udzielanej odpowiedzi dla rekordów, które należą do tego węzła.

Po prawej stronie zestawu reguł na karcie Model wyświetlany jest wykres Ważność predyktorów, który przedstawia względną ważność poszczególnych predyktorów w oszacowaniu modelu. Można tutaj zauważyć, że zmienna *Income level* jest w tym przypadku najbardziej istotna, a innym istotnym czynnikiem jest jedynie zmienna *Number of credit cards*.



Rysunek 10. Wykres ważności predyktorów

Na karcie Przegląd w modelu użytkowym wyświetlany jest ten sam model w postaci drzewa, z węzłem w każdym punkcie decyzyjnym. Elementy sterujące zmiany wielkości na pasku narzędzi umożliwiają powiększenie konkretnego węzła lub pomniejszanie obrazu, tak aby widoczny był większy obszar drzewa.



Rysunek 11. Karta Przegląd w modelu użytkowym, z wybraną opcją pomniejszania

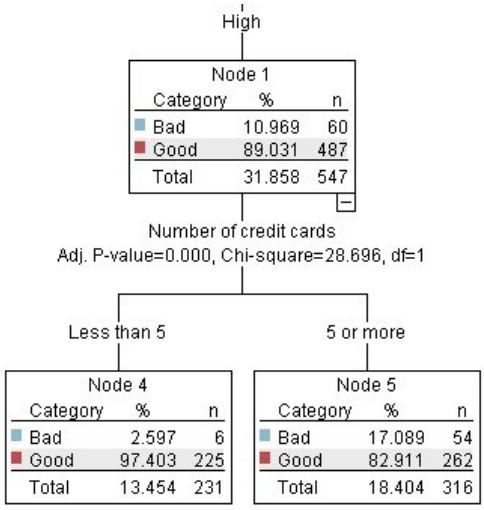
Patrząc na górną część drzewa w pierwszym węźle (węzeł 0) widzimy podsumowanie dla wszystkich rekordów ze zbioru danych. Tylko ponad 40% obserwacji ze zbioru danych jest klasyfikowanych jako wysokie ryzyko. Jest to dość duża proporcja, dlatego sprawdźmy, czy drzewo zawiera jakieś wskazówki, jakie czynniki mogą za to odpowiadać.

Pierwszy podział jest dokonany wg zmiennej *Income level*. Rekordy, w których poziom dochodu należy do kategorii *Low* są przypisane do węzła 2, dlatego nie powinno dziwić, że ta kategoria zawiera najwyższy procent osób, które nie spłacają kredytu. Niewątpliwie udzielenie kredytów klientom należącym do tej kategorii wiąże się z wysokim ryzykiem.

Jednak 16% klientów z tej kategorii faktycznie *nie ma* zaległości kredytowych, dlatego predykcja nie zawsze będzie poprawna. Żaden model nie może realnie przewidzieć każdej odpowiedzi, jednak dobry model powinien umożliwiać przewidzenie odpowiedzi *najbardziej prawdopodobnej* dla każdego rekordu w oparciu o dostępne dane.

Podobnie, jeśli spojrzymy na klientów z wysokim dochodem (węzeł 1), zauważymy, że duża większość (89%) jest obarczona małym ryzykiem. Jednak więcej niż 1 na 10 z tych klientów również nie spłaca kredytu. Czy można udoskonalić kryteria udzielania kredytu, aby zminimalizować ryzyko?

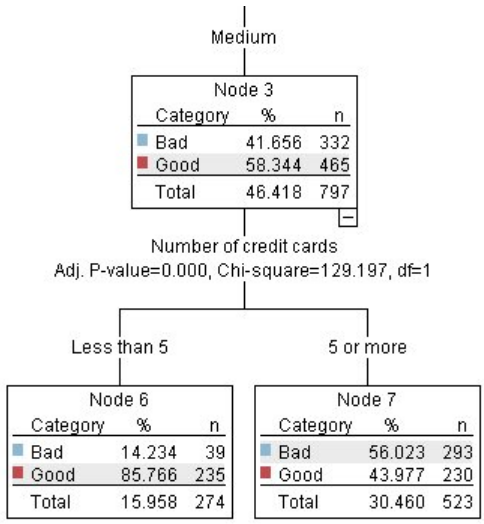
Należy zwrócić uwagę, jak model podzielił klientów na dwie podkategorie (węzły 4 i 5) w oparciu o liczbę posiadanych kart kredytowych. W przypadku klientów z wysokim dochodem, jeśli kredyt zostanie udzielony tylko tym osobom, które mają mniej niż 5 kart kredytowych, można zwiększyć wskaźnik sukcesu z 89% do 97% i uzyskać jeszcze bardziej zadowalający wynik.



Rysunek 12. Widok drzewa klientów z wysokim dochodem

Co jednak z klientami należącymi do kategorii osób ze średnim dochodem (węzeł 3)? Są oni dużo bardziej równomiernie podzieleni pomiędzy ocenami wysokiego i niskiego ryzyka.

Ponownie pomóc mogą podkategorie (w tym przypadku węzły 6 i 7). Tym razem udzielenie kredytu tylko klientom ze średnim dochodem posiadającym mniej niż 5 kart kredytowych zwiększy procent ocen niskiego ryzyka z 58% do 85%, co stanowi znaczne udoskonalenie.



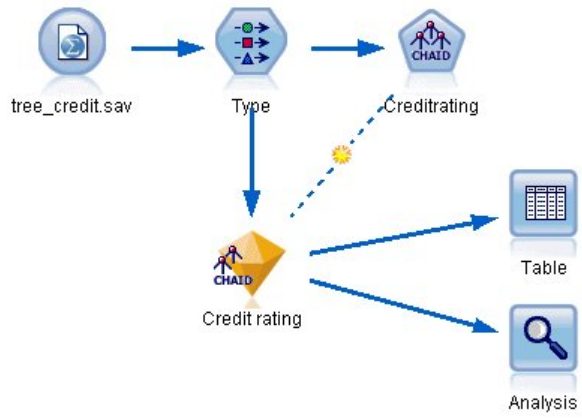
Rysunek 13. Widok drzewa klientów ze średnim dochodem

Dowiedzieliśmy się zatem, że każdy rekord dodany do tego modelu, będzie przypisany do konkretnego węzła, a do każdego węzła przypisana zostanie predykcja *Good* lub *Bad* w oparciu o najczęściej udzielaną odpowiedź.

Ten proces przypisywania predykcji do poszczególnych rekordów nazywany jest **ocenianiem**. Poprzez ocenianie tych samych rekordów użytych do oszacowania modelu można określić ich dokładność w odniesieniu do danych uczących — danych, dla których wynik jest znany. Dowiedzmy się, jak można to zrobić.

Ewaluacja modelu

Przeglądaliśmy model, aby zrozumieć, jak działa ocenianie. Jednak do ewaluacji *dokładności* tego procesu, konieczna jest ocena niektórych rekordów i porównanie odpowiedzi przewidzianych przez model z rzeczywistymi wynikami. Przeprowadzimy ocenę tych samych rekordów, jakie zostały użyte do oszacowania modelu, co pozwoli nam na porównanie obserwowanych i przewidzianych odpowiedzi.

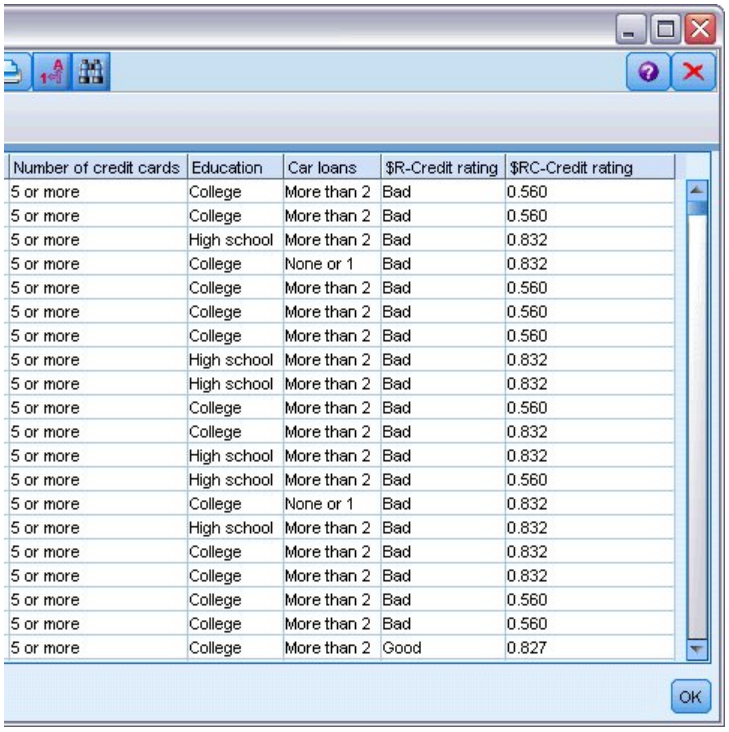


Rysunek 14. Dołączanie modelu użytkowego do węzłów wynikowych w celu przeprowadzenia ewaluacji modelu

1. Aby zobaczyć oceny lub predykcje, należy dołączyć węzeł tabeli do modelu użytkowego, kliknąć dwukrotnie węzeł tabeli, a następnie kliknąć przycisk **Wykonaj**.

   W tabeli przewidziane oceny są wyświetlane w postaci zmiennej o nazwie *$R-Credit rating*, która została utworzona przez model. Można porównać te wartości z oryginalną zmienną *Credit rating*, która zawiera rzeczywiste odpowiedzi.

   Zgodnie z konwencją nazwy zmiennych wygenerowanych podczas oceniania są tworzone na podstawie zmiennej przewidywanej, ale dodawany jest standardowy przedrostek. Przedrostki *$G* i *$GE* są generowane przez uogólniony model liniowy, *$R* to przedrostek używany dla predykcji wygenerowanych przez model CHAID, *$RC* dotyczy współczynnika ufności, przedrostek *$X* jest zwykle generowany w przypadku użycia zespołów, a przedrostki *$XR*, *$XS* i *$XF* są używane, w przypadku gdy zmienna przewidywana jest odpowiednio zmienną ilościową, jakościową, zmienną typu zbiór lub zmienną typu flaga. Różne typy modeli używają różnych zestawów przedrostków. **Współczynnik ufności** to własne oszacowanie modelu, w skali od 0,0 do 1,0, określające dokładność poszczególnych przewidywanych wartości.



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	College	None or 1	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.560
5 or more	College	None or 1	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Good	0.827

OK

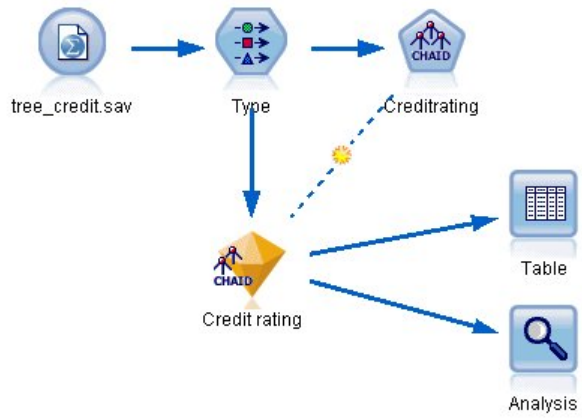
Rysunek 15. Tabela przedstawiająca wygenerowane oceny i współczynniki ufności

Zgodnie z oczekiwaniami przewidywana wartość jest zgodna z rzeczywistymi odpowiedziami dla wielu rekordów, ale nie dla wszystkich. Przyczyną jest fakt, że w każdym końcowym węźle CHAID znajdują się różne odpowiedzi. Predykcja jest zgodna z tą *najczęściej udzielaną*, ale będzie zła dla wszystkich pozostałych z tego węzła. (Przypominamy o 16-procentowej mniejszości klientów z niskim dochodem, którzy nie mają zaległości w spłatach).

Aby tego uniknąć, można kontynuować podział drzewa na coraz to mniejsze gałęzie, aż każdy węzeł będzie w 100% czysty — tylko wartości *Good* lub *Bad*, bez pomieszanych odpowiedzi. Jednak taki model będzie niezwykle skomplikowany i prawdopodobnie nie będzie na tyle uogólniony, aby mógł być zastosowany do innych zbiorów danych.

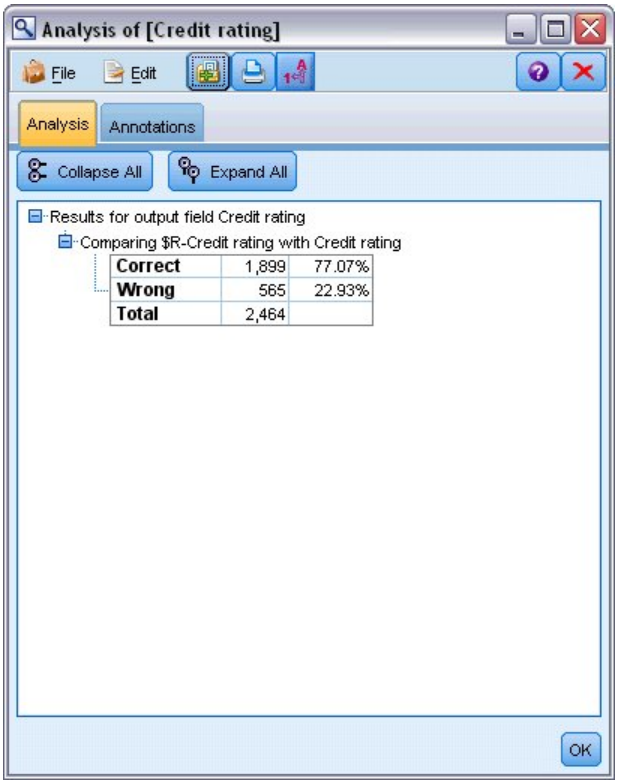
Aby dokładnie dowiedzieć się, ile predykcji jest poprawnych, można przejrzeć tabelę i zliczyć liczbę rekordów, w których wartość przewidzianej zmiennej *$R-Credit rating* jest zgodna z wartością zmiennej *Credit rating*. Na szczęście istnieje dużo łatwiejszy sposób — można użyć węzła Analiza, który zrobi to automatycznie.

2. Połącz model użytkowy z węzłem Analiza.
3. Kliknij dwukrotnie węzeł Analiza i kliknij przycisk **Wykonaj**.



Rysunek 16. Dołączanie węzła Analiza

Analiza przedstawia, że 1899 z 2464 rekordów — ponad 77% — wartości przewidzianych przez model jest zgodnych z rzeczywistymi odpowiedziami.



Rysunek 17. Porównywanie wyników analizy z odpowiedziami obserwowanymi i rzeczywistymi

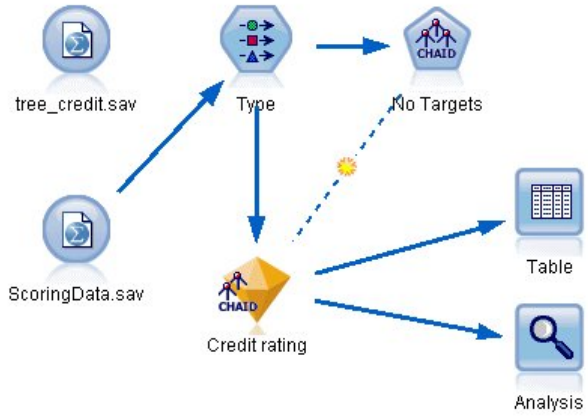
Wynik jest ograniczony przez fakt, że oceniane są te same rekordy, jakie zostały użyte do oszacowania modelu. W rzeczywistości można użyć węzła Podział, aby podzielić dane na osobne próby do uczenia i ewaluacji.

Użycie jednego przykładowego podziału do wygenerowania modelu i drugiego do przetestowania go pozwoli dużo lepiej wskazać poziom uogólnienia modelu dla innych zbiorów danych.

Węzeł Analiza umożliwia przetestowanie modelu w odniesieniu do rekordów, dla których wynik rzeczywisty jest już znany. Kolejny etap przedstawia sposób użycia modelu do oceny rekordów, dla których wynik nie jest znany. Przykładowo, może to dotyczyć osób, które obecnie nie są klientami banku, ale są potencjalnymi adresatami korespondencji promocyjnej.

Ocenianie rekordów

Wcześniej ocenialiśmy te same rekordy, jakie zostały użyte do oszacowania modelu w celu ewaluacji jego dokładności. Teraz dowiemy się, jak przeprowadzić ocenę innego zestawu rekordów niż te, których użyto do utworzenia modelu. Celem modelowania z użyciem rekordów zmiennej przewidywanej: Study, dla których wynik jest znany, jest określenie wzorów, które pozwolą przewidzieć wyniki, które jeszcze nie są znane.



Rysunek 18. Dołączanie nowych danych do oceny

Istnieje możliwość zaktualizowania węzła źródłowego Plik Statistics, tak aby wskazywał inny plik danych lub dodania nowego węzła źródłowego, który będzie odczytywał dane, jakie mają zostać poddane ocenie. Niezależnie od metody nowy zbiór danych musi zawierać te same zmienne wejściowe, jakie zostały użyte przez model (*Age*, *Income level*, *Education* itd.), ale bez zmiennej przewidywanej *Credit rating*.

Alternatywnie, można dodać model użytkowy do dowolnego strumienia, który obejmuje oczekiwane zmienne wejściowe. Niezależnie od tego, czy odczyt będzie z pliku czy z bazy danych, typ źródła nie ma znaczenia, o ile nazwy i typy zmiennych są zgodne z użytymi przez model.

Można również zapisać model użytkowy jako osobny plik lub wyeksportować model w formacie PMML do użycia z innymi aplikacjami, które ten format obsługują, albo zapisać model w repozytorium IBM SPSS Collaboration and Deployment Services, które umożliwia wdrożenie, analizowanie i zarządzanie modelami w całym przedsiębiorstwie.

Niezależnie od zastosowanej infrastruktury sam model działa w taki sam sposób.

Podsumowanie

Ten przykład przedstawia podstawowe etapy tworzenia, ewaluacji i oceniania modelu.

- Węzeł modelowania dokonuje oszacowania modelu poprzez badanie rekordów, dla których wynik jest znany i tworzy model użytkowy. Czasami ten proces jest nazywany uczeniem modelu.
- Model użytkowy może zostać dodany do dowolnego strumienia z oczekiwanymi zmiennymi w celu przeprowadzenia oceny rekordów. Ocenianie rekordów, dla których wynik jest już znany (np. dla istniejących klientów), pozwala na ocenę poprawności działania.
- Jeśli działanie modelu jest satysfakcjonujące, można ocenić nowe dane (np. dla potencjalnych klientów), aby przewidzieć ich odpowiedzi.

- Dane użyte do uczenia lub oszacowania modelu mogą być określane jako dane analityczne lub historyczne; dane oceniania mogą być również określane jako dane operacyjne.

Rozdział 3. Przegląd modelowania

Przegląd węzłów modelowania

Oprogramowanie IBM SPSS Modeler umożliwia korzystanie z wielu metod modelowania opartych na sztucznej inteligencji, uczeniu maszynowym i statystykach. Metody dostępne na palecie Modelowanie pozwalają na ekstrahowanie nowych informacji z danych i tworzenie modeli predykcyjnych. Każda metoda ma określone mocne strony i jest dostosowana do rozwiązywania określonych problemów.

Publikacja *IBM SPSS Modeler — podręcznik zastosowań* zawiera przykłady zastosowania wielu z tych metod wraz z ogólnym wprowadzeniem do procesu modelowania. Ten podręcznik jest dostępny jako samouczek onlineoraz jako plik w formacie PDF. Więcej informacji można znaleźć w temacie “Przykłady zastosowań” na stronie 4.

Metody modelowania dzielą się na niniejsze kategorie:

- Nadzorowane
- Związek
- Segmentacja

Modele nadzorowane

Modele nadzorowane korzystają z wartości jednej lub większej liczby zmiennych **wejściowych** do przewidywania wartości jednej lub większej liczby zmiennych wyjściowych lub **przewidywanych**. Niektóre z przykładów takich technik to: drzewa decyzyjne (C&RT, QUEST, CHAID i algorytmy C5.0), regresja (liniowa, logistyczna, uogólniona liniowa oraz algorytmy regresji Coxa), sieci neuronowe, algorytmy SVM oraz sieci Bayesowskie.

Modele nadzorowane pomagają organizacjom przewidywać znany wynik, np. fakt albo rezygnacji z zakupu bądź też dopasowania transakcji do znanego wzorca oszustwa. Techniki modelowania obejmują także uczenie maszynowe, wywodzenie reguł, identyfikację podgrup, metody statystyczne i generowanie wielu modeli.

Węzły nadzorowane

Węzeł Auto Klasyfikacja tworzy i porównuje różne modele pod kątem wyników binarnych (tak lub nie, odejścia lub brak odejścia itd.), umożliwiając użytkownikowi wybór optymalnego podejścia do danej analizy. Obsługiwana jest pewna liczba algorytmów modelowania, co umożliwia wybór metod, które mają zostać użyte, konkretnych opcji dla każdej z nich oraz kryteriów porównywania wyników. Węzeł generuje zestaw modeli w oparciu o określone opcje i nadaje rangi najlepszym kandydatom wybranym według wskazanych kryteriów.

Węzeł Auto Predykcja estymuje i porównuje modele zwracające wyniki w formie ciągłego przedziału liczbowego, korzystając z szeregu różnych metod. Węzeł działa tak samo, jak węzeł Auto Klasyfikacja, umożliwiając użytkownikowi wybór używanych algorytmów oraz eksperymentowanie z wieloma kombinacjami opcji w pojedynczym przebiegu modelowania. Obsługiwane algorytmy obejmują sieci neuronowe, C&RT, CHAID, regresję liniową, uogólnioną regresję liniową oraz algorytmy SVM. Modele można porównywać na podstawie korelacji, błędu względnego lub liczby używanych zmiennych.

Węzeł klasyfikacji i regresji (C&RT) generuje drzewo decyzyjne umożliwiające predykcję lub klasyfikację przyszłych obserwacji. W metodzie tej stosowany jest rekursywny podział rekordów na segmenty przez minimalizację zanieczyszczeń w każdym kroku, przy czym węzeł w drzewie jest uważany za „czysty”, jeśli 100% obserwacji w węźle przypada na konkretną kategorię zmiennej przewidywanej. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi lub jakościowymi (nominalnymi, porządkowymi lub flagami); wszystkie podziały są binarne (tylko dwie podgrupy).

Węzeł QUEST oferuje metodę klasyfikacji binarnej służącą do budowania drzew decyzyjnych, zaprojektowaną w celu redukcji czasu przetwarzania analiz dużych drzew decyzyjnych C&R, a jednocześnie w celu redukcji tendencji obecnej w metodach drzew klasyfikacji do preferowania danych wejściowych dopuszczających więcej podziałów. Zmienne wejściowe mogą być zakresami liczbowymi (ciągłymi), lecz zmienna przewidywana musi być jakościowa. Wszystkie podziały są binarne.

Węzeł CHAID generuje drzewa decyzyjne, korzystając ze statystyk chi-kwadrat w celu identyfikacji optymalnych podziałów. W odróżnieniu od węzłów C&RT i węzłów QUEST, CHAID może generować drzewa niebinarne, co oznacza, że niektóre podziały mają więcej niż dwie gałęzie. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi (ilościowymi) lub jakościowymi. Wyczerpujący CHAID stanowi modyfikację CHAID umożliwiającą dokładniejsze badanie wszystkich możliwych podziałów, lecz obliczenia w jego przypadku zajmują więcej czasu.

Węzeł C5.0 tworzy drzewo decyzyjne lub zestaw reguł. Model działa w oparciu o podział próby na podstawie zmiennej oferującej maksimum korzyści z informacji na każdym z poziomów. Zmienna przewidywana musi być jakościowa. Dozwolonych jest wiele podziałów na więcej niż dwie podgrupy.

Węzeł Lista decyzyjna identyfikuje podgrupy lub segmenty wskazujące wyższe lub niższe prawdopodobieństwo danego wyniku binarnego względem całej populacji. Można na przykład wyszukać klientów, których prawdopodobieństwo odejścia jest niewielkie, lub którzy z dużym prawdopodobieństwem pozytywnie zareagują na kampanię. Istnieje możliwość zastosowania posiadanej wiedzy biznesowej w modelu przez dodanie własnych, niestandardowych segmentów i przejrzenie modeli alternatywnych jeden obok drugiego w celu porównania wyników. Modele Lista decyzyjna składają się z list reguł, w których każda reguła ma warunek i wynik. Reguły są stosowane w kolejności wprowadzania, a pierwsza reguła spełniona określa wynik.

Modele regresji liniowej przewidują docelową wartość ilościową na podstawie liniowych relacji między docelową wartością ilościową a jednym lub większą liczbą predyktorów.

Węzeł analizy PCA/czynnikowej udostępnia wydajne techniki redukcji danych pozwalające obniżyć stopień ich złożoności. Analiza głównych składowych (ang. Principal Components Analysis, PCA) znajduje kombinacje liniowe zmiennych wejściowych, które umożliwiają określenie wariancji w całym zestawie zmiennych, pod warunkiem że składowe są zlokalizowane ortogonalnie (prostopadle) do siebie. Analiza czynnikowa próbuje zidentyfikować współczynniki objaśniające wzory korelacji występujące w ramach zbiorów obserwowanych zmiennych. W przypadku obu podejść celem jest znalezienie niewielkiej liczby zmiennych wyliczanych w efektywny sposób, która podsumowuje informacje w oryginalnym zestawie zmiennych.

Węzeł Dobór predyktorów przegląda zmienne wejściowe do usunięcia w oparciu o zbiór kryteriów (takich jak procent braków danych); następnie nadaje rangę istotności pozostałych danych wejściowych względem określonej zmiennej przewidywanej. Na przykład, jeśli mamy zbiór danych z setkami potencjalnych danych wejściowych, to które z nich z dużym prawdopodobieństwem okażą się użyteczne w modelowaniu wyników leczenia pacjenta?

Analiza dyskryminacyjna opiera się na ściślejszych założeniach niż regresja logistyczna, lecz może stanowić wartościową alternatywę lub uzupełnienie analizy metodą regresji logistycznej w przypadku spełnienia tych założeń.

Regresja logistyczna to technika statystyczna umożliwiająca klasyfikację rekordów na podstawie wartości zmiennych wejściowych. Jest ona analogiczna do regresji liniowej, lecz bazuje na przewidywanej zmiennej jakościowej zamiast na przedziale liczbowym.

Węzeł Modele uogólnione rozszerza ogólny model liniowy w taki sposób, że zmienna zależna jest liniowo powiązana z czynnikami i współzmiennymi za pośrednictwem określonej funkcji łączenia. Model pozwala ponadto, aby zmienna zależna nie miała rozkładu normalnego. Obejmuje ona funkcjonalność dużej liczby modeli statystycznych, m.in. regresji liniowej, regresji logistycznej, modeli logarytmiczno-liniowych dla danych o liczebności.

Uogólniony liniowy model mieszany (GLMM) stanowi wersję modelu liniowego rozszerzoną w taki sposób, że zmienna przewidywana może mieć rozkład inny niż normalny, jest liniowo powiązana z czynnikami i współzmiennymi za pośrednictwem określonej funkcji łączenia, a obserwacje mogą być skorelowane. Uogólnione liniowe modele mieszane obejmują szeroki wachlarz modeli, począwszy od prostych modeli regresji liniowej, aż po złożone wielopoziomowe modele dla danych z obserwacji długofalowych nieposiadających rozkładu normalnego.

Węzeł regresji Coxa umożliwia utworzenie modelu przeżycia dla danych określających czasy do wystąpienia zdarzeń i zawierających ocenzurowane rekordy. Model generuje funkcję przeżycia przewidującą prawdopodobieństwo, że zdarzenie będące przedmiotem zainteresowania wystąpiło w określonym czasie (*t*) dla danych wartości zmiennych wejściowych.

Węzeł SVM umożliwia szybką klasyfikację danych do jednej lub dwu grup bez przeuczenia. Algorytm SVM działa prawidłowo dla szerokiego zbioru danych, na przykład takiego o bardzo dużej liczbie zmiennych wejściowych.

Węzeł Sieci Bayesa umożliwia utworzenie modelu prawdopodobieństwa przez połączenie zaobserwowanych i zarejestrowanych dowodów z wiedzą rzeczywistą w celu ustanowienia prawdopodobieństwa występowania. Węzeł koncentruje się na sieciach Tree Augmented Naïve Bayes (TAN) i Markov Blanket, używanych głównie do klasyfikacji.

Węzeł SLRM (model odpowiedzi samonauczania) umożliwia utworzenie modelu, w którym pojedyncza nowa obserwacja lub niewielka liczba nowych obserwacji może zostać użyta do ponownej oceny modelu bez konieczności ponownego uczenia modelu z wykorzystaniem wszystkich danych.

Węzeł Szereg czasowy umożliwia estymację modelu wygładzania wykładniczego, modelu autoregresyjnej zintegrowanej średniej ruchomej (ARIMA) jednej zmiennej oraz modelu ARIMA wielu zmiennych (lub funkcji przenoszenia) dla danych szeregów czasowych i generuje prognozy przyszłych wyników. Ten węzeł Szereg czasowy jest podobny do poprzedniego węzła Szereg czasowy, odrzuconego w produkcie SPSS Modeler, wersja 18. Jednak ten nowszy Szereg czasowy jest przeznaczony do wykorzystania możliwości programu IBM SPSS Analytic Server w zakresie przetwarzania dużych zbiorów danych oraz wyświetlania modelu wynikowego w przeglądarce wyników, która została dodana w programie SPSS Modeler, wersja 17.

Węzeł KNN (*k*-najbliższego sąsiedztwa) wiąże nową obserwację z kategorią lub wartością *k* (gdzie *k* jest liczbą całkowitą) najbliższych obiektów w przestrzeni predyktora. Podobne obserwacje znajdują się blisko siebie, a niepodobne — daleko.

Węzeł STP (predykcji przestrzenno-czasowej) używa danych zawierających informacje o lokalizacji, zmiennych wejściowych predykcji (predyktorów), zmiennej czasu i zmiennej przewidywanej. W danych z każdą lokalizacją powiązany jest szereg wierszy, które odzwierciedlają wartości predyktorów w różnych punktach w czasie. Po przeanalizowaniu danych mogą być one używane do przewidywania wartości w dowolnej lokalizacji w danych kształtu używanych w analizie.

Modele asocjacyjne

Modele asocjacyjne znajdują wzorce w danych, w których jeden lub więcej obiektów (takich jak zdarzenia, zakupy czy atrybuty) jest powiązanych z jednym lub większą liczbą z pozostałych obiektów. Modele te tworzą zestawy reguł definiujące relacje. W tym miejscu zmienne w ramach danych pełnią rolę zarówno danych wejściowych, jak i docelowych. Związki te można znaleźć również ręcznie, lecz algorytmy reguł asocjacyjnych pozwalają wykonać te operacje znacznie szybciej i umożliwiają eksplorację bardziej złożonych wzorców. Modele Apriori i Carma stanowią przykłady użycia takich algorytmów. Jednym z kolejnych typów modeli asocjacyjnych jest model wykrywania kolejności, który znajduje wzorce sekwencyjne w danych ustrukturyzowanych względem czasu.

Modele asocjacyjne są najbardziej użyteczne w przypadku przewidywania wielokrotnych danych wynikowych — na przykład klienci, którzy kupili produkt X, kupili także produkty Y i Z. Modele asocjacyjne umożliwiają powiązanie konkretnego wniosku (np. decyzji o zakupie) z zestawem warunków. Przewagą algorytmów reguł asocjacyjnych wobec bardziej standardowych algorytmów drzew decyzyjnych (C5.0 i C&RT) jest fakt, że dozwolone są w nich związki między dowolnymi atrybutami. Algorytm drzewa decyzyjnego pozwala utworzyć reguły z tylko jednym wnioskiem, podczas gdy algorytmy powiązań próbują znaleźć wiele reguł, z których każda może mieć inny wniosek.

Węzły powiązań

Węzeł Apriori pozwala wyodrębnić zestaw reguł na podstawie danych, pobierając reguły o najwyższej możliwej zawartości informacji. Apriori oferuje pięć różnych metod reguł wybierania i korzysta ze złożonego schematu indeksowania do efektywnego przetwarzania dużych zbiorów danych. W przypadku dużych problemów czas uczenia Apriori jest zwykle krótszy. Brak jest arbitralnego limitu co do liczby reguł do utrzymania, możliwa jest obsługa reguł z maksymalnie 32 predykcjami. Apriori wymaga, aby wszystkie zmienne wejściowe i wyjściowe były zmiennymi jakościowymi, lecz oferuje wyższą wydajność z uwagi na optymalizację pod kątem tego typu danych.

Model CARMA pozwala wyodrębnić zestaw reguł na podstawie danych bez konieczności określania zmiennych wejściowych lub przewidywanych. Inaczej niż Apriori, węzeł CARMA oferuje ustawienia tworzenia dla obsługi reguł (pokrycie poprzedników i następników) zamiast pokrycia tylko poprzedników. Oznacza to, że wygenerowane reguły mogą być używane w szerszym spektrum zastosowań — na przykład w celu znalezienia listy produktów lub usług (poprzedników), z których wynikać będzie decyzja o promowaniu konkretnego produktu (następnika) w tegorocznym sezonie świątecznym.

Węzeł Sekwencje wykrywa reguły asocjacyjne w danych sekwencyjnych lub zorientowanych czasowo. Sekwencja to lista zbiorów elementów z tendencją do występowania w przewidywalnej kolejności. Na przykład klient dokonujący zakupu brzytwy i balsamu po goleniu przy następnej wizycie w sklepie może dokonać zakupu kremu po goleniu. Węzeł Sekwencje bazuje na algorytmie reguł asocjacyjnych CARMA, który korzysta z efektywnej metody dwu przejść do znajdowania sekwencji.

Węzeł Reguły asocjacyjne jest podobny do węzła Apriori; jednak inaczej niż w przypadku Apriori, węzeł Reguły asocjacyjne umożliwia przetwarzanie danych w postaci listy. Ponadto węzeł Reguły asocjacyjne może być używany wraz z IBM SPSS Analytic Server do przetwarzania dużych zbiorów danych i korzystania z szybszego przetwarzania równoległego.

Modele segmentacji

Modele segmentacji dzielą dane na segmenty lub grupy rekordów o podobnych wzorcach zmiennych wejściowych. Ponieważ modele segmentacji przetwarzają jedynie zmienne wejściowe, nie mają one żadnych informacji na temat zmiennych wyjściowych ani przewidywanych. Przykłady modeli segmentacji to sieci Kohonen, grupowanie K-średnich, grupowanie dwustopniowe i wykrywanie anomalii.

Modele segmentacji (zwane również „modelami grupowania”) są szczególnie przydatne w przypadkach, gdzie konkretny wynik jest nieznany (na przykład, przy identyfikacji nowych wzorców oszustw lub identyfikacji będących potencjalnie przedmiotem zainteresowania grup w bazie danych klientów). Modele skupień koncentrują się na identyfikacji grup o podobnych rekordach i oznaczaniu rekordów etykietami zgodnie z grupą, do której należą. Jest to realizowane mimo braku wstępnej wiedzy o grupach i ich charakterystykach, i pozwala odróżnić modele grupowania od innych technik modelowania, w których brak wstępnie zdefiniowanego wyniku czy zmiennej docelowej dla modelu objętego predykcją. W przypadku tych modeli nie ma poprawnych czy niepoprawnych odpowiedzi. Ich wartość określana jest przez zdolność do przechwytywania interesujących skupień w danych i oferowania użytecznych opisów tych skupień. Modele grupowania są często używane do tworzenia grup lub segmentów, które są następnie często używane jako dane wejściowe w kolejnych analizach (na przykład dzięki segmentacji potencjalnych klientów w jednorodne podgrupy).

Węzły segmentacji

Węzeł Auto Grupowanie szacuje i porównuje modele skupień identyfikujące grupy rekordów o podobnej charakterystyce. Węzeł działa tak samo, jak pozostałe zautomatyzowane węzły modelowania, umożliwiając eksperymentowanie z wieloma kombinacjami opcji w pojedynczym przebiegu modelowania. Modele można porównywać, korzystając z miar bazowych, które pozwalają podejmować próby filtrowania i oceny przydatności modelu skupień oraz udostępniają miary bazujące na istotności poszczególnych zmiennych.

Węzeł Metoda k-średnich grupuje zbiór danych w osobne grupy (lub skupienia). Metoda ta definiuje stałą liczbę skupień, w sposób iteracyjny przypisuje rekordy do skupień i dopasowuje centra skupień do chwili, gdy dalsze pokrycie nie będzie miało wpływu na ulepszenie modelu. Zamiast prób predykcji danych wynikowych *k*-średnia korzysta z procesu znanego jako nienadzorowane uczenie w celu ujawnienia wzorców w zbiorze zmiennych wejściowych.

Węzeł Sieć Kohonena generuje typ sieci neuronowej, którą można wykorzystać do grupowania zbioru danych w osobne grupy. Po pełnym przeszkoleniu sieci rekordy podobne do siebie powinny znajdować się blisko siebie na mapie wyników, podczas gdy rekordy różne od siebie powinny znajdować się daleko od siebie. Na podstawie liczby obserwacji przechwyconych przez każdą jednostkę w modelu użytkowym można rozpoznać silne jednostki. Może to dać pojęcie o odpowiedniej liczbie skupień.

Węzeł Dwustopniowa korzysta z dwustopniowej metody grupowania. Pierwszy krok stanowi pojedynczy przebieg danych z myślą o kompresji surowych danych wejściowych w łatwy w zarządzaniu zestaw podgrup. Drugi krok korzysta z hierarchicznej metody grupowania w celu progresywnego scalania podgrup w coraz większe grupy. Metoda Dwustopniowa oferuje korzyści wynikające z automatycznego szacowania optymalnej liczby grup na potrzeby danych szkoleniowych. Pozwala ona skutecznie obsługiwać mieszane typy zmiennych i duże zbiory danych.

Węzeł Anomalie umożliwia identyfikację nietypowych obserwacji lub wartości odstających, które są niezgodne z wzorcami dla „normalnych” danych. Korzystając z tego węzła, można zidentyfikować wartości odstające nawet, jeśli nie pasują one do żadnego z wcześniej znanych wzorców oraz jeśli brak pewności co do charakteru poszukiwanych danych.

Modele eksploracji w bazie danych

Program IBM SPSS Modeler oferuje integrację narzędzi do eksploracji danych i modelowania, dostępnych w bazach danych, takich jak Oracle Data Miner i Microsoft Analysis Services. Można tworzyć, oceniać i składować modele w bazie danych — a wszystko to w ramach aplikacji IBM SPSS Modeler. Bardziej szczegółowe informacje można znaleźć w publikacji *IBM SPSS Modeler — podręcznik eksploracji w bazie danych*.

Modele IBM SPSS Statistics

Jeśli na komputerze zainstalowano kopię produktu IBM SPSS Statistics z licencją, wówczas w celu tworzenia i oceny modeli źródłowych można uzyskiwać dostęp do konkretnych procedur programu IBM SPSS Statistics i uruchamiać je z poziomu programu IBM SPSS Modeler.

Budowanie modeli rozdzielonych

Modelowanie rozdzielone umożliwia użycie jednego strumienia w celu zbudowania osobnych modeli dla każdej możliwej wartości flagi, nominalnej lub ilościowej zmiennej wejściowej, przy czym modele wynikowe pozostają dostępne z pojedynczego modelu użytkowego. Możliwe wartości dla zmiennych wejściowych mogą na różne sposoby wpływać na model. Modelowanie rozdzielone pozwala na łatwe zbudowanie modelu najlepiej dopasowanego dla każdej możliwej wartości zmiennej w pojedynczym wykonaniu strumienia.

Należy zwrócić uwagę na to, że w sesjach modelowania interaktywnego nie można stosować rozdzielania. W przypadku modelowania interaktywnego należy określić każdy model osobno, co powoduje, że stosowanie rozdzielania, które prowadzi do automatycznego zbudowania wielu modeli, nie będzie miało sensu.

Modelowanie rozdzielone działa poprzez wyznaczanie konkretnej zmiennej wejściowej jako zmiennej podziału. W tym celu można ustawić rolę zmiennej na **Separacja** w specyfikacji typu.

Na zmienne podziału można wyznaczyć tylko zmienne z poziomem pomiaru **Flaga**, **Nominalna**, **Porządkowa** lub **Ilościowa**.

Jako zmienną podziału można przypisać więcej niż jedną zmienną wejściową. W takim przypadku jednak liczba utworzonych modeli może znacznie wzrosnąć. Model jest budowany dla każdej możliwej kombinacji wartości wybranych zmiennych podziału. Na przykład, jeśli trzy zmienne wejściowe, z których każda posiada trzy możliwe wartości, zostaną wyznaczone jako zmienne podziału, spowoduje to utworzenie 27 różnych modeli.

Nawet po przypisaniu jednej lub większej liczby zmiennych jako zmiennych podziału nadal możliwe jest określenie, czy zostaną utworzone modele rozdzielone, czy model pojedynczy — do tego służy pole wyboru dostępne w oknie dialogowym modelowania.

Jeśli zdefiniowane są zmienne podziału, ale pole wyboru nie jest zaznaczone, wówczas wygenerowany zostanie tylko jeden model. I podobnie — jeśli pole wyboru jest zaznaczone, ale nie zdefiniowano żadnej zmiennej podziału, wówczas podział zostanie zignorowany i zostanie wygenerowany jeden model.

W momencie wykonania strumienia w tle budowane są osobne modele dla każdej możliwej wartości zmiennej lub zmiennych podziału, ale w palecie modeli i obszarze roboczym strumienia umieszczany jest tylko jeden model użytkowy. Rozdzielony model użytkowy jest oznaczony symbolem podziału; symbol tworzą dwa szare prostokąty nałożone na obraz modelu użytkowego.

Podczas przeglądania rozdzielonego modelu użytkowego widoczna jest lista wszystkich osobnych modeli, które zostały zbudowane.

Pojedynczy model z listy można zbadać, klikając dwukrotnie ikonę jego modelu użytkowego w przeglądarce. W konsekwencji zostanie otwarte standardowe okno przeglądarki przeznaczone dla pojedynczego modelu. Gdy model użytkowy znajduje się w obszarze roboczym, dwukrotne kliknięcie miniatury wykresu powoduje otwarcie wykresu pełnego rozmiaru. Więcej informacji można znaleźć w temacie “Przeglądarka podzielonych modeli” na stronie 47.

Jeśli model został utworzony jako rozdzielony, nie można usunąć rozdziału z tego modelu ani nie można cofnąć rozdziału na dalszych etapach, takich jak węzeł lub model użytkowy.

Przykład. Firma sprzedająca na terenie całego kraju zamierza oszacować wartość sprzedaży wg kategorii produktów w każdym sklepie na terenie kraju. Korzystając z modeli rozdzielonych pracownicy firmy przypisują zmienną Sklep z danych wejściowych jako zmienną podziału, co umożliwia zbudowanie osobnych modeli dla każdej kategorii w każdym sklepie, w toku jednej operacji. Następnie pracownicy firmy mogą użyć wynikowych informacji w celu kontrolowania poziomów zapasów w sposób dalece bardziej dokładny niż w przypadku pojedynczego modelu.

Rozdział i dzielenie na podzbiory

Rozdział ma pewne cechy wspólne z dzieleniem na podzbiory, ale każdy z tych procesów jest stosowany inaczej.

Dzielenie na podzbiory powoduje podzielenie zbioru danych w sposób losowy na dwie lub trzy części: dane uczące, testowe i (opcjonalnie) walidacyjne. Jest stosowane w celu oceny wydajności jednego modelu.

Rozdział powoduje podzielenie zbioru danych na taką ilość części, ile istnieje możliwych wartości dla zmiennej podziału. Jest stosowane w celu budowania wielu modeli.

Rozdział i dzielenie na podzbiory działają w sposób całkowicie niezależny. W węźle modelowania można wybrać dowolny z tych procesów lub oba. Można również nie wybierać żadnego z nich.

Węzły modelowania obsługujące modele rozdzielone

Modele rozdzielone mogą być tworzone przez szereg węzłów modelowania. Do wyjątków należą następujące węzły: Auto Grupowanie, analizy PCA/czynnikowej, Dobór predyktorów, SLRM, Modele drzew losowych, Drzewo-AS, Liniowy-AS, LSVM, modele asocjacyjne (Apriori, Carma oraz Sekwencje), węzły modeli skupień (K-średnich, Kohonena, Dwustopniowe oraz Anomalie), węzły modeli Statistics, a także węzły używane na potrzeby modelowania w bazach danych.

Węzły modelowania, które obsługują modelowanie rozdzielone, są następujące:

C&RT


Sieci Bayesa


Liniowy

QUEST


Modele uogólnione


GLMM

CHAID


KNN


STP

C5.0

Model Coxa

SVM z jedną klasą

Sieci neuronowe

Auto Klasyfikacja

Drzewo XGBoost

Lista decyzyjna

Auto Predykcja

Liniowy XGBoost

Regresja

Regresja logistyczna

HDBSCAN

Analiza dyskryminacyjna

SVM

Szeregi czasowe

Zmienne, na które wpływa rozdział

Stosowanie modeli rozdzielonych wpływa na szereg zmiennych w produkcie IBM SPSS Modeler na różne sposoby. W niniejszej sekcji dostępne są wskazówki dotyczące stosowania modeli rozdzielonych z innymi węzłami w strumieniu.

Węzły Rekordy

Gdy modele rozdzielone są używane w strumieniu zawierającym węzeł Losowanie, należy rozdzielić rekordy na warstwy, stosując zmienną podziału, aby uzyskać równomierne próbkowanie rekordów. Ta opcja jest dostępna po wybraniu opcji *Złożone* jako metody próbkowania.

Jeśli strumień zawiera węzeł Zrównoważenie, wówczas równoważenie ma zastosowanie do całości zestawu rekordów wejściowych, a nie do podzbioru rekordów w podziale.

Jeśli w przypadku agregacji rekordów za pomocą węzła Agregacja planowane jest obliczenie agregacji dla każdego podziału, należy ustawić zmienne podziału w taki sposób, aby były zmiennymi kluczowymi.

Węzły Zmienne

Węzeł Typ to miejsce, w którym należy określić zmienną lub zmienne, które będą używane jako zmienne podziału.

Uwaga: Jeśli węzeł Zespół jest stosowany w celu połączenia dwóch lub większej liczby modeli użytkowych, nie można użyć tego węzła w celu odwrócenia podziału, ponieważ modele rozdzielone są zawarte w pojedynczym modelu użytkowym.

Węzły modelowania

Modele rozdzielone nie obsługują obliczenia ważności predyktora (względnej ważności zmiennych wejściowych predyktora podczas oszacowania modelu). Ustawienia ważności predyktora są ignorowane podczas budowania modeli rozdzielonych.

Uwaga: Ustawienia skorygowanej oceny skłonności są ignorowane w przypadku modelu rozdzielonego.

Węzeł KNN (najbliższe sąsiedztwo) obsługuje modele rozdzielone tylko wówczas, gdy jest ustawiony w taki sposób, aby przewidywał zmienne przewidywane. Ustawienie alternatywne (tylko identyfikacja najbliższego sąsiedztwa) nie powoduje utworzenia modelu. Jeśli zostanie wybrana opcja **Automatycznie wybierz wartość k**, wówczas każdy z modeli rozdzielonych może zawierać inną liczbę najbliższych sąsiadów. Zatem model ogólny zawiera wygenerowane kolumny w liczbie równej największej liczbie najbliższych sąsiadów znalezionych we wszystkich modelach rozdzielonych. W przypadku tych modeli rozdzielonych, w których liczba najbliższych sąsiadów jest mniejsza niż maksimum, jest odpowiednia liczba kolumn wypełnionych wartościami $null. Więcej informacji można znaleźć w temacie “Węzeł KNN” na stronie 343.

Węzły modelowania w bazie danych

Węzły modelowania w bazie danych nie obsługują modeli rozdzielonych.

Modele użytkowe

Eksport do PMML z modelu użytkowego modelu rozdzielonego jest niemożliwy, ponieważ model użytkowy zawiera wiele modeli, a PMML nie obsługuje takich pakietów. Eksport do formatu tekstowego lub HTML jest możliwy.

Opcje zmiennych węzła modelowania

Wszystkie węzły modelowania zawierają kartę Zmienne, na której można określić zmienne, które będą używane podczas budowania modelu.

Aby możliwe było zbudowanie modelu, konieczne jest określenie, które zmienne mają być używane jako zmienne przewidywane, a które jako dane wejściowe. We wszystkich węzłach modelowania (z kilkoma wyjątkami) stosowane są informacje na temat zmiennych z wcześniejszego węzła Typ. Korzystając z węzła Typ do wyboru zmiennych wejściowych i przewidywanych, nie trzeba zmieniać żadnych ustawień na tej karcie. (Wyjątki obejmują węzeł Kolejność oraz węzeł wyodrębniania tekstu, które wymagają, aby ustawienia zmiennej zostały określone w węźle modelowania).

Użyj ustawień węzła Typ. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej z wcześniejszego węzła Typ. Jest to ustawienie domyślne.

Użyj ustawień niestandardowych. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej określonych w tym miejscu, a nie w żadnym wcześniejszym węźle Typ. Po wybraniu tej opcji określ poniższe zmienne odpowiednio do potrzeb.

Uwaga: nie wszystkie zmienne są wyświetlane dla wszystkich węzłów.

- **Użyj formatu transakcyjnego (tylko węzły Apriori, CARMA, Reguły asocjacyjne MS i Model aprioryczny Oracle).** To pole wyboru należy zaznaczyć, jeśli dane źródłowe są dostępne w **formacie transakcyjnym**. Rekordy w tym formacie mają dwie zmienne — jedną dla identyfikatora, a drugą dla treści. Każdy rekord reprezentuje pojedynczą transakcję lub element, a skojarzone elementy są powiązane, ponieważ mają ten sam identyfikator. Usuń zaznaczenie tego pola wyboru, jeśli dane mają **format tabel**, w których elementy są reprezentowane przez osobne flagi, a każda zmienna flagi reprezentuje obecność lub brak konkretnego elementu, a każdy rekord reprezentuje pełny zestaw skojarzonych elementów. Więcej informacji można znaleźć w temacie “Dane tabelaryczne a dane transakcyjne” na stronie 258.

– **Identyfikator.** W przypadku danych transakcyjnych należy wybrać z listy zmienną identyfikacyjną. Jako zmienna identyfikacyjna mogą być używane zmienne numeryczne lub symboliczne. Każda unikalna wartość tej zmiennej powinna wskazywać na określoną jednostkę analizy. Na przykład w aplikacji do obsługi koszyka zakupów każdy identyfikator może reprezentować jednego klienta. W przypadku aplikacji do analizy dzienników sieciowych każdy identyfikator może reprezentować komputer (wg adresu IP) lub użytkownika (wg danych logowania).
– **Wartości identyfikatorów są posortowane.** (Tylko węzły Apriori i CARMA) Jeśli dane zostały wstępnie posortowane tak, że wszystkie rekordy o tym samym identyfikatorze są zgrupowane w strumieniu danych, należy wybrać tę opcję w celu przyspieszenia przetwarzania. Jeśli dane nie zostały wstępnie posortowane (lub nie ma co do tego pewności), należy pozostawić tę opcję niezaznaczoną. Węzeł posortuje dane automatycznie.

  Uwaga: Jeśli dane nie są posortowane, a użytkownik wybierze tę opcję, model może zwrócić niepoprawne wyniki.
– **Zawartość.** Należy określić zmienne zawartości dla modelu. Zmienne te zawierają interesujące elementy w procesie modelowania sekwencji. Można określić wiele zmiennych typu flaga (jeśli dane mają format tabelaryczny) lub jedną zmienną nominalną (jeśli dane mają format transakcyjny).

- **Zmienna przewidywana.** W przypadku modeli, które wymagają jednej lub większej liczby zmiennych przewidywanych, należy wybrać zmienną lub zmienne przewidywane. Działanie jest podobne, jak w przypadku ustawienia roli zmiennej na wartość *Zmienna przewidywana* w węźle Typ.
- **Ewaluacja.** (Tylko w przypadku modeli Auto Grupowanie). Dla modeli skupień nie jest określona żadna zmienna przewidywana; jednak można wybrać zmienną ewaluacyjną, aby ustalić jej poziom ważności. Dodatkowo można ocenić to, jak skutecznie grupy rozróżniają wartości tej zmiennej, co z kolei wskazuje, czy grupy mogą być używane w celu przewidzenia tej zmiennej przewidywanej. *Uwaga* Zmienna ewaluacyjna musi być łańcuchem z więcej niż jedną wartością.
  – **Zmienne wejściowe.** Umożliwia wybór zmiennej wejściowej lub kilku zmiennych. Działanie jest podobne, jak w przypadku ustawienia roli zmiennej na *Zmienna wejściowa* w węźle Typ.
  – **Podział.** To pole umożliwia określenie zmiennej używanej do podziału danych na osobne próby do uczenia, testowania i walidacji podczas budowania modelu. Korzystając z jednej próby do generowania modelu oraz innej do testowania go, można uzyskać wiarygodne informacje o tym, jak dobrze model pozwala uogólnić większe zbiory danych, podobne do danych bieżących. Jeśli korzystając z węzłów Typ lub Partycja, zdefiniowano wiele zmiennych dzielących na podzbiory, na karcie Zmienne każdego węzła modelowania korzystającego z tego podziału na podzbiory należy wybrać jedną zmienną dzielącą na podzbiory. (Jeśli obecna jest tylko jedna zmienna dzieląca na podzbiory, jest ona automatycznie używana po aktywowaniu dzielenia). Należy również pamiętać, że aby zastosować wybrany podział w analizie, dzielenie musi być również włączone na karcie Opcje modelu danego węzła. (Usunięcie zaznaczenia tej opcji umożliwia wyłączenie podziału bez zmiany ustawień zmiennych).
- **Rozdzielone.** W przypadku modeli rozdzielonych należy wybrać zmienne lub zmienną podziału. Działanie jest podobne, jak w przypadku ustawienia roli zmiennej na wartość *Rozdzielone* w węźle Typ. Na zmienne podziału można wyznaczyć tylko zmienne z poziomem pomiaru **Flaga**, **Nominalna**, **Porządkowa** lub **Ilościowa**. Zmienne wybrane jako zmienne podziału nie mogą być używane jako zmienne przewidywane, wejściowe, zmienne dzielące na podzbiory, zmienne częstości ani zmienne ważące. Więcej informacji można znaleźć w temacie “Budowanie modeli rozdzielonych” na stronie 28.
- **Użyj zmiennej częstości.** Ta opcja umożliwia wybranie zmiennej jako wagi częstości. Tej opcji należy użyć, jeśli każdy rekord w danych uczących reprezentuje więcej niż jedną jednostkę — na przykład jeśli stosowane są dane zagregowane. Wartości zmiennych powinny odpowiadać liczbom jednostek reprezentowanych przez poszczególne rekordy. Więcej informacji można znaleźć w temacie “Użycie zmiennych częstości i ważących” na stronie 33.

Uwaga: jeśli pojawi się komunikat o błędzie **Nieprawidłowe metadane (w zmiennych wejściowych/wyjściowych)**, należy się upewnić, że określono wszystkie wymagane zmienne, np. zmienną częstości.

- **Użyj zmiennej ważącej.** Ta opcja umożliwia wybranie zmiennej jako wagi obserwacji. Wagi obserwacji są stosowane w celu uwzględniania różnic w wariancji między poziomami zmiennej wyjściowej. Więcej informacji można znaleźć w temacie “Użycie zmiennych częstości i ważących” na stronie 33.

- **Następniki.** W przypadku węzłów wywołujących reguły (węzłów apriori) należy wybrać zmienne, które będą stosowane jako następniki w wynikowym zestawie reguł. (To odpowiada zmiennym o roli *Zmienna przewidywana* lub *Łącznie* w węźle Typ).
- **Poprzedniki.** W przypadku węzłów wywołujących reguły (węzłów apriori) należy wybrać zmienne, które będą stosowane jako poprzedniki w wynikowym zestawie reguł. (To odpowiada zmiennym o roli *Zmienna wejściowa* lub *Łącznie* w węźle Typ).

Niektóre modele zawierają kartę Zmienne, która różni się od kart opisanych w niniejszej sekcji.

- Więcej informacji można znaleźć w temacie “Opcje zmiennych węzła Sekwencje” na stronie 273.
- Więcej informacji można znaleźć w temacie “Opcje zmiennych węzła CARMA” na stronie 262.

Użycie zmiennych częstości i ważących

Zmienne częstości i ważące pozwalają nadać niektórym rekordom większe znaczenie niż innym, na przykład, jeśli wiadomo, że jedna część populacji jest niedostatecznie reprezentowana w danych uczących (waga) lub ponieważ jeden rekord reprezentuje pewną liczbę identycznych obserwacji (częstość).

- Wartości dla zmiennej częstości powinny być dodatnimi liczbami całkowitymi. Rekordy z wagą częstości, która ma wartość zero lub jest liczbą ujemną, są wykluczane z analizy. Wagi częstości, które nie są liczbami całkowitymi, są zaokrąglane do najbliższej liczby całkowitej.
- Wartości wagi obserwacji powinny być dodatnie, ale nie muszą być wartościami całkowitymi. Rekordy z wagą obserwacji, która ma wartość zero lub jest liczbą ujemną, są wykluczane z analizy.

Ocenianie zmiennych częstości i ważących

Zmienne częstości i ważące są używane w modelach uczących, ale nie są używane do oceniania, ponieważ ocena każdego rekordu jest przeprowadzana na podstawie jego charakterystyki niezależnie od tego, ile obserwacji on reprezentuje. Załóżmy na przykład, że dostępne są dane w następującej tabeli.

Tabela 1. Przykład danych

Zamężna/żonaty	**Odpowiedź**
Tak	Tak
Tak	Tak
Tak	Tak
Tak	Nie
Nie	Tak
Nie	Nie
Nie	Nie

Na podstawie tych informacji można wywnioskować, że trzy z czterech osób zamężnych/żonatych odpowiedziało na promocję, zaś dwie z trzech osób niezamężnych/nieżonatych nie odpowiedziało. Na tej podstawie można ocenić dowolne nowe rekordy, co przedstawia poniższa tabela.

Tabela 2. Przykład ocenionych rekordów

Zamężna/żonaty	**$-Odpowiedź**	**$RP-Odpowiedź**
Tak	Tak	0,75 (trzy/cztery)
Nie	Nie	0,67 (dwa/trzy)

Można również zapisać dane uczące w sposób bardziej kompaktowy, używając zmiennej częstości, co przedstawiono w następującej tabeli.

Tabela 3. Alternatywny przykład ocenionych rekordów

Zamężna/żonaty	Odpowiedź	Częstość
Tak	Tak	3
Tak	Nie	1
Nie	Tak	1
Nie	Nie	2

Ponieważ jest to reprezentacja dokładnie tego samego zbioru danych, zbudowany zostanie taki sam model, a predykcja odpowiedzi będzie przeprowadzona wyłącznie w oparciu o stan cywilny. Jeśli oceniane dane będą dotyczyły dziesięciu osób zamężnych/żonatych, przewidziana zostanie odpowiedź *Tak* dla każdej z nich, niezależnie od tego, czy zaprezentowanych zostanie dziesięć osobnych rekordów czy jeden z wartością częstości wynoszącą 10. Waga, choć ogólnie zwykle nie jest liczbą całkowitą, może być traktowana jako wartość, która podobnie wskazuje ważność rekordu. Dlatego właśnie zmienne częstości i ważące nie są używane podczas oceniania rekordów.

Ocenianie i porównywanie modeli

Niektóre typy modeli obsługują zmienne częstości, inne zmienne ważące, a jeszcze inne obie te zmienne. Jednak we wszystkich przypadkach, w jakich mają zastosowanie, są używane wyłącznie do budowania modelu i nie są brane pod uwagę podczas ewaluacji modeli za pomocą węzła ewaluacji lub analizy, ani podczas rangowania modeli za pomocą większości metod obsługiwanych przez węzły Auto Klasyfikacja i Auto Predykcja.

- Podczas podsumowywania modeli (na przykład za pomocą wykresów ewaluacyjnych) wartości częstości i wagi będą ignorowane. Dzięki temu możliwe jest porównanie poziomów pomiędzy modelami, które korzystają z tych zmiennych oraz modelami, które z nich nie korzystają; oznacza to jednak, że w celu dokładnego oszacowania konieczne jest użycie zbioru danych, który dokładnie reprezentuje populację bez korzystania ze zmiennych częstości lub ważących. W praktyce jest to możliwe po upewnieniu się, że ewaluacja modeli odbywa się przy użyciu próby testu, w której wartość zmiennych częstości i ważących zawsze wynosi null lub 1. (To ograniczenie ma zastosowanie tylko podczas ewaluacji modeli; jeśli wartości częstości lub wagi zawsze wynosiłyby 1 dla prób uczących i testowych, użycie ich w pierwszej kolejności nie byłoby uzasadnione).
- W przypadku użycia opcji Auto Klasyfikacja częstość może być brana pod uwagę, jeśli rangowanie modeli odbywa się w oparciu o zysk, dlatego w takim przypadku ta metoda jest zalecana.
- W razie potrzeby można podzielić dane na próby uczące i testowe, używając węzła Podział.

Opcje analizowania węzła modelowania

Wiele węzłów modelowania zawiera kartę Analiza, która umożliwia uzyskanie informacji o ważności predyktora razem z surowymi i skorygowanymi ocenami skłonności.

Ocena modelu

Oblicz ważność predyktora. W przypadku modeli generujących odpowiednią miarę ważności możliwe jest wyświetlenie tabeli wskazującej ważność względną każdego predyktora w procesie estymacji modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Należy zauważyć, że obliczenie ważności predyktora może potrwać dłużej dla niektórych modeli, szczególnie w przypadku pracy z dużymi zbiorami danych, i domyślnie ta opcja dla niektórych modeli jest wyłączona. Ważność predyktorów jest niedostępna dla modeli listy decyzyjnej. Więcej informacji można znaleźć w “Ważność predyktorów” na stronie 43.

Oceny skłonności

Oceny skłonności można aktywować w węźle modelowania oraz na karcie Ustawienia w modelu użytkowym. Ta funkcja jest dostępna tylko wówczas, gdy wybrana zmienna przewidywana jest zmienną typu flaga. Więcej informacji można znaleźć w temacie “Oceny skłonności” na stronie 35.

Wylicz surowe oceny skłonności. Surowe oceny skłonności są wyznaczane z modelu wyłącznie w oparciu o dane uczące. Jeśli model przewiduje wartość *true* (udzieli odpowiedzi), wówczas skłonność jest taka sama jak P, gdzie P to prawdopodobieństwo predykcji. Jeśli model przewidzi wartość typu false, wówczas skłonność jest obliczana jako (1 – P).

- W przypadku wybrania tej opcji podczas budowania modelu oceny skłonności będą domyślnie aktywowane w modelu użytkowym. Surowe oceny skłonności można jednak aktywować w modelu użytkowym w dowolnym czasie, niezależnie od tego, czy zostały wybrane w węźle modelowania.
- Podczas oceniania modelu surowe oceny skłonności zostaną dodane do zmiennej z literami *RP* dodanymi do standardowego przedrostka. Przykładowo, jeśli predykcje znajdują się w zmiennej o nazwie *$R-churn*, wprowadzona nazwa zmiennej oceny skłonności będzie następująca: *$RRP-churn*.

Wylicz skorygowane oceny skłonności. Surowe skłonności są wyznaczane wyłącznie w oparciu o oszacowania udostępnione przez model, które mogą być nadmiernie dopasowane, co może doprowadzić do zbyt optymistycznych oszacowań skłonności. Skorygowane skłonności spróbują przeprowadzić wyrównanie, sprawdzając, jak model działa w podzbiorze testowym lub walidacyjnym i korygując skłonności, tak aby uzyskać lepsze oszacowanie.

- To ustawienie wymaga, aby w strumieniu obecna była poprawna zmienna dzieląca na podzbiory.
- W przeciwieństwie do surowych ocen ufności skorygowane oceny skłonności muszą być obliczone podczas budowania modelu; w przeciwnym razie nie będą dostępne podczas oceniania modelu użytkowego.
- Podczas oceniania modelu skorygowane oceny skłonności zostaną dodane do zmiennej z literami *AP* dodanymi do standardowego przedrostka. Przykładowo, jeśli predykcje znajdują się w zmiennej o nazwie *$R-churn*, wprowadzona nazwa zmiennej oceny skłonności będzie następująca: *$RAP-churn*. Skorygowane oceny skłonności są niedostępne dla modeli regresji logistycznej.
- Podczas obliczania skorygowanych ocen skłonności podzbiór testowy lub walidacyjny używany do obliczeń nie może być zrównoważony. Aby tego uniknąć, należy sprawdzić, czy opcja **Równoważ tylko dane uczące** jest zaznaczona w którymkolwiek poprzedzającym węźle ważenia. Ponadto, jeśli w poprzedzającej części strumienia przeprowadzona została złożona próba, spowoduje to unieważnienie skorygowanych ocen skłonności.
- Skorygowane oceny skłonności są niedostępne w przypadku modeli drzewa wzmacnianego i zestawu reguł. Więcej informacji można znaleźć w temacie “Wzmacniane modele C5.0” na stronie 124.

Na podstawie. Aby możliwe było obliczenie skorygowanych ocen skłonności, w strumieniu musi znajdować się zmienna dzieląca na podzbiory. Można określić, czy do obliczenia ma być używany podzbiór testowy czy walidacyjny. Aby uzyskać jak najlepsze wyniki, podzbiór testowy lub walidacyjny powinien zawierać co najmniej tyle rekordów, ile podzbiór użyty do uczenia oryginalnego modelu.

Oceny skłonności

W przypadku modeli, które zwracają predykcję *tak* lub *nie* można zażądać ocen skłonności obok standardowych wartości predykcji i ufności. Oceny skłonności wskazują wiarygodność konkretnego wyniku lub odpowiedzi. Poniższa tabela zawiera przykład.

Tabela 4. Oceny skłonności

Klient	Skłonność do odpowiedzi
Joe Smith	35%
Jane Smith	15%

Oceny skłonności są dostępne tylko dla modeli zawierających zmienne przewidywane typu flaga i oznaczają wiarygodność wartości *Prawda* zdefiniowanej dla zmiennej, co jest określone w źródle lub węźle Typ.

Oceny skłonności i oceny ufności

Oceny skłonności różnią się od ocen ufności, które mają zastosowanie do bieżącej predykcji, bez względu na to, czy wynik ma wartość *tak*, czy *nie*. Na przykład w sytuacjach, gdy predykcja ma wartość *nie*, wysoka ufność oznacza

wysoką wiarygodność tego, że odpowiedź *nie* zostanie uzyskana. Oceny skłonności obchodzą to ograniczenie, aby umożliwić łatwiejsze porównywanie wśród wszystkich rekordów. Na przykład skłonność *nie* z ufnością *0,85* oznacza surową skłonność *0,15* (lub *1 minus 0,85*).

Tabela 5. Oceny ufności

Klient	Predykcja	Ufność
Joe Smith	Odpowie	0,35
Jane Smith	Nie odpowie	0,85

Uzyskiwanie ocen skłonności

- Oceny skłonności można włączyć na karcie Analiza w węźle modelowania lub na karcie Ustawienia w modelu użytkowym. Ta funkcja jest dostępna tylko wówczas, gdy wybrana zmienna przewidywana jest zmienną typu flaga. Więcej informacji można znaleźć w temacie “Opcje analizowania węzła modelowania” na stronie 34.
- Oceny skłonności mogą być również obliczane przez węzeł Zespół — w zależności od stosowanej metody zespolenia.

Obliczanie skorygowanych ocen skłonności

Skorygowane oceny skłonności są obliczane jako część procesu budowania modelu i nie są dostępne w innych sytuacjach. Po zbudowaniu modelu następuje jego ocena przy użyciu danych z podzbioru testowego lub walidacyjnego, a następnie tworzony jest nowy model w celu uzyskania skorygowanych ocen skłonności. Ten model jest konstruowany poprzez analizę wydajności pierwotnego modelu w tym podzbiorze. W zależności od typu modelu można użyć jednej lub dwóch metod w celu obliczenia skorygowanych ocen skłonności.

- W przypadku modeli zestawu reguł i modeli drzewa skorygowane oceny skłonności są generowane poprzez ponowne obliczenie częstości każdej kategorii w każdym węźle drzewa (dla modeli drzewa) albo obliczenie pokrycia i ufności każdej reguły (dla modeli zestawu reguł). W wyniku powstaje nowy zestaw reguł lub model drzewa, który jest przechowywany w oryginalnym modelu i używany zawsze, gdy wymagane są skorygowane oceny skłonności. Za każdym razem, gdy oryginalny model jest stosowany względem danych, nowy model może następnie być stosowany względem surowych ocen skłonności w celu wygenerowania ocen skorygowanych.
- W przypadku innych modeli rekordy generowane przez ocenianie oryginalnego modelu w podzbiorze testowym lub walidacyjnym są następnie kategoryzowane na podstawie ich surowej oceny skłonności. Następnie uczony jest model sieci neuronowej, co definiuje nieliniową funkcję, która odwzorowuje ze średniej surowej skłonności w każdej kategorii do średniej obserwowanej skłonności w tej samej kategorii. Zgodnie z tym, co zostało powiedziane wcześniej w odniesieniu do modeli drzew, wynikowy model sieci neuronowej jest przechowywany z modelem oryginalnym i może być stosowany względem surowych ocen skłonności każdorazowo, gdy żądane są skorygowane oceny skłonności.

Przestroga dotycząca brakujących wartości w podzbiorze testowym. Postępowanie w przypadku brakujących wartości wejściowych w podzbiorze testowym/walidacyjnym różni się w zależności od modelu (szczegółowe informacje zawierają poszczególne algorytmy oceniania modeli). Model C5 nie może obliczyć skorygowanych skłonności, gdy brakuje zmiennych wejściowych.

Koszty błędnej klasyfikacji

W niektórych kontekstach pewne błędy są bardziej kosztowne od innych. Przykładowo, bardziej kosztowne może być sklasyfikowanie osób składających wniosek kredytowy z wysokim poziomem ryzyka jako osób z niskim poziomem ryzyka (jeden rodzaj błędu) niż sklasyfikowanie osób składających wniosek z niskim poziomem ryzyka jako osób z wysokim poziomem ryzyka (inny rodzaj błędu). Koszty błędnej klasyfikacji umożliwiają określenie względnej ważności różnych rodzajów błędów predykcji.

Kosztami błędnej klasyfikacji zwykle są wagi zastosowane do określonych danych wynikowych. Wagi te są uwzględniane w modelu i rzeczywiście mogą zmienić predykcję (jako sposób ochrony przed kosztownymi błędami).

Z wyjątkiem modeli C5.0 koszty błędnej klasyfikacji nie mają zastosowania podczas oceniania modelu i nie są brane pod uwagę podczas rangowania lub porównywania modeli za pomocą węzła Auto Klasyfikacja, wykresu ewaluacyjnego lub węzła analizy. Model, który uwzględnia koszty, może nie wygenerować mniejszej liczby błędów niż ten, który ich nie uwzględnia, i może nie mieć wyższej rangi pod względem całkowitej dokładności, ale prawdopodobnie lepiej sprawdzi się w warunkach praktycznych, ponieważ generuje błędy *mniej kosztowne*.

Macierz kosztów przedstawia koszty dla każdej możliwej kombinacji przewidywanej kategorii rzeczywistej. Domyślnie wszystkie koszty błędnej klasyfikacji są ustawione na wartość 1,0. Aby wprowadzić niestandardowe wartości kosztów, należy wybrać opcję **Stosuj koszty błędnej klasyfikacji** i wprowadzić do macierzy kosztów niestandardowe wartości.

Aby zmienić koszt błędnej klasyfikacji, należy zaznaczyć komórkę odpowiadającą odpowiedniej kombinacji wartości przewidywanych i rzeczywistych, usunąć istniejącą zawartość komórki i wprowadzić do niej żądany koszt. Koszty nie są automatycznie symetryczne. Przykładowo, jeśli koszt błędnej klasyfikacji *A* jako *B* zostanie ustawiony na 2,0, koszt błędnej klasyfikacji *B* jako *A* nadal będzie miał domyślną wartość 1,0, chyba że zostanie ona również jawnie zmieniona.

Uwaga: Tylko w przypadku modelu drzewa decyzyjnego możliwe jest określenie kosztów już na etapie budowania modelu.

Modele użytkowe

Rysunek 19. Model użytkowy

Model użytkowy to kontener dla modelu, czyli zbiór reguł, formuł lub równań, który reprezentuje wyniki operacji wykonywanych podczas budowania modelu w programie SPSS Modeler. Głównym przeznaczeniem modelu użytkowego jest dokonywanie oceny danych w celu wygenerowania predykcji lub umożliwienie przeprowadzenia dalszej analizy właściwości modelu. Po otwarciu modelu użytkowego na ekranie widoczne są różne szczegóły dotyczące modelu, takie jak względna ważność zmiennych wejściowych w procesie tworzenia modelu. Aby możliwe było wyświetlenie predykcji, konieczne jest włączenie i wykonanie dalszych procesów lub węzła wynikowego. Więcej informacji można znaleźć w temacie “Używanie modeli użytkowych w strumieniach” na stronie 48.

Rysunek 20. Łącze modelu z węzła modelowania do modelu użytkowego

Po uruchomieniu węzła modelowania na obszarze roboczym strumienia umieszczany jest odpowiedni model użytkowy. Jest on oznaczany za pomocą złotej ikony w kształcie rombu. W obszarze roboczym strumienia model użytkowy jest wyświetlany wraz z połączeniem (linia ciągła) z najbliższym odpowiednim węzłem znajdującym się przed węzłem modelowania oraz z łączem (linia przerywana) z samym węzłem modelowania.

Model użytkowy jest również umieszczany na palecie modeli w prawym górnym rogu okna IBM SPSS Modeler. Z dowolnej lokalizacji modele użytkowe mogą być wybierane i przeglądane w celu wyświetlenia szczegółów modelu.

Modele użytkowe zawsze umieszczane są na palecie modeli, jeśli węzeł modelowania zostanie poprawnie wykonany. Użytkownik może ustawić własne opcje, które określą, kiedy model użytkowy będzie dodatkowo umieszczony w obszarze roboczym strumienia.

Przedstawione poniżej tematy udostępniają informacje dotyczące korzystania z modeli użytkowych w programie IBM SPSS Modeler. Aby dokładnie zrozumieć zastosowane algorytmy, należy zapoznać się z publikacją *IBM SPSS Modeler Algorithms Guide*, dostępną w postaci pliku PDF wraz z pobieranym produktem.

Łącza modelu

Domyślnie model użytkowy jest przedstawiany w obszarze roboczym z łączem do tworzonego węzła modelowania. Jest to szczególnie użyteczne w przypadku złożonych strumieni z kilkoma modelami użytkowymi i umożliwia zidentyfikowanie modelu użytkowego, który będzie aktualizowany przez poszczególne węzły modelowania. Każde łącze zawiera symbol, który wskazuje, czy model jest zastępowany w momencie wykonania węzła modelowania. Więcej informacji można znaleźć w temacie “Zastępowanie modelu” na stronie 39.

Definiowanie i usuwanie łącz modelu

Łącza modelu można definiować i usuwać ręcznie w obszarze roboczym. Podczas definiowania nowego łącza kursor przyjmuje kształt kursora łącza.

Rysunek 21. Kursor łącza

Definiowanie nowego łącza (menu kontekstowe)

1. Kliknij prawym przyciskiem myszy węzeł modelowania, z którego chcesz rozpocząć łącze.
2. Z menu kontekstowego wybierz opcję **Definiuj łącze modelu**.
3. Kliknij model użytkowy, w którym łącze ma się zakończyć.

Definiowanie nowego łącza (menu główne)

1. Kliknij węzeł modelowania, z którego chcesz rozpocząć łącze.
2. Z menu głównego wybierz opcje:
   Edycja > Węzeł > Definiuj łącze modelu
3. Kliknij model użytkowy, w którym łącze ma się zakończyć.

Usuwanie istniejącego łącza (menu kontekstowe)

1. Kliknij prawym przyciskiem myszy model użytkowy na końcu łącza.
2. Z menu kontekstowego wybierz opcję **Usuń łącze modelu**.

Lub:

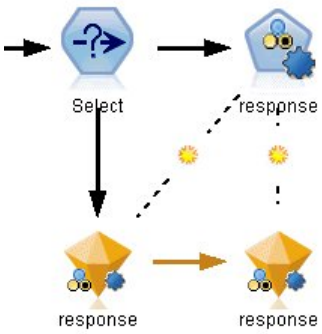
1. Kliknij prawym przyciskiem myszy symbol w środku łącza.
2. Z menu kontekstowego wybierz opcję **Usuń łącze**.

Usuwanie istniejącego łącza (menu główne)

1. Kliknij węzeł modelowania lub model użytkowy, z którego chcesz usunąć łącze.
2. Z menu głównego wybierz opcje:
   Edycja > Węzeł > Usuń łącze modelu

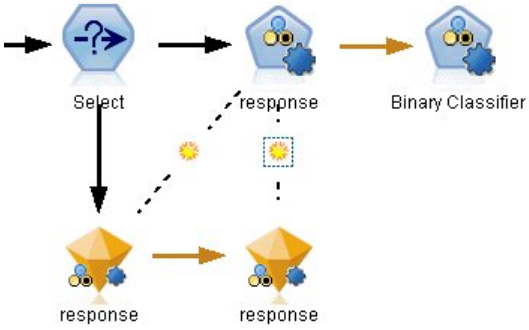
Kopiowanie i wklejanie łącz modelu

Jeśli skopiujesz połączony model użytkowy bez jego węzła modelowania, a następnie wkleisz go do tego samego strumienia, wówczas model użytkowy zostanie wklejony z łączem do węzła modelowania. Nowe łącze ma ten sam status zastępowania (patrz “Zastępowanie modelu” na stronie 39), co łącze oryginalne.



Rysunek 22. Kopiowanie i wklejanie połączonego modelu użytkowego

Jeśli skopiujesz i wkleisz model użytkowy razem z połączonym z nim węzłem modelowania, wówczas połączenie zostanie zachowane bez względu na to, czy obiekty są wklejane do tego samego, czy do nowego strumienia.



Rysunek 23. Kopiowanie i wklejanie połączonego modelu użytkowego

Uwaga: jeśli skopiujesz połączony model użytkowy bez jego węzła modelowania, a następnie wkleisz model użytkowy do nowego strumienia (albo do superwęzła, który nie zawiera węzła modelowania), wówczas łącze zostanie przerwane i wklejony zostanie tylko model użytkowy.

Łącza modelu i superwęzły

Jeśli zdefiniujesz superwęzeł w taki sposób, aby zawierał węzeł modelowania lub model użytkowy połączonego modelu (ale nie obydwa te elementy), wówczas łącze zostanie zerwane. Rozwinięcie superwęzła nie spowoduje przywrócenia łącza; jest to możliwe tylko poprzez cofnięcie tworzenia superwęzła.

Zastępowanie modelu

W przypadku ponownego wykonania węzła modelowania, który utworzył model użytkowy, można wybrać, czy istniejący model użytkowy zostanie zastąpiony (czyli zaktualizowany). Jeśli opcja zastępowania zostanie wyłączona, podczas ponownego wykonywania węzła modelowania zostanie utworzony nowy model użytkowy.

Każde łącze z węzła modelowania do modelu użytkowego zawiera symbol, który wskazuje, czy model jest zastępowany w przypadku ponownego wykonywania węzła modelowania.

Rysunek 24. Łącze modelu z włączonym zastępowaniem modelu

Gdy zastępowanie modelu jest włączone, łącze jest początkowo pokazane i przedstawione jako symbol niewielkiego słoneczka w łączu. W tym stanie ponowne wykonanie węzła modelowania na jednym końcu łącza powoduje po prostu aktualizację modelu użytkowego na drugim końcu.

Rysunek 25. Łącze modelu z wyłączonym zastępowaniem modelu

Jeśli zastępowanie modelu jest wyłączone, symbol łącza jest zastąpiony szarą kropką. W takim stanie ponowne wykonanie węzła modelowania na jednym końcu łącza spowoduje dodanie nowej, zaktualizowanej wersji modelu użytkowego do obszaru roboczego.

W każdym z tych przypadków w palecie Modele aktualizowany jest istniejący model użytkowy lub dodawany jest nowy model użytkowy — w zależności od ustawienia opcji systemowej **Zastąp poprzedni model**.

Kolejność wykonywania

W przypadku wykonywania strumienia z wieloma gałęziami zawierającymi modele użytkowe najpierw następuje ocena strumienia w celu zapewnienia, że gałąź, w której włączona jest opcja zastępowania modelu jest wykonywana przed innymi gałęziami, które korzystają z wynikowego modelu użytkowego.

Jeśli wymagania w przypadku konkretnej sytuacji są bardziej złożone, można ustawić kolejność wykonywania ręcznie, stosując skrypty.

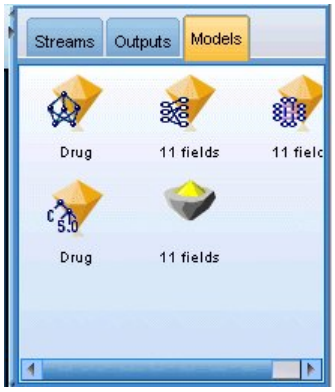
Zmiana ustawienia zastępowania modelu

1. Kliknij prawym przyciskiem myszy symbol łącza.
2. W razie potrzeby wybierz opcję **Wyłącz podmianę modelu**.

Uwaga: Ustawienie zastępowania modelu w łączu modelu zastępuje ustawienie na karcie Powiadomienia w oknie dialogowym Opcje użytkownika (Narzędzia > Opcje > Opcje użytkownika).

Paleta modeli

Paleta modeli (na karcie Modele w oknie menedżerów) umożliwia używanie, analizowanie i modyfikowanie modeli użytkowych w różny sposób.



Rysunek 26. Paleta modeli

Kliknięcie modelu użytkowego w palecie modeli prawym przyciskiem myszy spowoduje otwarcie menu kontekstowego zawierającego następujące opcje:

- **Dodaj do strumienia.** Dodaje model użytkowy do aktualnie aktywnego strumienia. Jeśli w strumieniu zaznaczono węzeł, model użytkowy zostanie połączony z tym węzłem, o ile takie połączenie jest możliwe, lub w przeciwnym razie z najbliższym możliwym węzłem. Model użytkowy jest wyświetlany wraz z łączem do węzła modelowania, który utworzył model, o ile ten węzeł nadal znajduje się w strumieniu.
- **Przeglądaj.** Otwiera przeglądarkę modelu dla modelu użytkowego.
- **Zmień nazwę i skomentuj.** Umożliwia zmianę nazwy modelu użytkowego i/lub zmodyfikowanie komentarzy do modelu.
- **Utwórz węzeł modelowania.** Jeśli istnieje model użytkowy, który ma zostać zmodyfikowany lub aktualizowany, a strumień użyty do utworzenia modelu jest niedostępny, można użyć tej opcji, aby ponownie utworzyć węzeł modelowania z zastosowaniem takich samych opcji, jakie zostały użyte do utworzenia oryginalnego modelu.
- **Zapisz model, Zapisz model jako.** Umożliwia zapisanie modelu użytkowego w zewnętrznie wygenerowanym pliku binarnym (.gm) modelu.
- **Składuj model.** Umożliwia zapisanie modelu użytkowego w repozytorium IBM SPSS Collaboration and Deployment Services Repository.
- **Eksportuj PMML.** Umożliwia wyeksportowanie modelu użytkowego w formacie PMML (Predictive Model Markup Language), który może być użyty do oceniania nowych danych poza programem IBM SPSS Modeler. Opcja **Eksportuj PMML** jest dostępna dla wszystkich wygenerowanych węzłów modelu.
- **Dodaj do projektu.** Zapisuje model użytkowy i dodaje go do bieżącego projektu. Na karcie Klasy model użytkowy zostanie dodany do folderu Wygenerowane modele. Na karcie CRISP-DM zostanie on dodany do domyślnej fazy projektu.
- **Usuń.** Usuwa model użytkowy z palety.

Kliknięcie pustego obszaru w palecie modeli prawym przyciskiem myszy spowoduje otwarcie menu kontekstowego zawierającego następujące opcje:

- **Otwórz model.** Ładuje model użytkowy wcześniej utworzony w programie IBM SPSS Modeler.
- **Pobierz model.** Pobiera zapisany model z repozytorium IBM SPSS Collaboration and Deployment Services.
- **Załaduj paletę.** Ładuje zapisaną paletę modeli z pliku zewnętrznego.
- **Pobierz paletę.** Pobiera zapisaną paletę modeli z repozytorium IBM SPSS Collaboration and Deployment Services.
- **Zapisz paletę.** Zapisuje całą zawartość palety modeli w zewnętrznie wygenerowanym pliku (.gen) palety modeli.
- **Składuj paletę.** Składuje całą zawartość palety modeli w repozytorium IBM SPSS Collaboration and Deployment Services.
- **Wyczyść paletę.** Usuwa wszystkie modele użytkowe z palety.
- **Dodaj paletę do projektu.** Zapisuje paletę modeli i dodaje ją do bieżącego projektu. Na karcie Klasy model użytkowy zostanie dodany do folderu Wygenerowane modele. Na karcie CRISP-DM zostanie on dodany do domyślnej fazy projektu.

- **Importuj PMML.** Ładuje model z pliku zewnętrznego. Można otworzyć, przeglądać i oceniać modele PMML utworzone w programie IBM SPSS Statistics lub za pomocą innych aplikacji obsługujących ten format. Więcej informacji można znaleźć w temacie “Importowanie i eksportowanie modeli w formacie PMML” na stronie 49.

Przeglądanie modeli użytkowych

Przeglądarki modeli użytkowych umożliwiają analizowanie i korzystanie z wyników modeli. Za pośrednictwem przeglądarki można zapisywać, drukować lub eksportować wygenerowany model, analizować podsumowanie modelu i wyświetlać lub edytować komentarze do modelu. W przypadku niektórych typów modeli użytkowych można również wygenerować nowe węzły, takie jak węzeł Filtr lub węzeł Zestaw reguł. W przypadku pewnych modeli można również wyświetlać ich parametry, takie jak reguły lub centra skupienia. Dla niektórych typów modeli (modele oparte na drzewie i modele skupień) możliwe jest wyświetlanie graficznej reprezentacji struktury modelu. Poniżej opisano elementy sterujące służące do korzystania z przeglądarek modeli użytkowych.

Menu

Menu Plik. Wszystkie modele użytkowe mają menu Plik, które zawiera dodatkowy zestaw następujących opcji:

- **Zapisz węzeł.** Zapisuje model użytkowy w pliku węzła (.nod).
- **Składuj węzeł.** Umożliwia zapisanie modelu użytkowego w repozytorium IBM SPSS Collaboration and Deployment Services.
- **Nagłówek i stopka.** Umożliwia edytowanie nagłówka i stopki strony do wydrukowania z modelu użytkowego.
- **Ustawienia strony.** Umożliwia zmianę ustawień strony do wydrukowania z modelu użytkowego.
- **Podgląd wydruku.** Wyświetla, w jaki sposób węzeł użytkowy będzie wyglądał po wydrukowaniu. Informacje, jakie mają się znaleźć na podglądzie, można wybrać z menu podrzędnego.
- **Drukuj.** Umożliwia wydrukowanie zawartości modelu użytkowego. Informacje, jakie mają się znaleźć na wydruku, można wybrać z menu podrzędnego.
- **Drukuj widok.** Umożliwia wydrukowanie bieżącego widoku lub wszystkich widoków.
- **Eksportuj do pliku tekstowego.** Umożliwia wyeksportowanie zawartości modelu użytkowego do pliku tekstowego. Informacje, jakie mają zostać wyeksportowane, można wybrać z menu podrzędnego.
- **Eksportuj do HTML.** Umożliwia wyeksportowanie zawartości modelu użytkowego do pliku HTML. Informacje, jakie mają zostać wyeksportowane, można wybrać z menu podrzędnego.
- **Eksportuj PMML.** Umożliwia wyeksportowanie modelu w formacie PMML (Predictive Model Markup Language), który może zostać użyty z innym oprogramowaniem obsługującym format PMML. Więcej informacji można znaleźć w temacie “Importowanie i eksportowanie modeli w formacie PMML” na stronie 49.
- **Eksportuj SQL.** Umożliwia wyeksportowanie modelu w formacie SQL (Structured Query Language), który może być edytowany i używany za pomocą innych baz danych.

  Uwaga: Opcja Eksportuj SQL jest dostępna tylko dla następujących modeli: C5, C&RT, CHAID, QUEST, Regresja liniowa, Regresja logistyczna, Sieci neuronowe, analizy PCA/czynnikowej i Lista decyzyjna.
- **Publikacja dla Server Scoring Adapter.** Umożliwia opublikowanie modelu w bazie danych, w której zainstalowano składnik Scoring Adapter, który umożliwia ocenianie modelu w bazie danych. Więcej informacji można znaleźć w temacie “Publikowanie modeli dla adaptera oceniania” na stronie 51.

Menu Utwórz. Większość modeli użytkowych ma również menu Utwórz, które umożliwia wygenerowanie nowych węzłów na podstawie modelu użytkowego. Opcje dostępne w tym menu będą zależały od typu przeglądanego modelu. Szczegółowe informacje o tym, co można wygenerować z poszczególnych modeli, można znaleźć w konkretnym typie modelu użytkowego.

Menu Widok. Na karcie Model modelu użytkowego to menu umożliwia wyświetlenie lub ukrycie różnych pasków narzędzi wizualizacji, jakie są dostępne w bieżącym trybie. Aby udostępnić pełny zestaw pasków narzędzi, należy wybrać opcję Tryb edycji (ikona pędzla) z paska narzędzi opcji ogólnych.

Przycisk Podgląd. Niektóre modele użytkowe mają przycisk Podgląd, który umożliwia wyświetlenie przykładowych danych modelu, z uwzględnieniem zmiennych dodatkowych utworzonych w procesie modelowania. Domyślnie wyświetlanych jest 10 wierszy; ustawienie to można jednak zmienić we właściwościach strumienia.

Przycisk Dodaj do bieżącego projektu. Zapisuje model użytkowy i dodaje go do bieżącego projektu. Na karcie Klasy model użytkowy zostanie dodany do folderu Wygenerowane modele. Na karcie CRISP-DM zostanie on dodany do domyślnej fazy projektu.

Podsumowanie modelu użytkowego/informacje

Karta Podsumowanie lub widok Informacje dla modelu użytkowego umożliwiają wyświetlenie informacji na temat zmiennych, ustawień budowania i procesu estymacji modelu. Wyniki są prezentowane w widoku drzewa, które może być rozwijane i zwijane poprzez kliknięcie konkretnych elementów.

Analiza. Wyświetla informacje na temat modelu. Konkretne szczegóły różnią się w zależności od typu modelu i są zawarte w sekcjach dotyczących poszczególnych modeli użytkowych. Ponadto jeśli wykonano węzeł analizy dołączony do tego węzła modelowania, informacje z tej analizy również będą wyświetlane w tej sekcji.

Pola. Na liście znajdują się zmienne użyte jako zmienne przewidywane i wejściowe podczas budowania modelu. W przypadku modeli rozdzielonych wyświetlana jest również lista zmiennych określających podziały.

Uwaga: W widoku Informacje modelu sieci neuronowych, modelu liniowego oraz innych modeli w trybach wartości logicznej lub agregacji bootstrap wyświetlana ikona jest taka sama (nominalna), niezależnie od typu: flaga, nominalny, czy porządkowy.

Ustawienia/Opcje budowania. Zawiera informacje na temat ustawień użytych podczas budowania modelu.

Podsumowanie uczenia. Przedstawia typ modelu, strumień użyty do jego utworzenia, użytkownika, który go utworzył, informację, kiedy został utworzony oraz czas, jaki był potrzebny do zbudowania modelu. Należy pamiętać, że czas, jaki był potrzebny do zbudowania modelu, jest dostępny tylko na karcie Podsumowanie, a nie jest dostępny w widoku Informacje.

Ważność predyktorów

Zazwyczaj działania modelujące mają koncentrować się na zmiennych predyktorów, które są najważniejsze, a opuszczane lub ignorowane mają być te zmienne, które są najmniej ważne. Wykres ważności predyktorów pomaga osiągnąć ten cel przez wskazanie względnej ważności każdego predyktora przy szacowaniu modelu. Ponieważ wartości są względne, suma wartości wszystkich wyświetlanych predyktorów wynosi 1,0. Ważność predyktora nie jest powiązana z dokładnością modelu. Jest powiązana z ważnością każdego predyktora przy prognozach, a nie z tym, czy taka prognoza jest dokładna.

Ważność predyktora jest dostępna dla modeli, które generują odpowiednią miarę statystyczną ważności — są to między innymi sieci neuronowe, drzewa decyzyjne (C&RT, C5.0, CHAID oraz QUEST), sieci bayesowskie, funkcje dyskryminacyjne, SVM oraz modele SLRM, a także regresja liniowa i logistyczna, uogólnione modele liniowe oraz modele najbliższego sąsiedztwa (KNN). W przypadku większości tych modeli ważność predyktora można włączyć na karcie Analiza w węźle modelowania. Więcej informacji można znaleźć w temacie “Opcje analizowania węzła modelowania” na stronie 34. Informacje na temat modeli, patrz “Sąsiedzi” na stronie 345.

Uwaga: Ważność predyktora nie jest obsługiwana dla modeli rozdzielonych. Ustawienia ważności predyktora są ignorowane podczas budowania modeli rozdzielonych. Więcej informacji można znaleźć w temacie “Budowanie modeli rozdzielonych” na stronie 28.

Obliczanie ważności predyktorów może trwać znacznie dłużej niż budowanie modelu, szczególnie w sytuacji, gdy stosowane są duże zbiory danych. Dłużej trwa obliczanie modelu SVM i regresji logistycznej niż obliczanie innych

modeli i dlatego domyślnie jest wyłączone dla tych modeli. Jeśli używany jest zbiór danych z dużą liczbą predyktorów, wówczas szybsze uzyskanie wyników jest możliwe w przypadku użycia węzła Dobór predyktorów (patrz poniżej).

- Ważność predyktorów jest obliczana z podzbioru testowego, jeśli jest dostępny. W przeciwnym wypadku używane są dane uczące.
- W przypadku modeli SLRM ważność predyktorów jest dostępna, ale jest obliczana przez algorytm SLRM. Więcej informacji można znaleźć w temacie “Modele użytkowe SLRM” na stronie 331.
- Za pomocą narzędzi produktu IBM SPSS Modeler przeznaczonych do tworzenia wykresów można korzystać z wykresów, edytować je i zapisywać.
- Opcjonalnie można wygenerować węzeł Filtr na podstawie informacji z wykresu ważności predyktorów. Więcej informacji można znaleźć w temacie “Filtrowanie zmiennych na podstawie ważności”.

Ważność predyktorów i dobór predyktorów

Może się wydawać, że w niektórych przypadkach wykres ważność predyktora wyświetlany w modelu użytkowym przedstawia wyniki podobne do węzła Dobór predyktorów. Dobór predyktorów ocenia każdą zmienną wejściową na podstawie siły jej relacji z określoną zmienną przewidywaną, a niezależne od innych zmiennych wejściowych wykres ważności predyktora przedstawia względną ważność każdej zmiennej wejściowej dla *tego* konkretnego modelu. Dlatego dobór predyktorów jest bardziej konserwatywny w ocenie zmiennych wejściowych. Na przykład, jeśli *nazwa stanowiska* i *kategoria zadania* tworzą silne powiązanie z wynagrodzeniem, wówczas dobór predyktorów wskaże, że obie te zmienne są ważne. Jednak podczas modelowania uwzględniane są również interakcje i korelacje. Zatem może się okazać, że używana jest tylko jedna zmienna wejściowa z dwóch, jeśli obie powielają znaczną część informacji. W praktyce dobór predyktorów jest najbardziej użyteczny w przypadku początkowego monitorowania, szczególnie wówczas, gdy istnieją duże zbiory danych z dużymi ilościami zmiennych, a ważność predyktorów jest bardziej użyteczna niż precyzyjne dostosowywanie modelu.

Różnice między pojedynczymi modelami a węzłami zautomatyzowanego modelowania w kontekście ważności predyktorów

W ważności predyktorów mogą pojawić się różnice i jest to zależne od tego, czy pojedynczy model jest tworzony z osobnego węzła, czy w celu uzyskania wyników używany jest węzeł zautomatyzowanego modelowania. Takie różnice w implementacji są wywołane ograniczeniami stosowanej technologii.

Na przykład w przypadku pojedynczych klasyfikatorów, takich jak CHAID, obliczenie stosuje regułę zatrzymującą i używa wartości prawdopodobieństwa podczas obliczania wartości ważności. Z kolei automatyczny klasyfikator nie stosuje reguły zatrzymującej i używa etykiet przewidywanych bezpośrednio w obliczeniu. Te różnice mogą oznaczać, że w przypadku wygenerowania pojedynczego modelu z użyciem klasyfikatora automatycznego wartość ważności będzie traktowana jako oszacowanie zgrubne — inaczej niż w przypadku obliczenia dla pojedynczego klasyfikatora. W celu uzyskania najbardziej dokładnych wartości ważności predyktorów sugerujemy użycie pojedynczego węzła zamiast węzłów z modelowaniem zautomatyzowanym.

Filtrowanie zmiennych na podstawie ważności

Opcjonalnie można wygenerować węzeł Filtr na podstawie informacji z wykresu ważności predyktorów.

W razie potrzeby zaznacz predyktory, które zamierzasz uwzględnić w wykresie, a następnie wybierz w menu następujące opcje:

Utwórz > Węzeł filtrowania (ważność predyktora)

LUB

> Węzeł filtrowania (ważność predyktora)

Maksymalna liczba zmiennych. Dołącza lub usuwa najważniejsze predyktory do osiągnięcia maksymalnej liczby zmiennych.

Ważność wyższa niż. Dołącza lub usuwa wszystkie predyktory o względnej ważności większej niż określona wartość.

Przeglądarka zespołów

Modele dla zestawów

Model do zestawu dostarcza informacje o modelach składowych w zestawie oraz o wydajności zestawu jako całości.

Główny (niezależny od widoku) pasek narzędzi pozwala na wybranie zestawu lub modelu odniesienia do oceniania. Jeśli do oceniania wybierze się zestaw, można również wybrać regułę łączenia. Zmiany te nie wymagają ponownego wykonania modelu; jednak wybory te zapisują się w (wartościowych informacjach) modelu w celu oceny i/lub podrzędnej ewaluacji modelu. Mają one również wpływ na PMML, wyeksportowane z przeglądarki zespołów.

Reguła łączenia. Podczas wybierania zestawu jest to reguła służąca do łączenia przewidywanych wartości z modeli podstawowych w celu wyliczenia wartości oceny zestawu.

- Przewidywane wartości zestawu dla przewidywanych zmiennych **jakościowych** mogą być połączone przy pomocy głosowania, największego prawdopodobieństwa lub największego, średniego prawdopodobieństwa. **Głosowanie** wybiera kategorię, która ma największe prawdopodobieństwo, najczęściej wśród modeli podstawowych. **Największe prawdopodobieństwo** wybiera kategorię, która uzyskuje największe, pojedyncze prawdopodobieństwo wśród modeli podstawowych. **Największe średnie prawdopodobieństwo** wybiera kategorię z najwyższą wartością, gdy prawdopodobieństwa kategorii wśród modeli podstawowych są uśrednione.
- Przewidywane wartości zestawu dla przewidywanych zmiennych **ilościowych** mogą być połączone przy pomocy średniej lub mediany przewidywanych wartości z modeli podstawowych.

Wartość domyślna jest pobierana ze specyfikacji wykonanych podczas tworzenia modelu. Zmiana reguły łączenia powoduje przeliczenie dokładności modelu i aktualizację wszystkich widoków dokładności modelu. Wykres ważności predyktora jest również aktualizowany. Ten element sterujący jest wyłączony, jeśli do oceny zostanie wybrany model odniesienia.

Pokaż wszystkie łączące reguły. Po wybraniu tej opcji, w tabeli jakościowej modeli pokazane są wyniki dostępnych reguł łączenia. Aktualizowany jest również Wykres dokładności modeli składowych tak, aby wyświetlał linie odniesienia dla każdej metody głosowania.

Podsumowanie modelu: Widok Podsumowanie modelu to szybkie podsumowanie jakości i różnorodności zestawu.

Jakość. Wykres ten przedstawia dokładność modelu finalnego w porównaniu z modelem odniesienia i z modelem naturalnym. Dokładność jest przedstawiona w większym i lepszym formacie; „najlepszy model” będzie miał największą dokładność. Dla jakościowej zmiennej docelowej dokładność jest zwykłą wartością procentową rekordów, dla których przewidywana wartość jest zgodna z zaobserwowaną wartością. Dla jakościowej zmiennej docelowej dokładność wynosi 1 minus iloraz średniego błędu względnego w predykcji (średnia wartości bezwzględnych przewidywanych wartości minus zaobserwowane wartości) i zakres przewidywanych wartości (maksymalna przewidywana wartość minus minimalna przewidywana wartość).

Dla spakowanych zestawów, model odniesienia jest standardowym modelem utworzonym w całym podziale szkoleniowym. Dla wzmocnionych zestawów, model odniesienia jest pierwszym modelem składowym.

Model naturalny przedstawia dokładność, jeśli żaden model nie został zbudowany i przypisuje wszystkie rekordy do kategorii modalnej. Dla jakościowej zmiennej przewidywanej model naturalny nie jest obliczany.

Różnorodność. Wykres ten przedstawia „różnorodność opinii” wśród modeli składowych, używanych do tworzenia zestawu, przedstawioną w większym i bardziej różnorodnym formacie. Jest on miarą tego, jak bardzo predykcje różnią się w ramach modeli podstawowych. Różnorodność jest niedostępna dla modeli zespolonych wzmocnionych, ani nie jest pokazana dla jakościowych zmiennych przewidywanych.

Ważność predyktorów: Zazwyczaj działania modelujące mają koncentrować się na zmiennych predyktorów, które są najważniejsze, a opuszczane lub ignorowane mają być te zmienne, które są najmniej ważne. Wykres ważności

predyktorów pomaga osiągnąć ten cel przez wskazanie względnej ważności każdego predyktora przy szacowaniu modelu. Ponieważ wartości są względne, suma wartości wszystkich wyświetlanych predyktorów wynosi 1,0. Ważność predyktora nie jest powiązana z dokładnością modelu. Jest powiązana z ważnością każdego predyktora przy prognozach, a nie z tym, czy taka prognoza jest dokładna.

Ważność predyktorów jest niedostępna dla wszystkich modeli zespolonych. Zestaw predyktorów może się różnić w poszczególnych modelach składowych, ale ważność można przeliczyć dla predyktorów używanych w co najmniej jednym modelu składowym.

Częstotliwość predyktorów: Zestaw predyktorów może się różnić w poszczególnych modelach składowych z uwagi na wybór metody modelowania lub wybór predyktora. Wykres częstotliwości predyktorów jest wykresem punktowym, przedstawiającym rozkład predyktorów w poszczególnych modelach składowych w zestawie. Każdy punkt przedstawia jeden lub więcej modeli składowych, zawierających predyktor. Predyktory są wykreślone na osi y i są posortowane w kolejności malejącej według częstotliwości; wskutek tego predyktor znajdujący się najwyżej jest tym, który jest używany w największej liczbie modeli składowych, a predyktor znajdujący się najniżej jest tym, który został użyty najmniejszą ilość razy. Wyświetlonych jest 10 najwyższych predyktorów.

Predyktory pojawiające się najczęściej są zwykle najistotniejsze. Ten wykres jest nieprzydatny dla metod, w których zestaw predyktorów nie może się różnić w poszczególnych modelach składowych.

Dokładność modeli składowych: Jest to wykres punktowy dokładności przewidywania dla modeli składowych. Każdy punkt przedstawia jeden lub więcej modeli składowych, przy czym poziom dokładności jest wykreślony na osi y. Najedź na dowolny punkt, aby uzyskać informacje dotyczące odpowiedniego, indywidualnego modelu składowego.

Linie odniesienia. Wykres przedstawia oznaczone kolorami linie zestawów oraz model odniesienia i model naturalny. Obok linii odpowiadającej modelowi, który zostanie wykorzystany do oceniania pojawia się znacznik.

Interaktywność. Wykres aktualizuje się po zmianie reguły łączenia.

Wzmocnione zestawy. Dla wzmocnionych zestawów wyświetla się wykres liniowy.

Szczegóły dotyczące modeli składowych: Tabela pokazuje informacje dotyczące modeli składowych, zestawionych według wierszy. Domyślnie modele składowe są posortowane rosnąco według kolejności numerów modeli. Można posortować wiersze w kolejności rosnącej lub malejącej, według wartości w dowolnej kolumnie.

Model. Liczba przedstawiająca kolejność utworzenia modeli składowych.

Dokładność. Całkowita dokładność w formacie procentowym.

Metoda. Metoda modelowania.

Predyktory. Liczba predyktorów użytych w modelu składowym.

Rozmiar modelu. Rozmiar modelu zależy od metody modelowania. Dla drzew jest to liczba węzłów w drzewie; dla modeli liniowych jest to liczba współczynników; dla sieci neuronowych jest to liczba synaps.

Rekordy. Ważona liczba rekordów wejściowych w próbie szkoleniowej.

Automatyczne przygotowanie danych:

Widok ten przedstawia informacje o tym, które zmienne zostały wyłączone i w jaki sposób przekształcone zmienne zostały uwzględnione w kroku automatycznego przygotowania danych (ADP). Dla każdej zmiennej, która została przekształcona lub wyłączona, tabela zawiera nazwę zmiennej, jej rolę w analizie i działanie podjęte przez krok ADP. Zmienne są posortowane alfabetycznie, w kolejności rosnącej, według nazw zmiennych.

Działanie **Obcięcie wartości odstających**, jeśli się wyświetli, wskazuje, że wartości predyktorów ciągłych, które znajdują się poza wartością odcięcia (3 standardowe odchylenia od średniej) zostały ustawione na wartość odcięcia.

Modele użytkowe dla modeli rozdzielonych

Model użytkowy dla modelu rozdzielonego zapewnia dostęp do wszystkich osobnych modeli tworzonych w wyniku rozdzielania.

Model użytkowy dla modelu rozdzielonego zawiera:

- listę wszystkich utworzonych modeli rozdzielonych razem z zestawem statystyk na temat każdego modelu
- informacje o modelu ogólnym

Z listy modeli rozdzielonych można otwierać poszczególne modele, a następnie je analizować.

Przeglądarka podzielonych modeli

Karta Model zawiera wszystkie modele zawarte w modelu użytkowym oraz statystyki w różnych formach, dotyczące modeli rozdzielonych. Przyjmuje dwie formy ogólne — w zależności od wykonanych etapów modelowania.

Sortuj według. Za pomocą tej listy można wybrać kolejność, w jakiej wyświetlane będą modele. Listę można posortować na podstawie wartości zawartych w dowolnych wyświetlanych kolumnach, w kolejności rosnącej lub malejącej. Alternatywnie należy kliknąć nagłówek kolumny, aby posortować listę według tej kolumny. Domyślnie wartości są posortowane w kolejności malejącej według ogólnej dokładności.

Pokaż/Ukryj kolumny. Kliknij ten przycisk, aby wyświetlić menu, z którego można wybrać poszczególne kolumny do pokazania lub ukrycia.

Widok. Jeśli używane jest dzielenie na podzbiory, można wybrać wyświetlanie wyników dla danych uczących lub danych testujących.

W przypadku każdego podziału na podzbiory szczegóły są przedstawiane w następujący sposób:

Wykres. Miniatura wskazująca rozkład danych dla danego modelu. Jeśli model użytkowy znajduje się w obszarze roboczym, kliknij dwukrotnie miniaturę, aby wyświetlić wykres w pełnym rozmiarze.

Model. Ikona typu modelu. Kliknij dwukrotnie ikonę, aby otworzyć model użytkowy dla konkretnego rozdziału.

Zmienne podziału. Zmienne wyznaczone w węźle modelowania jako zmienne podziału z różnymi wartościami, jakie mogą przyjmować.

Rekordy w podzbiorze. Liczba rekordów należących do konkretnego podzbioru.

Liczba użytych zmiennych. Nadaje rangi modelom rozdzielonym na podstawie liczby używanych zmiennych wejściowych.

Ogólna dokładność (%). Procent rekordów, które są poprawnie przewidywane przez model rozdzielony względem łącznej liczby rekordów w tym rozdziale.

Podział. Nagłówek kolumny przedstawia zmienne wykorzystane do utworzenia podziałów, a komórki są wartościami podziału. Kliknij dwukrotnie na dowolny podział, aby otworzyć przeglądarkę modelu dla modelu utworzonego dla tego podziału.

Dokładność. Całkowita dokładność w formacie procentowym.

Rozmiar modelu. Rozmiar modelu zależy od metody modelowania. Dla drzew jest to liczba węzłów w drzewie; dla modeli liniowych jest to liczba współczynników; dla sieci neuronowych jest to liczba synaps.

Rekordy. Ważona liczba rekordów wejściowych w próbie szkoleniowej.

Używanie modeli użytkowych w strumieniach

Modele użytkowe są umieszczane w strumieniach, aby umożliwić dokonanie oceny nowych danych i wygenerowanie nowych węzłów. **Ocenianie** danych pozwala na użycie informacji uzyskanych podczas budowania modelu w celu utworzenia predykcji dla nowych rekordów. Aby wyświetlić wyniki oceny, konieczne jest dołączenie węzła końcowego (czyli przetwarzającego lub wynikowego) do modelu użytkowego i wykonanie tego węzła.

W niektórych modelach modele użytkowe mogą również udostępnić dodatkowe informacje na temat jakości predykcji, takie jak współczynniki ufności lub odległości od centrum skupienia. Generowanie nowych węzłów pozwala na łatwe utworzenie nowych węzłów w oparciu o strukturę generowanego modelu. Na przykład, większość modeli, które dokonują wyboru zmiennej wejściowej, umożliwia wygenerowanie węzłów Filtr, zezwalających na przekazanie tylko tych zmiennych wejściowych, które w modelu określono jako ważne.

Uwaga: Mogą istnieć niewielkie różnice w ocenach przypisanych do danej obserwacji przez określony model, jeśli oceny są wykonywane za pomocą różnych wersji programu IBM SPSS Modeler. Wynika to zwykle z udoskonaleń oprogramowania wprowadzanych w różnych wersjach.

Aby użyć modelu użytkowego do oceniania danych

1. Połącz model użytkowy ze źródłem danych lub strumieniem, który będzie przekazywał dane.
2. Dodaj lub połącz co najmniej jeden węzeł przetwarzania lub wynikowy (taki jak Tabela lub Analiza) do modelu użytkowego.
3. Wykonaj jeden węzeł poniżej w odniesieniu do węzła modelu.

Uwaga: Do oceniania danych nie można użyć węzła Reguła surowa. Aby ocenić dane na podstawie powiązanego modelu reguł, należy za pomocą węzła Reguła surowa wygenerować model użytkowy Zestaw reguł i użyć tego węzła do oceniania. Więcej informacji można znaleźć w temacie “Generowanie zestawu reguł z powiązanego modelu użytkowego” na stronie 269.

Aby użyć modelu użytkowego do generowania węzłów przetwarzania

1. Rozpocznij przeglądanie modelu na palecie lub jego edycję w obszarze roboczym strumienia.
2. Wybierz żądany typ węzła z menu Utwórz w oknie przeglądarki modeli użytkowych. Dostępne opcje będą się różniły w zależności od typu modelu użytkowego. Szczegółowe informacje o tym, co można wygenerować z poszczególnych modeli, można znaleźć w konkretnym typie modelu użytkowego.

Ponowne generowanie węzła modelowania

Jeśli istnieje model użytkowy, który ma zostać zmodyfikowany lub zaktualizowany, a strumień użyty do utworzenia modelu jest niedostępny, można ponownie wygenerować węzeł modelowania z zastosowaniem takich samych opcji, jakie zostały użyte do utworzenia oryginalnego modelu.

Aby ponownie zbudować model, należy prawym przyciskiem myszy kliknąć model w palecie modeli i wybrać opcję **Utwórz węzeł modelowania**.

Alternatywnie, podczas przeglądania dowolnego modelu, należy wybrać opcję **Utwórz węzeł modelowania** z menu Utwórz.

W większości przypadków ponownie utworzony węzeł modelowania powinien działać identycznie jak ten, którego użyto do utworzenia oryginalnego modelu.

- W przypadku modeli drzewa decyzyjnego dodatkowe ustawienia określone podczas sesji interaktywnej również mogą być zapisane w węźle, a opcja **Stosuj dyrektywy drzewa** będzie aktywna w ponownie utworzonym węźle modelowania.
- W przypadku modeli listy decyzyjnej aktywna będzie opcja **Użyj informacji o zapisanej sesji interaktywnej**. Więcej informacji można znaleźć w temacie “Opcje modelu Lista decyzyjna” na stronie 152.
- W modelach szeregów czasowych aktywna jest opcja **Kontynuuj oszacowanie za pomocą istniejących modeli**, która umożliwia ponowne utworzenie poprzedniego modelu z użyciem bieżących danych. Więcej informacji można znaleźć w temacie Opcje modelu szeregów czasowych.

Importowanie i eksportowanie modeli w formacie PMML

PMML, czyli Predictive Model Markup Language, to format XML służący do opisu modeli eksploracji danych i modeli statystycznych, a w szczególności danych wejściowych modeli, transformacji używanych w celu przygotowania danych do eksploracji oraz parametrów definiujących same modele. IBM SPSS Modeler może importować i eksportować modele opisane w języku PMML, przez co pozwala na współużytkowanie modeli z innymi aplikacjami obsługującymi ten format, na przykład IBM SPSS Statistics.

Więcej informacji na temat języka PMML można znaleźć w serwisie WWW Data Mining Group (*http://www.dmg.org*).

Aby wyeksportować model

Większość typów modeli wygenerowanych w programie IBM SPSS Modeler można eksportować w języku PMML. Więcej informacji można znaleźć w temacie “Typy modeli obsługujące język PMML” na stronie 50.

1. Prawym przyciskiem myszy kliknij model użytkowy na palecie modeli. (Zamiast tego można kliknąć model użytkowy w obszarze roboczym i wybrać menu Plik).
2. W menu kliknij opcję **Eksportuj PMML**.
3. W oknie dialogowym eksportu (lub zapisywania) określ katalog docelowy i unikalną nazwę modelu.

Uwaga:

Opcje eksportu PMML można zmienić w oknie dialogowym Opcje użytkownika. W menu głównym kliknij opcje:

Narzędzia > Opcje > Opcje użytkownika

i kliknij kartę PMML.

Aby zaimportować model zapisany w formacie PMML

Modele wyeksportowane w formacie PMML z programu IBM SPSS Modeler lub innych aplikacji można importować do palety modeli. Więcej informacji można znaleźć w temacie “Typy modeli obsługujące język PMML” na stronie 50.

1. Kliknij paletę modeli prawym przyciskiem myszy i z menu wybierz polecenie **Importuj PMML**.
2. Wybierz plik do zaimportowania i określ opcje dotyczące etykiet zmiennych, odpowiednio do potrzeb.
3. Kliknij przycisk **Otwórz**.

Stosuj etykiety zmiennych. W pliku PMML mogą być określone zarówno nazwy, jak i etykiety zmiennych ze słownika danych (na przykład Identyfikator kierującego jako etykieta zmiennej *RefID*). Wybranie tej opcji spowoduje, że używane będą etykiety zmiennych, jeśli będą obecne w wyeksportowanym pliku PMML.

Jeśli wybrana jest opcja stosowania etykiet zmiennych, ale plik PMML nie zawiera etykiet zmiennych, to używane będą jak zwykle nazwy zmiennych.

Typy modeli obsługujące język PMML
Eksport PMML

Modele IBM SPSS Modeler. Następujące modele utworzone w programie IBM SPSS Modeler można eksportować w języku PMML 4.0:

- C&RT
- QUEST
- CHAID
- Sieci neuronowe
- C5.0
- Regresja logistyczna
- Genlin
- SVM
- Apriori
- Carma
- K-średnie
- Sieć Kohonena
- Dwustopniowa
- Dwustopniowa-AS
- GLMM (PMML jest eksportowany dla wszystkich modeli GLMM, ale PMML ma jedynie efekty stałe)
- Lista decyzyjna
- Model Coxa
- Sekwencja (ocenianie modeli PMML typu Sekwencja nie jest obsługiwane)
- Drzewa losowe
- Drzewo-AS
- Liniowy
- Liniowy-AS
- Regresja
- Regresja logistyczna
- GLE
- LSVM
- Wykrywanie anomalii
- KNN
- Reguły asocjacyjne

Modele rodzime bazy danych. Spośród modeli wygenerowanych przez algorytmy rodzime bazy danych, eksport PMML jest niedostępny. Nie można eksportować modeli utworzonych za pomocą programu Analysis Services firmy Microsoft lub programu Oracle Data Miner.

Import modeli PMML

IBM SPSS Modeler może importować i oceniać modele PMML wygenerowane przez aktualne wersje wszystkich produktów IBM SPSS Statistics, w tym modele wyeksportowane z programu IBM SPSS Modeler oraz kod PMML modelu lub transformacji wygenerowany przez IBM SPSS Statistics w wersji 17.0 lub nowszej. Zasadniczo kryteria te spełnia każdy model PMML, który mechanizm oceniania jest w stanie ocenić, z następującymi wyjątkami:

- Apriori, CARMA, Wykrywanie anomalii, Sekwencje i Reguły asocjacyjne — tych modeli nie można importować.

- Modeli PMML nie można przeglądać po zaimportowaniu do programu IBM SPSS Modeler, mimo że mogą być używane w ocenianiu. (Uwaga: dotyczy to także modeli, które pierwotnie wyeksportowano z programu IBM SPSS Modeler. Aby uniknąć tego ograniczenia, należy wyeksportować model jako wygenerowany plik modelu [*.gm], a nie w języku PMML).
- Przy importowaniu poprawność jest sprawdzana w ograniczonym zakresie, ale przy próbie oceny modelu przeprowadzane jest pełne sprawdzanie poprawności. Dlatego może się zdarzyć, że mimo udanego importu ocenianie nie powiedzie się lub przyniesie nieprawidłowe wyniki.

Uwaga: W przypadku modeli PMML z innych programów zaimportowanych do programu IBM SPSS Modeler podejmowana jest próba oceny poprawnego modelu PMML, który da się rozpoznać i ocenić. Nie ma jednak gwarancji, że wszystkie modele PMML dadzą się ocenić i że będą oceniane tak samo, jak w aplikacji, która je wygenerowała.

Publikowanie modeli dla adaptera oceniania

Modele można publikować do serwera bazy danych, na którym zainstalowany jest adapter oceniania. Adapter oceniania umożliwia przeprowadzenie oceniania modelu w bazie danych przy użyciu możliwości funkcji bazy danych zdefiniowanych przez użytkownika (UDF). Wykonywanie oceniania w bazie danych eliminuje potrzebę wyodrębniania danych przed ocenianiem. Publikowanie do adaptera oceniania powoduje również wygenerowanie przykładowego kodu SQL w celu wykonania UDF.

Aby opublikować składnik Scoring Adapter

1. Kliknij dwukrotnie model użytkowy, aby go otworzyć.
2. Z menu modelu użytkowego wybierz opcje:

   Plik > Publikacja dla Server Scoring Adapter
3. Wypełnij wymagane pola w oknie dialogowym i kliknij przycisk **OK**.

Połączenie z bazą danych. Szczegóły połączenia z bazą danych, która będzie używana na potrzeby modelu.

Identyfikator publikacji. (Tylko w przypadku baz danych Db2 for z/OS) Identyfikator dla modelu. Jeśli ten sam model zostanie przebudowany i używany będzie ten sam identyfikator publikacji, wówczas wygenerowany kod SQL pozostanie taki sam, dlatego możliwe będzie przebudowanie modelu bez konieczności zmiany aplikacji, która używa poprzednio wygenerowanego kodu SQL. (W przypadku innych baz danych generowany kod SQL jest unikatowy dla modelu).

Wygeneruj przykładowy SQL. Jeśli ta opcja zostanie wybrana, wówczas przykładowy kod SQL zostanie wygenerowany do pliku określonego w polu **Plik**.

Modele surowe

Model surowy zawiera informacje wyodrębnione z danych, ale nie jest przeznaczony do bezpośredniego generowania predykcji. Oznacza to, że nie można go dodawać do strumieni. Modele surowe reprezentuje ikona przedstawiająca „nieobrobiony diament”, widoczna w palecie modeli.

Rysunek 27. Ikona modelu surowego

Aby wyświetlić informacje o surowym modelu reguły, kliknij prawym przyciskiem myszy model i wybierz opcję **Przeglądaj** z menu kontekstowego. Na różnych kartach — podobnie jak w przypadku innych modeli wygenerowanych w produkcie IBM SPSS Modeler — dostępne jest podsumowanie modelu i informacje o regule.

Generowanie węzłów. Menu Utwórz umożliwia tworzenie nowych węzłów na podstawie reguł.

- **Węzeł wyboru.** Generuje węzeł selekcji w celu wybierania rekordów, do których ma zastosowanie aktualnie wybrana reguła. Ta opcja jest wyłączona, jeśli nie wybrano reguły.

- **Zestaw reguł.** Generuje węzeł Zestaw reguł w celu przewidywania wartości dla pojedynczej zmiennej przewidywanej. Więcej informacji można znaleźć w temacie “Generowanie zestawu reguł z powiązanego modelu użytkowego” na stronie 269.

Rozdział 4. Modele monitorujące

Zmienne monitorowania i rekordy

W trakcie wstępnych etapów analizy może być używanych wiele węzłów modelowania; pozwala to lokalizować zmienne i rekordy, które z dużym prawdopodobieństwem okażą się istotne dla modelowania. Istnieje możliwość użycia węzła Dobór predyktorów do monitorowania i rangowania zmiennych wg ważności, oraz węzła Wykrywanie anomalii do odszukania nietypowych rekordów, niezgodnych ze znanymi wzorcami „zwykłych" danych.

Węzeł Dobór predyktorów przegląda zmienne wejściowe do usunięcia w oparciu o zbiór kryteriów (takich jak procent braków danych); następnie nadaje rangę istotności pozostałych danych wejściowych względem określonej zmiennej przewidywanej. Na przykład, jeśli mamy zbiór danych z setkami potencjalnych danych wejściowych, to które z nich z dużym prawdopodobieństwem okażą się użyteczne w modelowaniu wyników leczenia pacjenta?

Węzeł Anomalie umożliwia identyfikację nietypowych obserwacji lub wartości odstających, które są niezgodne z wzorcami dla „normalnych" danych. Korzystając z tego węzła, można zidentyfikować wartości odstające nawet, jeśli nie pasują one do żadnego z wcześniej znanych wzorców oraz jeśli brak pewności co do charakteru poszukiwanych danych.

Należy zwrócić uwagę, że ta opcja wykrywania anomalii identyfikuje nietypowe rekordy lub obserwacje w oparciu o zestaw zmiennych wybranych w modelu bez względu na jakąkolwiek konkretną zmienną przewidywaną (zależną) oraz niezależnie od tego, czy te zmienne są istotne dla przewidywanego wzorca. Z tego względu może okazać się wskazane zastosowanie wykrywania anomalii w połączeniu z wyborem predyktora lub inną technika monitorowania i rangowania zmiennych. Można na przykład zastosować wybór predyktora do identyfikacji najistotniejszych zmiennych względem określonej zmiennej przewidywanej, a następnie, korzystając z wykrywania anomalii, zlokalizować rekordy najbardziej nietypowe przy uwzględnieniu tych zmiennych. (Alternatywą byłoby utworzenie modelu drzewa decyzyjnego, a następnie zbadanie wszelkich błędnie sklasyfikowanych rekordów jako potencjalnych anomalii. Metoda ta byłaby jednak znacznie trudniejsza do zreplikowania lub zautomatyzowania na większą skalę).

Węzeł wyboru predyktora

Problemy z eksploracją danych mogą obejmować setki, a nawet tysiące zmiennych, które mogą potencjalnie stanowić wartości wejściowe. W wyniku tego sprawdzenie, które zmienne mogą być uwzględnione w modelu, może być bardzo czasochłonne i wymagać wiele wysiłku. Aby zawęzić możliwość wyboru, można użyć algorytmu Dobór predyktorów, który pozwoli zidentyfikować zmienne najbardziej istotne dla danej analizy. Przykładowo, jeśli podejmowana jest próba predykcji danych wynikowych pacjenta w oparciu o liczbę czynników, które czynniki najprawdopodobniej będą istotne?

Wybór predyktora przeprowadzany jest w trzech krokach:

- **Monitorowanie.** Usuwa nieistotne i problematyczne zmienne wejściowe i rekordy lub obserwacje, takie jak zmienne wejściowe ze zbyt dużą liczbą braków wartości lub ze zbyt dużą lub zbyt małą zmiennością.
- **Rangowanie.** Umożliwia sortowanie pozostałych zmiennych i przypisanie rang na podstawie ważności.
- **Wybór.** Określa podzbiór predyktorów, jakie będą używane w kolejnych modelach — na przykład poprzez zachowanie wyłącznie najistotniejszych zmiennych wejściowych i odfiltrowanie lub wykluczenie pozostałych.

W czasach, w których wiele organizacji operuje zbyt dużą ilością danych, korzyści, jakie zapewnia wybór predyktora dla uproszczenia i przyspieszenia procesu modelowania, mogą być bardzo wymierne. Szybkie przeniesienie zainteresowania na najbardziej istotne zmienne pozwala zredukować liczbę koniecznych obliczeń; znacznie łatwiej jest zlokalizować niewielką liczbę istotnych relacji, które w innej sytuacji mogłyby zostać przeoczone, i ostatecznie

uzyskać prostsze, bardziej dokładne i łatwiejsze do objaśnienia modele. Zmniejszenie liczby zmiennych używanych w modelu może pozwolić na ograniczenie liczby przeprowadzanych ocen, jak również ilości danych zgromadzonych podczas przyszłych iteracji.

Przykład. Firma telekomunikacyjna dysponuje składnicą danych zawierającą informacje na temat odpowiedzi na specjalną promocję udzieloną przez 5000 klientów firmy. Dane obejmują dużą liczbę zmiennych dotyczących wieku, zatrudnienia, dochodów klientów oraz statystyki dot. korzystania z telefonu. Trzy zmienne przewidywane przedstawiają dane, czy klient odpowiedział na wszystkie trzy oferty. Firma chce użyć tych danych, aby lepiej przewidzieć, którzy klienci najprawdopodobniej odpowiedzą na podobne oferty w przyszłości.

Wymagania. Jedna zmienna przewidywana (której rola jest ustawiona jako *Przewidywana*) oraz wiele zmiennych wejściowych, jakie mają być monitorowane lub rangowane w odniesieniu do zmiennej przewidywanej. Dla zmiennych przewidywanych i wejściowych poziom pomiaru może być ustawiony jako *Ilościowa* (zakres liczbowy) lub *Jakościowa*.

Ustawienia modelu Dobór predyktorów

Ustawienie na karcie Model obejmuje opcje modelu standardowego oraz ustawienia, które umożliwiają dostosowanie kryteriów monitorowania zmiennych wejściowych.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Monitorowanie zmiennych wejściowych

Monitorowanie obejmuje usuwanie zmiennych wejściowych lub obserwacji, które nie dają przydatnych informacji w odniesieniu do reakcji zmiennej wejściowej/przewidywanej. Opcje monitorowania są tworzone w oparciu o atrybuty danej zmiennej bez uwzględniania jakości predykcji w odniesieniu do wybranej zmiennej przewidywanej. Zmienne poddawane monitorowaniu są wykluczane z obliczeń używanych do rangowania zmiennych wejściowych i opcjonalnie można je odfiltrować lub usunąć z danych użytych w modelowaniu.

Zmienne mogą być monitorowane wg następujących kryteriów:

- **Maksymalny procent braków danych.** Monitorowane są zmienne ze zbyt dużą liczbą braków danych, wyrażoną jako procent łącznej liczby rekordów. Zmienne z dużym procentem braków danych zapewniają mało przewidywalne informacje.
- **Maksymalny odsetek rekordów w pojedynczej kategorii.** Monitorowane są zmienne ze zbyt dużą liczbą rekordów należących do tej samej kategorii w odniesieniu do łącznej liczby rekordów. Na przykład, jeśli 95% klientów w bazie danych jeździ tym samym modelem samochodu, uwzględnienie tej informacji nie będzie przydatne dla rozróżnienia klientów. Monitorowane są wszystkie zmienne, które przekroczyły określone maksimum. Ta opcja dotyczy tylko zmiennych jakościowych.
- **Maksimum kategorii jako procent wszystkich rekordów.** Monitorowane są zmienne ze zbyt dużą liczbą kategorii w odniesieniu do łącznej liczby rekordów. Jeśli zbyt duży procent kategorii zawiera tylko jedną obserwację, użycie zmiennej może być ograniczone. Na przykład, jeśli każdy klient nosi inny kapelusz, istnieje małe prawdopodobieństwo, że informacja ta będzie użyteczna podczas modelowania wzorców zachowania. Ta opcja dotyczy tylko zmiennych jakościowych.
- **Minimalny współczynnik zmienności.** Monitorowane są zmienne o współczynniku zmienności z mniejszym od określonego minimum lub mu równym. Ta miara to stosunek standardowego odchylenia zmiennej wejściowej do średniej dla zmiennej wejściowej. Jeśli ta wartość jest bliska zeru, zmienność wartości dla tej zmiennej jest zbyt mała. Ta opcja dotyczy tylko zmiennych ilościowych (zakres liczbowy).
- **Minimalne odchylenie standardowe.** Monitorowane są zmienne z odchyleniem standardowym mniejszym od określonego minimum lub mu równym. Ta opcja dotyczy tylko zmiennych ilościowych (zakres liczbowy).

Rekordy z brakami danych. Rekordy lub obserwacje, które mają braki danych dla zmiennej przewidywanej lub braki wartości dla wszystkich zmiennych wejściowych, są automatycznie wykluczane ze wszystkich obliczeń używanych podczas rangowania.

Opcje doboru predyktorów

Na karcie Opcje można określić domyślne ustawienia wybierania lub wykluczania zmiennych wejściowych w modelu użytkowym. Następnie można dodać model do strumienia, aby wybrać podzbiór zmiennych, jakie będą używane w kolejnych działaniach związanych z budowaniem modelu. Można również zastąpić te ustawienia, zaznaczając lub usuwając zaznaczenie dodatkowych zmiennych w przeglądarce modelu po wygenerowaniu modelu. Ustawienia domyślne umożliwiają jednak zastosowanie modelu użytkowego bez wprowadzania zmian, co może być przydatne w przypadku tworzenia skryptów.

Więcej informacji można znaleźć w temacie “Wyniki dla modelu wyboru predyktora” na stronie 56.

Dostępne są następujące opcje:

Wszystkie zmienne z rangą. Wybiera zmienne na podstawie ich rangi: *important*, *marginal* lub *unimportant*. Można przeprowadzić edycję etykiety każdej rangi oraz wartości odcięcia użytych do przypisania rekordów do określonej rangi.

Określona liczba najważniejszych zmiennych. Wybiera pierwszych *n* zmiennych na podstawie ważności.

Ważność wyższa niż. Wybiera wszystkie zmienne, których ważność jest większa od określonej wartości.

Zmienna przewidywana jest zawsze zachowywana niezależnie od wyboru.

Opcje rangowania ważności

Wszystkie jakościowe. Jeśli wszystkie zmienne wejściowe i przewidywane są jakościowe, ważność może być rangowana na podstawie jednej z czterech miar:

- **Chi-kwadrat Pearsona.** Testuje niezależność zmiennej przewidywanej i zmiennej wejściowej bez wskazywania siły lub kierunku istniejących relacji.
- **Stosunek wiarygodności chi-kwadrat.** Działanie podobne jak w przypadku opcji Chi-kwadrat Pearsona, ale testowana jest również niezależność pomiędzy zmienną przewidywaną i wejściową.
- **V Kramera.** Miara związku na podstawie statystyki chi-kwadrat Pearsona. Wartości należą do zakresu od 0, co oznacza brak związku, do 1, co oznacza doskonały związek.
- **Lambda.** Miara związku odzwierciedlająca proporcjonalną redukcję błędu w przypadku użycia zmiennej do przewidywania wartości przewidywanych. Wartość 1 oznacza, że zmienna wejściowa doskonale przewiduje zmienną przewidywaną, a wartość 0 oznacza, że zmienna wejściowa nie zapewnia żadnych przydatnych informacji na temat zmiennej przewidywanej.

Niektóre jakościowe. Jeśli niektóre — ale nie wszystkie — zmienne wejściowe są jakościowe i zmienne przewidywane również są jakościowe, ważność może być rangowana na podstawie metody chi-kwadrat Pearsona lub ilorazu wiarygodności. (Metody *V* Cramera i lambda są niedostępne, jeśli nie wszystkie zmienne wejściowe są jakościowe).

Jakościowe a ilościowe. Podczas rangowania jakościowych zmiennych wejściowych w odniesieniu do ilościowych zmiennych przewidywanych lub odwrotnie (jedna lub druga jest jakościowa, ale nie obie) używana jest statystyka *F*.

Obie ilościowe. Podczas rangowania jakościowych zmiennych wejściowych w odniesieniu do jakościowych zmiennych przewidywanych używana jest statystyka *t* w oparciu o współczynnik korelacji.

Modele użytkowe wyboru predyktora

Modele użytkowe wyboru predyktora przedstawiają ważność każdej zmiennej wejściowej w odniesieniu do wybranej zmiennej przewidywanej, zgodnie z rangą określoną przez węzeł Dobór predyktorów. Na liście są również wyświetlane wszystkie zmienne, które były monitorowane przed rangowaniem. Więcej informacji można znaleźć w temacie “Węzeł wyboru predyktora” na stronie 53.

Po uruchomieniu strumienia zawierającego model użytkowy wyboru predyktora model działa jak filtr, który zachowuje tylko wybrane zmienne wejściowe, zgodnie z bieżącym wyborem na karcie Model. Na przykład, można zaznaczyć wszystkie rangowane zmienne jako ważne (jedna z opcji domyślnych) lub ręcznie wybrać podzbiór zmiennych na karcie Model. Zmienna przewidywana jest również zachowywana niezależnie od wyboru. Wszystkie pozostałe zmienne są wykluczane.

Filtrowanie odbywa się wyłącznie na podstawie nazwy zmiennej; na przykład, jeśli wybrane zostaną zmienne *age* i *income*, wszystkie zmienne, które są zgodne z jedną z tych nazw, zostaną zachowane. Model nie aktualizuje rangowania zmiennej na podstawie nowych danych; po prostu filtruje zmienne na podstawie wybranych nazw. Dlatego podczas stosowania tego modelu dla nowych lub zaktualizowanych danych należy zachować ostrożność. W razie wątpliwości zalecane jest ponowne wygenerowanie modelu.

Wyniki dla modelu wyboru predyktora

Na karcie Model modelu użytkowego wyboru predyktora w górnym panelu wyświetlana jest ranga i ważność wszystkich zmiennych wejściowych; można tu wybrać zmienne do filtrowania, używając pól wyboru w kolumnie po lewej stronie. Po uruchomieniu strumienia zachowane zostaną tylko wybrane zmienne; pozostałe zmienne są odrzucane. Wybór domyślny dokonywany jest na podstawie opcji określonych w węźle budowania modelu, jednak w razie potrzeby można zaznaczyć lub usunąć zaznaczenie dodatkowych zmiennych.

W dolnym panelu wyświetlone są zmienne wejściowe, które zostały wykluczone z rangowania na podstawie wartości procentowej braków danych lub innych kryteriów określonych w węźle modelowania. Podobnie jak w przypadku zmiennych z rangą można dołączyć lub odrzucić te zmienne, używając pól wyboru w kolumnie po lewej stronie. Więcej informacji można znaleźć w temacie “Ustawienia modelu Dobór predyktorów” na stronie 54.

- Aby posortować listę według rangi, nazwy zmiennej, ważności lub innej wyświetlanej kolumny, należy kliknąć nagłówek kolumny. Lub, aby użyć paska narzędzi, należy zaznaczyć żądany element z listy Sortuj wg i za pomocą strzałek w górę i w dół zmienić kierunek sortowania.
- Pasek narzędzi pozwala zaznaczyć lub usunąć zaznaczenie wszystkich zmiennych oraz uzyskać dostęp do okna dialogowego Zaznacz zmienne, co pozwoli zaznaczyć zmienne według rangi lub ważności. Można także nacisnąć klawisze Shift i Ctrl, klikając jednocześnie zmienne, aby rozszerzyć wybór oraz użyć spacji, aby włączyć lub wyłączyć grupę wybranych zmiennych. Więcej informacji można znaleźć w temacie “Wybieranie zmiennych według ważności”.
- Wartości graniczne dla rangowania zmiennych wejściowych, takie jak ważne, brzegowe lub nieważne, są wyświetlane w legendzie pod tabelą. Wartości te są określane w węźle modelowania. Więcej informacji można znaleźć w temacie “Opcje doboru predyktorów” na stronie 55.

Wybieranie zmiennych według ważności

Podczas monitorowania danych za pomocą modelu użytkowego wyboru predyktora wszystkie zmienne wybrane z listy zmiennych z rangą lub monitorowanych — co wskazują pola wyboru w kolumnie po lewej stronie — zostaną zachowane. Pozostałe zmienne zostaną odrzucone. Aby zmienić wybór, można za pomocą paska narzędzi uzyskać dostęp do okna dialogowego Zaznacz zmienne, co pozwoli na wybranie zmiennych według rangi lub ważności.

Wszystkie zaznaczone zmienne. Wybiera wszystkie zmienne oznaczone jako ważne, brzegowe lub nieważne.

Określona liczba najważniejszych zmiennych. Umożliwia wybór *n* najważniejszych zmiennych na podstawie ważności.

Ważność wyższa niż. Wybiera wszystkie zmienne, których ważność jest większa od określonej wartości granicznej.

Generowanie filtru z modelu wyboru predyktora

Po uzyskaniu wyników z modelu wyboru predyktora można użyć okna dialogowego Generuj filtr na podstawie wyboru predyktorów, aby wygenerować co najmniej jeden węzeł filtrowania, który będzie uwzględniał lub wykluczał podzbiory zmiennych w oparciu o ważność w odniesieniu do określonej zmiennej przewidywanej. Możliwość użycia modelu użytkowego jako filtru zapewnia elastyczność przeprowadzania doświadczeń dla różnych podzbiorów zmiennych bez konieczności kopiowania lub modyfikowania modelu. Zmienna przewidywana jest zawsze zachowywana przez filtr, niezależnie od tego, czy wybrano opcję uwzględniania czy wykluczania.

Uwzględnij/Wyklucz. Zmienne można uwzględniać lub wykluczać — na przykład, aby uwzględnić 10 pierwszych zmiennych lub wykluczyć wszystkie zmienne oznaczone jako nieważne.

Wybrane zmienne. Uwzględnia lub wyklucza wszystkie zmienne aktualnie zaznaczone w tabeli.

Wszystkie zaznaczone zmienne. Wybiera wszystkie zmienne oznaczone jako ważne, brzegowe lub nieważne.

Określona liczba najważniejszych zmiennych. Umożliwia wybór *n* najważniejszych zmiennych na podstawie ważności.

Ważność wyższa niż. Wybiera wszystkie zmienne, których ważność jest większa od określonej wartości granicznej.

Węzeł Anomalie

Modele wykrywania anomalii służą do wykrywania wartości odstających lub nietypowych obserwacji w danych. W odróżnieniu od innych metod modelowania, które zapisują reguły dotyczące nietypowych obserwacji, modele wykrywania anomalii zapisują informację o zachowaniach normalnych. Dzięki temu możliwe jest zidentyfikowanie wartości odstających nawet wtedy, gdy nie pasują one do żadnego znanego wzorca, co może być szczególnie użyteczne w takich zastosowaniach, jak wykrywanie oszustw, w których wciąż pojawiają się nowe wzorce (metody popełnienia oszustwa). Wykrywanie anomalii jest metodą nienadzorowaną, co oznacza, że nie wymaga początkowego uczącego zbioru danych zawierającego znane obserwacje oszustw.

Podczas gdy tradycyjne metody wykrywania wartości odstających z reguły analizują jednocześnie jedną lub dwie zmienne, algorytm wykrywania anomalii może analizować duże liczby zmiennych, by wykryć skupienia lub grupy podobnych rekordów. Każdy rekord jest następnie porównywany z innymi rekordami w tej samej grupie w celu wykrycia ewentualnych anomalii. Im bardziej odległa jest obserwacja od normalnego środka grupy, tym bardziej prawdopodobne jest, iż mamy do czynienia z obserwacją nietypową. Algorytm może na przykład zgrupować rekordy w trzy osobne skupienia i oznaczyć rekordy, które wypadają daleko od środka swoich skupień.

Każdemu rekordowi przypisuje się indeks anomalii, czyli iloraz indeksu odchylenia grupy od średniej ze skupienia, do którego należy obserwacja. Im większy indeks, tym większe odchylenie obserwacji od średniej. W typowych warunkach obserwacje z indeksem anomalii poniżej 1, a nawet 1,5, nie są uznawane za anomalie, ponieważ odchylenie jest prawie takie samo lub nieznacznie większe od średniej. Jednak obserwacje z indeksem większym niż 2 są dobrymi kandydatami na anomalie, ponieważ odchylenie jest co najmniej dwukrotnie większe od średniej.

Wykrywanie anomalii jest metodą eksploracyjną pomyślaną jako sposób na szybkie wykrywanie nietypowych obserwacji lub rekordów będących kandydatami do dalszej analizy. Należy je traktować jako obserwacje/rekordy *podejrzane*, które po bliższym zbadaniu mogą, ale nie muszą okazać się rzeczywistymi anomaliami. Może okazać się, że rekord jest stuprocentowo poprawny, ale warto monitorować go na potrzeby budowania modelu. Może się też zdarzyć, że algorytm będzie stale zgłaszał fałszywe anomalie, co świadczyć może o błędzie lub artefakcie w procesie zbierania danych.

Należy zwrócić uwagę, że ta opcja wykrywania anomalii identyfikuje nietypowe rekordy lub obserwacje w oparciu o zestaw zmiennych wybranych w modelu bez względu na jakąkolwiek konkretną zmienną przewidywaną (zależną) oraz niezależnie od tego, czy te zmienne są istotne dla przewidywanego wzorca. Z tego względu może okazać się wskazane zastosowanie wykrywania anomalii w połączeniu z wyborem predyktora lub inną technika monitorowania i rangowania zmiennych. Można na przykład zastosować wybór predyktora do identyfikacji najistotniejszych zmiennych względem

określonej zmiennej przewidywanej, a następnie, korzystając z wykrywania anomalii, zlokalizować rekordy najbardziej nietypowe przy uwzględnieniu tych zmiennych. (Alternatywą byłoby utworzenie modelu drzewa decyzyjnego, a następnie zbadanie wszelkich błędnie sklasyfikowanych rekordów jako potencjalnych anomalii. Metoda ta byłaby jednak znacznie trudniejsza do zreplikowania lub zautomatyzowania na większą skalę).

Przykład. W procesie weryfikacji wniosków o dofinansowanie dla projektów rozwoju rolnictwa w celu wykrycia ewentualnych oszustw można zastosować technikę wykrywania anomalii, aby ujawniać odstępstwa od normy oraz wyróżniać rekordy nietypowe i warte dokładniejszego zbadania. Szczególnie interesują nas wnioski o dofinansowanie na kwotę zbyt wysoką (lub zbyt niską) w stosunku do rodzaju i wielkości gospodarstwa.

Wymagania. Jedna lub wiele zmiennych wejściowych. Należy zwrócić uwagę, że jako zmienne wejściowe można wykorzystać tylko zmienne o roli wejściowej z przypisanym źródłem lub węzłem wprowadzania danych. Zmienne przewidywane (rola zmiennej przewidywanej lub obie role) są ignorowane.

Mocne strony. Oznaczając obserwacje, które *nie* spełniają znanego zestawu kryteriów, a nie te, które kryteria spełniają, modele wykrywania anomalii mogą rozpoznać nietypowe obserwacje nawet wówczas, gdy nie są one zgodne ze znanymi wcześniej wzorcami. W połączeniu z wyborem predyktorów wykrywanie anomalii umożliwia analizowanie dużych ilości danych w celu stosunkowo szybkiego wykrycia najbardziej interesujących rekordów.

Opcje modelu Wykrywanie anomalii

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Ustal wartość odcięcia dla anomalii w oparciu o. Określa metodę używaną do określania wartości odcięcia przy oznaczaniu anomalii. Dostępne są następujące opcje:

- **Minimalny poziom indeksu anomalii.** Określa minimalną wartość odcięcia używaną do oznaczania anomalii. Oznaczane są rekordy z wartością równą lub większą od wartości odcięcia.
- **Procent najbardziej nietypowych rekordów w danych uczących.** Automatycznie określa próg na takim poziomie, by oznaczony został określony odsetek rekordów w danych uczących. Uzyskana wartość odcięcia jest włączana do modelu jako parametr. Należy pamiętać, że ta opcja określa sposób ustalania wartości odcięcia, a *nie* faktyczny odsetek rekordów, które zostaną oznaczone w procesie oceniania. Rzeczywiste wyniki oceniania mogą być różne, w zależności od danych.
- **Liczba najbardziej nietypowych rekordów w danych uczących.** Automatycznie określa próg na takim poziomie, by oznaczona została określona liczba rekordów w danych uczących. Uzyskana wartość progowa jest włączana do modelu jako parametr. Należy pamiętać, że ta opcja określa sposób ustalania wartości odcięcia, a *nie* konkretna liczba rekordów, które zostaną oznaczone w procesie oceniania. Rzeczywiste wyniki oceniania mogą być różne, w zależności od danych.

Uwaga: Niezależnie od sposobu określenia wartości odcięcia nie wpływa ona na indeks anomalii przypisywany do każdego z rekordów. Określa ona jedynie próg oznaczania rekordów jako nietypowe podczas szacowania lub oceny modelu. Aby później przeanalizować większą lub mniejszą liczbę rekordów, można użyć węzła selekcji do wybrania podzbioru rekordów na podstawie wartości indeksu anomalii (`$O-AnomalyIndex > X`).

Liczba nieprawidłowych zmiennych w raporcie. Określa liczbę zmiennych, która ma być podawana jako uzasadnienie oznaczenia konkretnego rekordu jako anomalii. Zgłaszane są najbardziej nietypowe zmienne, tj. te, które najbardziej odbiegają od normy danej zmiennej dla grupy, do której przypisany jest rekord.

Zaawansowane opcje wykrywania anomalii

Aby określić opcje dotyczące braków danych i inne ustawienia, włącz tryb **Zaawansowany** na karcie Zaawansowany.

Współczynnik regulacji. Wartość używana do równoważenia względnej wagi nadanej zmiennym ciągłym (przedział liczbowy) i jakościowym przy obliczaniu odległości. Większe wartości zwiększają wpływ na zmienne ciągłe. Musi to być wartość różna od zera.

Automatycznie wylicz liczbę grup elementów równorzędnych. Wykrywanie anomalii można wykorzystać do szybkiego analizowania dużej liczby możliwych rozwiązań w celu wybrania optymalnej liczby grup elementów równorzędnych dla danych uczących. Można poszerzyć lub zawęzić przedział, określając minimalną i maksymalną liczbę grup elementów równorzędnych. Większe wartości umożliwią systemowi eksplorowanie szerszego zbioru możliwych rozwiązań, jednak kosztem większego czasu przetwarzania.

Zdefiniuj liczbę grup elementów równorzędnych. Jeśli wiesz, ile grup należy uwzględnić w modelu, wybierz tę opcję i wprowadź liczbę grup elementów równorzędnych. Wybranie tej opcji z reguły korzystnie wpływa na wydajność.

Poziom i współczynnik szumów. Ustawienia te określają sposób traktowania wartości odstających podczas grupowania dwustopniowego. W pierwszym etapie używane jest drzewo predyktorów grupy w celu skondensowania danych — tj. zastąpienia bardzo dużej liczby odrębnych rekordów przez mniejszą liczbę grup, które będzie można łatwiej analizować. Drzewo budowane jest na podstawie miar podobieństwa, a gdy węzeł drzewa zawiera zbyt wiele rekordów, dzieli się na węzły podrzędne. W drugim etapie na węzłach końcowych drzewa predyktorów skupień realizowane jest grupowanie hierarchiczne. Obsługa szumu jest włączona przy pierwszym przejściu przez dane, a wyłączona przy drugim przejściu. Obserwacje z grupy szumu z pierwszego przebiegu są przypisywane do zwykłych grup w drugim przebiegu.

- **Poziom szumów.** Podaj wartość w zakresie od 0 do 0,5. To ustawienie ma znaczenie tylko wtedy, gdy drzewo predyktorów grupy zapełni się podczas faz wzrostu, tj. gdy nie będzie mogło przyjąć więcej obserwacji w węźle-liściu i nie będzie możliwości podziału żadnego z takich węzłów.

  Jeśli drzewo predyktorów grupy zapełni się, a poziom szumów będzie ustawiony na 0, to próg zostanie podwyższony, a drzewo predyktorów grupy urośnie od nowa ze wszystkimi obserwacjami. Po zakończeniu finalnego grupowania wartości, których nie można przypisać do żadnego skupienia, zostaną oznaczone jako odstające. Grupie odstającej przypisywany jest numer identyfikacyjny -1. Grupa odstająca nie jest uwzględniana w ogólnej liczbie grup. A zatem, jeśli użytkownik określi *n* grup i włączy obsługę szumu, algorytm wygeneruje *n* grup plus jedną grupę szumów. W praktyce zwiększenie tej wartości daje algorytmowi więcej swobody, pozwalając na ulokowanie nietypowych rekordów w drzewie zamiast przypisywania ich do odrębnej grupy odstającej.

  Jeśli drzewo predyktorów grupy zapełni się, a poziom szumów będzie wyższy od 0, drzewo urośnie od nowa po umieszczeniu wszelkich danych z liści rzadkich w ich własnym liściu szumu. Liść jest uznawany za rzadki, jeśli stosunek liczby obserwacji w liściu rzadkim do liczby obserwacji w największym liściu jest mniejszy od poziomu szumów. Po zbudowaniu drzewa predyktorów skupień zostaną w nim umieszczone obserwacje odstające, jeśli istnieć będzie taka możliwość. W przeciwnym razie przed drugą fazą grupowania wartości odstające zostaną odrzucone.
- **Współczynnik szumów.** Określa, jaka część pamięci przydzielonej dla komponentu ma być używana do buforowania szumów. Wartość ta należy do przedziału od 0,0 do 0,5. Jeśli umieszczenie danej obserwacji w liściu drzewa powodowałoby takie zagęszczenie obserwacji, że odległość między nimi byłaby niższa od progu, liść nie jest dzielony. Jeśli zagęszczenie jest wyższe od progu, liść jest dzielony, co powoduje dodanie kolejnej małej grupy do drzewa predyktorów. W praktyce zwiększenie tej wartości może spowodować, że algorytm będzie szybciej zmierzał do budowy prostszego drzewa.

Podstaw braki danych. W przypadku zmiennych ciągłych powoduje wstawienie średniej wartości zmiennej w miejsce braków danych. W przypadku zmiennych jakościowych brakujące kategorie są łączone i traktowane jak jedna poprawna kategoria. Jeśli ta opcja nie będzie zaznaczona, to rekordy z brakami danych zostaną wykluczone z analizy.

Modele użytkowe wykrywania anomalii

Model użytkowy wykrywania anomalii zawiera wszystkie informacje zarejestrowane w modelu wykrywania anomalii, a także informacje o danych uczących i procesie estymacji.

Po uruchomieniu strumienia zawierającego model użytkowy wykrywania anomalii do strumienia tego dodawanych jest szereg nowych zmiennych, zgodnie z wyborami dokonanymi na karcie Ustawienia w modelu użytkowym. Więcej informacji można znaleźć w temacie “Ustawienia modelu wykrywania anomalii” na stronie 60. Nazwy nowych zmiennych tworzone są na podstawie nazwy modelu z przedrostkiem *$O*, tak jak przedstawiono to w poniższej tabeli.

Tabela 6. Generowanie nazw nowych zmiennych

Nazwa zmiennej	Opis
$O-Anomaly	Zmienna typu flaga określająca, czy rekord jest nietypowy (tj. czy jest anomalią).
$O-AnomalyIndex	Wartość indeksu anomalii dla rekordu.
$O-PeerGroup	Określa grupę elementów równorzędnych, do której przypisany jest rekord.
$O-Field-n	Nazwa *n*-tej zmiennej pod względem nietypowości, tj. odchylenia od normy grupy.
$O-FieldImpact-n	Indeks odchylenia zmiennej. Ta wartość jest miarą odchylenia od normy zmiennej w grupie, do której rekord jest przypisany.

Opcjonalnie można wyłączyć oceny węzłów niebędących anomaliami, aby wyniki były bardziej czytelne. Więcej informacji można znaleźć w temacie "Ustawienia modelu wykrywania anomalii".

Szczegóły modelu wykrywania anomalii

Na karcie Model wygenerowanego modelu Wykrywanie anomalii wyświetlane są informacje o grupach elementów równorzędnych w modelu.

Należy zwrócić uwagę, że podane wielkości i statystyki grup elementów równorzędnych są oszacowaniami opartymi na danych uczących i mogą nieznacznie różnić się od faktycznych wyników oceniania, nawet jeśli oceniane były te same dane.

Podsumowanie modelu wykrywania anomalii

Karta Podsumowanie modelu użytkowego wykrywania anomalii zawiera informacje o zmiennych, ustawieniach budowania i procesie szacowania. Podana jest także liczba grupy elementów równorzędnych wraz z wartością odcięcia używaną do oznaczania rekordów jako nietypowych.

Ustawienia modelu wykrywania anomalii

Na karcie Ustawienia określ opcje oceniania modelu użytkowego.

Wskaż nietypowe rekordy za pomocą Określa, w jaki sposób rekordy nietypowe mają być traktowane w wynikach.

- **Flaga i indeks** Tworzy flagę ustawioną na *Prawda* dla wszystkich rekordów przekraczających wartość odcięcia uwzględnioną w modelu. Dla każdego rekordu w osobnym polu podawany jest także indeks anomalii. Więcej informacji można znaleźć w temacie "Opcje modelu Wykrywanie anomalii" na stronie 58.
- **Tylko flaga** Tworzy pole flagi, ale bez podawania indeksu anomalii dla każdego rekordu.
- **Tylko indeks** Podaje indeks anomalii bez tworzenia flagi.

Liczba nieprawidłowych zmiennych w raporcie Określa liczbę zmiennych, która ma być podawana jako uzasadnienie oznaczenia konkretnego rekordu jako anomalii. Zgłaszane są najbardziej nietypowe zmienne, tj. te, które najbardziej odbiegają od normy danej zmiennej dla grupy, do której przypisany jest rekord.

Odrzuć takie rekordy Wybierz tę opcję, aby odrzucić ze strumienia wszystkie rekordy **Prawidłowe**. Dzięki temu łatwiej będzie skupić się na potencjalnych anomaliach w dalszych węzłach. Można też odrzucić wszystkie rekordy **Nietypowe**, aby ograniczyć dalszą analizę do rekordów nieoznaczonych na podstawie modelu jako potencjalne anomalie.

Uwaga: Ze względu na nieznaczne różnice w zaokrągleniach faktyczna liczba rekordów oznaczonych podczas oceniania może różnić się od liczby rekordów oznaczonych podczas uczenia modelu, nawet jeśli oba te procesy były realizowane na tych samych danych.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Rozdział 5. Zautomatyzowane węzły modelowania

Zautomatyzowane węzły modelowania estymują i porównują różne metody modelowania, co pozwala na wypróbowanie szeregu podejść w jednym przebiegu modelowania. Można wybrać algorytmy modelowania, które mają być używane, i konkretne opcje dla każdego algorytmu, w tym kombinacje, które w innych warunkach wykluczałyby się nawzajem. Na przykład zamiast dla sieci neuronowej wybierać między metodą szybką, dynamiczną albo przycinania, można wypróbować wszystkie te metody. Węzeł umożliwia eksplorację każdej możliwej kombinacji opcji, rangując każdy model kandydacki w oparciu o określoną przez użytkownika miarę, a następnie zapisuje najlepszy z nich do wykorzystania w ocenie lub do dalszej analizy.

Do wyboru są trzy zautomatyzowane węzły modelowania, które można stosować odpowiednio do potrzeb analitycznych:

Węzeł Auto Klasyfikacja tworzy i porównuje różne modele pod kątem wyników binarnych (tak lub nie, odejścia lub brak odejścia itd.), umożliwiając użytkownikowi wybór optymalnego podejścia do danej analizy. Obsługiwana jest pewna liczba algorytmów modelowania, co umożliwia wybór metod, które mają zostać użyte, konkretnych opcji dla każdej z nich oraz kryteriów porównywania wyników. Węzeł generuje zestaw modeli w oparciu o określone opcje i nadaje rangi najlepszym kandydatom wybranym według wskazanych kryteriów.

Węzeł Auto Predykcja estymuje i porównuje modele zwracające wyniki w formie ciągłego przedziału liczbowego, korzystając z szeregu różnych metod. Węzeł działa tak samo, jak węzeł Auto Klasyfikacja, umożliwiając użytkownikowi wybór używanych algorytmów oraz eksperymentowanie z wieloma kombinacjami opcji w pojedynczym przebiegu modelowania. Obsługiwane algorytmy obejmują sieci neuronowe, C&RT, CHAID, regresję liniową, uogólnioną regresję liniową oraz algorytmy SVM. Modele można porównywać na podstawie korelacji, błędu względnego lub liczby używanych zmiennych.

Węzeł Auto Grupowanie szacuje i porównuje modele skupień identyfikujące grupy rekordów o podobnej charakterystyce. Węzeł działa tak samo, jak pozostałe zautomatyzowane węzły modelowania, umożliwiając eksperymentowanie z wieloma kombinacjami opcji w pojedynczym przebiegu modelowania. Modele można porównywać, korzystając z miar bazowych, które pozwalają podejmować próby filtrowania i oceny przydatności modelu skupień oraz udostępniają miary bazujące na istotności poszczególnych zmiennych.

Najlepsze modele są zapisywane w jednym złożonym modelu użytkowym, dzięki czemu możliwe jest przeglądanie i porównywanie ich oraz wybór modeli do wykorzystania w ocenianiu.

- W przypadku binarnych, nominalnych i liczbowych zmiennych przewidywanych można wybrać wiele modeli do oceniania i połączyć wyniki w jeden, zespolony model użytkowy. Połączenie predykcji z wielu modeli umożliwia obejście ograniczeń w poszczególnych modelach, co często powoduje wyższą ogólną dokładność niż dokładność, jaką można uzyskać z dowolnego modelu.
- Opcjonalnie można analizować wyniki zstępująco i wygenerować węzły modelowania lub modele użytkowe dla dowolnych pojedynczych modeli, które chcemy zastosować lub dalej eksplorować.

Modele a czas wykonania

W zależności od zestawu danych i liczby modeli wykonywanie zautomatyzowanych węzłów modelowania może trwać wiele godzin lub nawet dłużej. Wybierając opcje, należy zwracać uwagę na liczbę generowanych modeli. O ile jest to możliwe ze względów organizacyjnych, celowe może być zaplanowanie uruchamiania modeli na godziny nocne lub weekendy, gdy zapotrzebowanie na zasoby obliczeniowe jest mniejsze.

- W razie potrzeby można zastosować węzeł podziału lub próby do ograniczenia liczby rekordów uwzględnionych przy pierwszym uczeniu. Po zawężeniu wyboru do kilku modeli kandydackich można wrócić do operowania na całym zbiorze danych.

- Aby ograniczyć liczbę zmiennych wejściowych, należy użyć funkcji Dobór predyktorów. Więcej informacji można znaleźć w temacie “Węzeł wyboru predyktora” na stronie 53. Można też wykorzystać pierwsze przebiegi modelu do ustalenia, które zmienne i opcje warto analizować dalej. Na przykład, jeśli najlepsze modele używają tych samych trzech zmiennych, to zdecydowanie wskazane jest zachowanie tych trzech zmiennych.
- Opcjonalnie można ograniczyć czas estymacji każdego z modeli i określić miary ewaluacyjne używane do monitorowania i rangowania modeli.

Ustawienia algorytmów zautomatyzowanych węzłów modelowania

Dla każdego typu modelu można użyć ustawień domyślnych lub wybrać opcje. Opcje są podobne do dostępnych w poszczególnych węzłach modelowania, z tym że zamiast wybierać tylko jedno ustawienie z kilku dostępnych, możemy wybrać dowolną liczbę ustawień obowiązujących w większości przypadków. Na przykład, porównując modele sieci neuronowych, można wybrać kilka różnych metod uczenia i wypróbować każdą z nich z wartością początkową generatora lub bez takiej wartości. Użyte zostaną wszystkie możliwe kombinacje wybranych opcji, dzięki czemu można bardzo łatwo wygenerować wiele różnych modeli w jednym przebiegu. Należy jednak zachować ostrożność, ponieważ wybranie wielu ustawień może doprowadzić do bardzo szybkiego zwielokrotnienia liczby modeli.

Aby wybrać opcje dla każdego typu modeli

1. W zautomatyzowanym węźle modelowania wybierz kartę **Zaawansowane**.
2. Kliknij kolumnę **Parametry modelu** dla typu modelu.
3. Z listy rozwijanej wybierz opcję **Określ**.
4. W oknie dialogowym **Ustawienia algorytmu** wybierz opcje z kolumny **Opcje**.

Uwaga: Dodatkowe opcje są dostępne na karcie Zaawansowane okna dialogowego **Ustawienia algorytmu**.

Reguły zatrzymujące zautomatyzowanego węzła modelowania

Reguły zatrzymujące określone dla zautomatyzowanych węzłów modelowania obowiązują względem wykonywania całego węzła, a nie działania poszczególnych modeli zbudowanych przez węzeł.

Ogranicz całkowity czas wykonania do. (tylko modele Sieci neuronowe, Metoda k-średnich, Sieć Kohonena, Dwustopniowa, SVM, KNN, Sieci Bayesa i C&RT) Zatrzymuje wykonanie po określonej liczbie godzin. Wszystkie modele wygenerowane do tego czasu zostaną uwzględnione w modelu użytkowym, ale nie będą już tworzone następne modele.

Zatrzymaj jak tylko utworzone zostaną ważne modele. Zatrzymuje wykonanie, gdy model spełni wszystkie kryteria określone na karcie Odrzuć (w przypadku węzła automatycznej klasyfikacji lub automatycznego grupowania) lub na karcie Model (w przypadku węzła automatycznej predykcji). Więcej informacji można znaleźć w temacie “Opcje odrzucania węzła Auto Klasyfikacja” na stronie 70. Więcej informacji można znaleźć w temacie “Opcje odrzucania węzła Auto Grupowanie” na stronie 78.

Węzeł Auto Klasyfikacja

Węzeł Auto Klasyfikacja estymuje i porównuje modele dla przewidywanych zmiennych nominalnych (zbiory) lub binarnych (tak/nie), stosując szereg różnych metod, co pozwala na wypróbowanie szeregu podejść w jednym przebiegu modelowania. Istnieje możliwość wyboru algorytmów, które mają być używane, oraz eksperymentowania z wieloma kombinacjami opcji. Na przykład zamiast dla modelu SVM wybierać między Radialną funkcją bazową, funkcją wielomianową, funkcją sigmoidalną lub funkcją liniową, można wypróbować wszystkie te metody. Węzeł umożliwia eksplorację każdej możliwej kombinacji opcji, rangując każdy model kandydacki w oparciu o określoną przez użytkownika miarę, a następnie zapisuje najlepsze modele do wykorzystania w ocenie lub do dalszej analizy. Aby uzyskać więcej informacji, patrz Rozdział 5, “Zautomatyzowane węzły modelowania”, na stronie 63.

Przykład

Załóżmy, że firma prowadząca handel detaliczny dysponuje danymi historycznymi o ofertach skierowanych do poszczególnych klientów w ramach wcześniejszych kampanii. Teraz firma chce uzyskać wyższy zysk, dobierając właściwą ofertę dla każdego klienta.

Wymagania

Zmienna przewidywana z poziomem pomiaru *Nominalna* albo *Flaga* (z rolą ustawioną na **Przewidywana**) i co najmniej jedna zmienna wejściowa (z rolą ustawioną na **Dane wejściowe**). Dla zmiennej typu flaga w obliczeniach zysku, wzrostu i pokrewnych statystyk przewidziana wartość *Prawda* oznacza trafienie. Zmienne wejściowe mogą mieć poziom pomiaru *Ilościowa* lub *Jakościowa*, przy zastrzeżeniu, że niektóre dane wejściowe mogą być nieodpowiednie w przypadku niektórych typów modeli. Na przykład zmienne porządkowe używane jako predyktory dla modeli C&RT, CHAID i QUEST muszą być zapisane w formie liczby (nie łańcucha), gdyż w przeciwnym razie będą przez te modele ignorowane. Podobnie, ilościowe zmienne wejściowe mogą być w niektórych przypadkach poddawane kategoryzacji. Wymagania są takie same, jak w przypadku użycia odrębnych węzłów modelowania; na przykład model sieci bayesowskiej działa tak samo niezależnie od tego, czy został wygenerowany z węzła Sieci Bayesa, czy z węzła Auto Klasyfikacja.

Zmienne częstości i ważąca.

Częstość i waga pozwalają nadać niektórym rekordom większe znaczenie niż innym, na przykład wówczas, kiedy użytkownik wie, że wbudowany zbiór danych nie zapewnia właściwej reprezentacji części populacji nadrzędnej (Waga) lub ponieważ jeden rekord reprezentuje pewną liczbę identycznych obserwacji (Częstość). W przypadku zaznaczenia tej opcji zmienna częstości może być wykorzystywane przez modele C&RT, CHAID, QUEST, Lista decyzyjna i Sieci Bayesa. Zmienna ważąca może być wykorzystywana przez modele C&RT, CHAID i C5.0. Inne typy modeli będą ignorować te zmienne, tworząc modele mimo to. Zmienne częstości i ważąca są używane tylko do tworzenia modeli i nie są uwzględniane podczas oceniania modeli. Aby uzyskać więcej informacji, patrz “Użycie zmiennych częstości i ważących” na stronie 33.

Prefiksy

W przypadku dołączenia węzła tabeli do modelu użytkowego dla węzła Auto Klasyfikacja tabeli dostępnych jest kilka nowych zmiennych o nazwach rozpoczynających się prefiksem $.

Nazwy zmiennych, które są wygenerowane podczas oceniania, są tworzone na podstawie zmiennej przewidywanej, ale dodawany jest standardowy przedrostek. Różne typy modeli używają różnych zestawów przedrostków.

Na przykład prefiksy $G, $R, $C są używane jako prefiksy dla predykcji generowanych odpowiednio przez modele Uogólniony liniowy, CHAID i C5.0. Przedrostek $X jest zwykle generowany w przypadku użycia zespołów, a przedrostki $XR, $XS i $XF są używane, w przypadku gdy zmienna przewidywana jest odpowiednio zmienną ilościową, jakościową lub zmienną typu flaga.

$..C są używane do predykcji ufności zmiennych przewidywanych Jakościowa lub Flaga; na przykład $XFC jest używana jako prefiks dla ufności predykcji zespołu Flaga. $RC i $CC to prefiksy dla pojedynczej predykcji ufności odpowiednio dla modelu CHAID i C5.0.

Obsługiwane typy modeli

Do obsługiwanych typów modeli należą: Sieci neuronowe, Drzewo C&R, QUEST, CHAID, C5.0, Regresja logistyczna, Lista decyzyjna, Sieć Bayesa, Analiza dyskryminacyjna, Najbliższy sąsiad, SVM, Drzewo XGBoost i XGBoost-AS. Więcej informacji można znaleźć w temacie “Opcje zaawansowane węzła Auto Klasyfikacja” na stronie 67.

Opcje modelu węzła Auto Klasyfikacja

Karta Model węzła Auto Klasyfikacja umożliwia określenie liczby modeli do utworzenia wraz z kryteriami używanymi do porównywania modeli.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Utwórz modele rozdzielone. Tworzy osobny model dla każdej możliwej wartości zmiennych wejściowych, jakie są określone jako zmienne podziału. Więcej informacji można znaleźć w “Budowanie modeli rozdzielonych” na stronie 28.

Ranguj modele według. Określa kryteria używane do porównywania i rangowania modeli. Do dostępnych opcji należy ogólna dokładność, pole pod krzywą ROC, zysk, wzrost i liczba zmiennych. Należy zwrócić uwagę, że wszystkie te miary będą dostępne w raporcie podsumowującym, niezależnie od tego, które z nich zostaną tutaj wybrane.

Uwaga: W przypadku nominalnej zmiennej przewidywanej (zbiór) rangowanie ograniczone jest do opcji **Ogólna dokładność** lub **Liczba zmiennych**.

Przy obliczaniu zysków, wzrostu i pokrewnych statystyk przyjmuje się, że wartość *Prawda* zmiennej przewidywanej oznacza trafienie.

- **Ogólna dokładność** Odsetek rekordów, które są poprawnie przewidywane przez model względem łącznej liczby rekordów.
- **Obszar pod krzywą ROC** Krzywa ROC jest wskaźnikiem wydajności modelu. Im wyżej nad linią odniesienia znajduje się krzywa, tym bardziej dokładny jest test.
- **Zysk (Skumulowane)** Suma zysków skumulowana percentylami (posortowanymi wg ufności predykcji) na podstawie określonych kryteriów kosztu, przychodów i wagi. Zwykle zysk zaczyna się w pobliżu 0 dla górnego percentyla, stabilnie rośnie, a potem się zmniejsza. W dobrym modelu krzywa zysku będzie miała wyraźnie zaznaczony szczyt, który zostanie podany wraz z percentylem, w którym wystąpił. W modelu, który nie dostarcza informacji, krzywa zysku będzie stosunkowo prosta i może wykazywać wzrost, spadek lub być pozioma, w zależności od obowiązującej struktury kosztów/przychodów.
- **Wzrost (Skumulowane)** Iloraz trafień w skumulowanych kwantylach względem całej próby (kwantyle posortowane są według ufności predykcji). Na przykład wzrost równy 3 w górnym kwantylu oznacza współczynnik trafień trzykrotnie wyższy niż w całej próbie. W dobrym modelu wzrost powinien zaczynać się wyraźnie powyżej wartości 1,0 w górnych kwantylach, a potem gwałtownie spadać do 1,0 w dolnych kwantylach. W modelu, który nie dostarcza informacji, wzrost będzie utrzymywał się w pobliżu wartości 1,0.
- **Liczba zmiennych** Nadaje rangi modelom na podstawie liczby używanych zmiennych wejściowych.

Ranguj modele wykorzystując. Jeśli używany jest podzbiór, możliwe jest określenie, czy rangi bazują na zbiorze danych uczących, czy na zbiorze testowym. W przypadku dużych zbiorów danych użycie podzbioru do wstępnego monitorowania modeli może znacząco poprawić wydajność.

Liczba modeli do wykorzystania. Określa maksymalną liczbę modeli do uwzględnienia w modelu użytkowym wygenerowanym przez węzeł. Modele rangowane najwyżej są wymienione zgodnie z określonym kryterium rangowania. Należy pamiętać, że zwiększenie tego limitu może spowodować spowolnienie działania. Maksymalna dozwolona wartość to 100.

Oblicz ważność predyktora. W przypadku modeli generujących odpowiednią miarę ważności możliwe jest wyświetlenie tabeli wskazującej ważność względną każdego predyktora w procesie estymacji modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Należy zwrócić uwagę, że ważność predyktora może wydłużyć czas potrzebny do obliczenia niektórych modeli i jej użycie nie jest zalecane w przypadku porównywania wielu różnych modeli. Jest to znacznie bardziej użyteczne w przypadku zawężenia analizy do niewielkiej liczby modeli, które mają być analizowane bardziej szczegółowo. Więcej informacji można znaleźć w temacie “Ważność predyktorów” na stronie 43.

Kryteria zysku. *Uwaga*. Tylko dla zmiennych przewidywanych typu flaga. Zysk równy jest przychodowi w każdym rekordzie pomniejszonemu o koszt w tym rekordzie. Zysk z kwantyla jest po prostu sumą zysków z wszystkich rekordów w tym kwantylu. Przyjmuje się, że zyski mają zastosowanie tylko do trafień, ale koszty — do wszystkich rekordów.

- **Koszty.** Pozwala określić koszty powiązane z poszczególnymi rekordami. Można wybrać przychód **Stały** lub **Zmienny**. W przypadku kosztów stałych należy określić wartość kosztu. W przypadku kosztów zmiennych należy kliknąć przycisk Selektor zmiennych, aby wybrać zmienną stanowiącą zmienną kosztu. (Opcja **Koszty** jest niedostępna dla wykresów ROC).
- **Przychód.** Określa przychód powiązany z poszczególnymi rekordami reprezentującymi trafienie. Można wybrać przychód **Stały** lub **Zmienny**. W przypadku przychodów stałych należy określić wartość przychodu. W przypadku przychodów zmiennych należy kliknąć przycisk Selektor zmiennych, aby wybrać zmienną stanowiącą zmienną przychodu. (Opcja **Przychód** jest niedostępna dla wykresów ROC).
- **Waga.** Jeśli rekordy w danych reprezentują więcej niż jedną jednostkę, można użyć wag częstości, aby skorygować wyniki. Należy określić wagę powiązaną z poszczególnymi rekordami, używając wag **Stała** lub **Zmienna**. W przypadku wag stałych należy określić wartość wagi (liczba jednostek na rekord). W przypadku wag zmiennych należy kliknąć przycisk Selektor zmiennych, aby wybrać zmienną stanowiącą zmienną wagi. (Opcja **Waga** jest niedostępna dla wykresów ROC).

Kryteria wzrostu. *Uwaga*. Tylko dla zmiennych przewidywanych typu flaga. Określa percentyl, który ma być używany w obliczeniach wzrostu. Tę wartość można zmienić także podczas porównywania wyników. Więcej informacji można znaleźć w temacie “Zautomatyzowane modele użytkowe” na stronie 79.

Opcje zaawansowane węzła Auto Klasyfikacja

Karta Zaawansowane węzła Auto Klasyfikacja umożliwia zastosowanie podzbioru (jeśli jest dostępny), wybór algorytmów i opcji, które mają być używane, oraz określenie reguł zatrzymujących.

Wybierz modele. Domyślnie wybierane są wszystkie modele; jednak w przypadku serwera Analytic Server można zdecydować o ograniczeniu modeli tylko do tych, które działają na serwerze Analytic Server, konfigurują je w taki sposób, aby tworzyły modele rozdzielone lub były gotowe na przetwarzanie bardzo dużych zbiorów danych.

Uwaga: Zbudowany lokalnie w węźle Auto Klasyfikacji model Analytic Server nie jest obsługiwany.

Używane modele. Zaznacz pola wyboru w kolumnach po lewej stronie typów modeli (algorytmów), które mają być uwzględnione w porównaniu. Im więcej typów zostanie wybranych, tym więcej zostanie utworzonych modeli i tym dłużej będzie trwało przetwarzanie.

Typ modelu. Lista dostępnych algorytmów (patrz niżej).

Parametry modelu. Dla każdego typu modelu można użyć ustawień domyślnych lub wybrać opcję **Określ** i samodzielnie określić opcje. Opcje są podobne do dostępnych w poszczególnych węzłach modelowania, z tym, że można wybrać dowolną liczbę ustawień obowiązujących w większości przypadków. Na przykład, porównując modele sieci neuronowych możemy, zamiast wybierać jedną z sześciu metod uczenia, wybrać je wszystkie i uczyć sześć modeli w jednym przebiegu.

Liczba modeli. Liczba modeli generowanych przez każdy z algorytmów na podstawie bieżących ustawień. Łączenie opcji może spowodować szybki wzrost liczby modeli, dlatego zdecydowanie zaleca się uważną obserwację tej liczby, zwłaszcza przy pracy z dużymi zbiorami danych.

Ogranicz maksymalny czas na budowanie jednego modelu. (Tylko modele Metoda k-średnich, Sieć Kohonena, Dwustopniowa, SVM, KNN, Sieci Bayesa i Lista decyzyjna) Określa limit czasu tworzenia jednego modelu. Na przykład, jeśli czas uczenia jednego konkretnego modelu jest nieprzewidywalny ze względu na pewne złożone interakcje, to nie chcemy, by ten model wstrzymywał cały przebieg modelowania.

Uwaga: Jeśli zmienna przewidywana jest nominalna (zbiór), to opcja Lista decyzyjna jest niedostępna.

Obsługiwane algorytmy

Węzeł SVM umożliwia szybką klasyfikację danych do jednej lub dwu grup bez przeuczenia. Algorytm SVM działa prawidłowo dla szerokiego zbioru danych, na przykład takiego o bardzo dużej liczbie zmiennych wejściowych.

Węzeł KNN (*k*-najbliższego sąsiedztwa) wiąże nową obserwację z kategorią lub wartością *k* (gdzie *k* jest liczbą całkowitą) najbliższych obiektów w przestrzeni predyktora. Podobne obserwacje znajdują się blisko siebie, a niepodobne — daleko.

Analiza dyskryminacyjna opiera się na ściślejszych założeniach niż regresja logistyczna, lecz może stanowić wartościową alternatywę lub uzupełnienie analizy metodą regresji logistycznej w przypadku spełnienia tych założeń.

Węzeł Sieci Bayesa umożliwia utworzenie modelu prawdopodobieństwa przez połączenie zaobserwowanych i zarejestrowanych dowodów z wiedzą rzeczywistą w celu ustanowienia prawdopodobieństwa występowania. Węzeł koncentruje się na sieciach Tree Augmented Naïve Bayes (TAN) i Markov Blanket, używanych głównie do klasyfikacji.

Węzeł Lista decyzyjna identyfikuje podgrupy lub segmenty wskazujące wyższe lub niższe prawdopodobieństwo danego wyniku binarnego względem całej populacji. Można na przykład wyszukać klientów, których prawdopodobieństwo odejścia jest niewielkie, lub którzy z dużym prawdopodobieństwem pozytywnie zareagują na kampanię. Istnieje możliwość zastosowania posiadanej wiedzy biznesowej w modelu przez dodanie własnych, niestandardowych segmentów i przejrzenie modeli alternatywnych jeden obok drugiego w celu porównania wyników. Modele Lista decyzyjna składają się z list reguł, w których każda reguła ma warunek i wynik. Reguły są stosowane w kolejności wprowadzania, a pierwsza reguła spełniona określa wynik.

Regresja logistyczna to technika statystyczna umożliwiająca klasyfikację rekordów na podstawie wartości zmiennych wejściowych. Jest ona analogiczna do regresji liniowej, lecz bazuje na przewidywanej zmiennej jakościowej zamiast na przedziale liczbowym.

Węzeł CHAID generuje drzewa decyzyjne, korzystając ze statystyk chi-kwadrat w celu identyfikacji optymalnych podziałów. W odróżnieniu od węzłów C&RT i węzłów QUEST, CHAID może generować drzewa niebinarne, co oznacza, że niektóre podziały mają więcej niż dwie gałęzie. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi (ilościowymi) lub jakościowymi. Wyczerpujący CHAID stanowi modyfikację CHAID umożliwiającą dokładniejsze badanie wszystkich możliwych podziałów, lecz obliczenia w jego przypadku zajmują więcej czasu.

Węzeł QUEST oferuje metodę klasyfikacji binarnej służącą do budowania drzew decyzyjnych, zaprojektowaną w celu redukcji czasu przetwarzania analiz dużych drzew decyzyjnych C&R, a jednocześnie w celu redukcji tendencji obecnej w metodach drzew klasyfikacji do preferowania danych wejściowych dopuszczających więcej podziałów. Zmienne wejściowe mogą być zakresami liczbowymi (ciągłymi), lecz zmienna przewidywana musi być jakościowa. Wszystkie podziały są binarne.

Węzeł klasyfikacji i regresji (C&RT) generuje drzewo decyzyjne umożliwiające predykcję lub klasyfikację przyszłych obserwacji. W metodzie tej stosowany jest rekursywny podział rekordów na segmenty przez minimalizację zanieczyszczeń w każdym kroku, przy czym węzeł w drzewie jest uważany za „czysty", jeśli 100% obserwacji w węźle przypada na konkretną kategorię zmiennej przewidywanej. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi lub jakościowymi (nominalnymi, porządkowymi lub flagami); wszystkie podziały są binarne (tylko dwie podgrupy).

Węzeł C5.0 tworzy drzewo decyzyjne lub zestaw reguł. Model działa w oparciu o podział próby na podstawie zmiennej oferującej maksimum korzyści z informacji na każdym z poziomów. Zmienna przewidywana musi być jakościowa. Dozwolonych jest wiele podziałów na więcej niż dwie podgrupy.

Węzeł Sieci neuronowe korzysta z uproszczonego modelu przetwarzania informacji przez ludzki umysł. Polega on na symulowaniu dużej liczby połączonych wzajemnie jednostek prostego przetwarzania, które przypominają abstrakcyjne wersje neuronów. Sieci neuronowe są estymatorami funkcji ogólnych o dużej wydajności, a do uczenia i stosowania ich wymagane jest tylko minimum wiedzy w zakresie statystyki lub matematyki.

Modele regresji liniowej przewidują docelową wartość ilościową na podstawie liniowych relacji między docelową wartością ilościową a jednym lub większą liczbą predyktorów.

Węzeł LSVM umożliwia klasyfikację danych do jednej lub dwu grup bez przeuczenia. Algorytm LSVM jest liniowy i działa prawidłowo z szerokimi zbiorami danych, na przykład zbiorami o bardzo dużej liczbie rekordów.

Węzeł Drzewa losowe jest podobny do istniejącego węzła C&RT; jednak węzeł Drzewa losowe jest przeznaczony do przetwarzania dużych zbiorów danych w celu utworzenia pojedynczego drzewa i wyświetla model wynikowy w przeglądarce wyników, która została dodana w programie SPSS Modeler, wersja 17. Węzeł Drzewa losowe generuje drzewo decyzyjne umożliwiające predykcję lub klasyfikację przyszłych obserwacji. W metodzie tej stosowany jest rekursywny podział rekordów na segmenty przez minimalizację zanieczyszczeń w każdym kroku, przy czym węzeł w drzewie jest uważany za *czysty*, jeśli 100% obserwacji w węźle przypada na konkretną kategorię zmiennej przewidywanej. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi lub jakościowymi (nominalnymi, porządkowymi lub flagami); wszystkie podziały są binarne (tylko dwie podgrupy).

Węzeł Drzewo-AS jest podobny do istniejącego węzła CHAID; jednak węzeł Drzewo-AS jest przeznaczony do przetwarzania dużych zbiorów w celu utworzenia pojedynczego drzewa i wyświetla model wynikowy w przeglądarce wyników, która została dodana w programie SPSS Modeler, wersja 17. Węzeł generuje drzewo decyzyjne używając statystyki chi-kwadrat (CHAID), aby określić optymalne podziały. CHAID może generować drzewa niebinarne, co oznacza, że niektóre podziały mają więcej niż dwie gałęzie. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi (ilościowymi) lub jakościowymi. Wyczerpujący CHAID stanowi modyfikację CHAID umożliwiającą dokładniejsze badanie wszystkich możliwych podziałów, lecz obliczenia w jego przypadku zajmują więcej czasu.

XGBoost Tree© to zaawansowana implementacja algorytmu wzmacniania gradientowego, który jako model bazowy wykorzystuje model drzewa. Algorytmy wzmacniania iteracyjnie ucząc się, wyznaczają słabe klasyfikatory i dodają je do ostatecznego silnego klasyfikatora. XGBoost Tree jest algorytmem bardzo elastycznym i oferuje liczne parametry, które mogą być trudne do praktycznego wykorzystania przez użytkowników. Dlatego węzeł Drzewo XGBoost w programie SPSS Modeler eksponuje tylko funkcje podstawowe i najczęściej używane parametry. Węzeł jest zaimplementowany w języku Python.

XGBoost© to zaawansowana implementacja algorytmu wzmacniania gradientowego. Algorytmy wzmacniania iteracyjnie ucząc się, wyznaczają słabe klasyfikatory i dodają je do ostatecznego silnego klasyfikatora. XGBoost jest algorytmem bardzo elastycznym i oferuje liczne parametry, które mogą być trudne do praktycznego wykorzystania przez większość użytkowników. Dlatego węzeł XGBoost-AS w programie SPSS Modeler eksponuje tylko funkcje podstawowe i najczęściej używane parametry. Węzeł XGBoost-AS jest zaimplementowany w środowisku Spark.

Uwaga: W przypadku uruchomienia Drzewa-AS w programie Analytic Server budowa modelu w poprzedzającym węźle podziału na podzbiory zakończy się niepowodzeniem. W takim wypadku, aby umożliwić działanie Auto Klasyfikacji z innymi węzłami modelowania, należy odznaczyć typ modelu Drzewo-AS.

Koszty błędnej klasyfikacji

W niektórych kontekstach pewne błędy są bardziej kosztowne od innych. Przykładowo, bardziej kosztowne może być sklasyfikowanie osób składających wniosek kredytowy z wysokim poziomem ryzyka jako osób z niskim poziomem ryzyka (jeden rodzaj błędu) niż sklasyfikowanie osób składających wniosek z niskim poziomem ryzyka jako osób z wysokim poziomem ryzyka (inny rodzaj błędu). Koszty błędnej klasyfikacji umożliwiają określenie względnej ważności różnych rodzajów błędów predykcji.

Kosztami błędnej klasyfikacji zwykle są wagi zastosowane do określonych danych wynikowych. Wagi te są uwzględniane w modelu i rzeczywiście mogą zmienić predykcję (jako sposób ochrony przed kosztownymi błędami).

Z wyjątkiem modeli C5.0 koszty błędnej klasyfikacji nie mają zastosowania podczas oceniania modelu i nie są brane pod uwagę podczas rangowania lub porównywania modeli za pomocą węzła Auto Klasyfikacja, wykresu ewaluacyjnego lub węzła analizy. Model, który uwzględnia koszty, może nie wygenerować mniejszej liczby błędów niż ten, który ich nie uwzględnia, i może nie mieć wyższej rangi pod względem całkowitej dokładności, ale prawdopodobnie lepiej sprawdzi się w warunkach praktycznych, ponieważ generuje błędy *mniej kosztowne*.

Macierz kosztów przedstawia koszty dla każdej możliwej kombinacji przewidywanej kategorii rzeczywistej. Domyślnie wszystkie koszty błędnej klasyfikacji są ustawione na wartość 1,0. Aby wprowadzić niestandardowe wartości kosztów, należy wybrać opcję **Stosuj koszty błędnej klasyfikacji** i wprowadzić do macierzy kosztów niestandardowe wartości.

Aby zmienić koszt błędnej klasyfikacji, należy zaznaczyć komórkę odpowiadającą odpowiedniej kombinacji wartości przewidywanych i rzeczywistych, usunąć istniejącą zawartość komórki i wprowadzić do niej żądany koszt. Koszty nie są automatycznie symetryczne. Przykładowo, jeśli koszt błędnej klasyfikacji *A* jako *B* zostanie ustawiony na 2,0, koszt błędnej klasyfikacji *B* jako *A* nadal będzie miał domyślną wartość 1,0, chyba że zostanie ona również jawnie zmieniona.

Opcje odrzucania węzła Auto Klasyfikacja

Karta Odrzuć węzła Auto Klasyfikacja umożliwia automatyczne odrzucanie modeli, które nie spełniają określonych kryteriów. Modele te nie będą wymienione w raporcie podsumowującym.

Można określić próg minimalnej dokładności ogólnej i próg maksymalnej liczby zmiennych używanych w modelu. Ponadto dla zmiennych przewidywanych typu flaga można określić próg minimalnego wzrostu, zysku i obszaru pod krzywą; wzrost i zysk określane są tak, jak określono to na karcie Model. Więcej informacji można znaleźć w temacie “Opcje modelu węzła Auto Klasyfikacja” na stronie 65.

Opcjonalnie można skonfigurować węzeł tak, by wykonywanie było zatrzymywane po wygenerowaniu pierwszego modelu spełniającego wszystkie określone kryteria. Więcej informacji można znaleźć w temacie “Reguły zatrzymujące zautomatyzowanego węzła modelowania” na stronie 64.

Opcje ustawień węzła Auto Klasyfikacja

Karta Ustawienia węzła Auto Klasyfikacja umożliwia wstępną konfigurację opcji ocena-czas dostępnych w modelu użytkowym.

Odfiltruj zmienne utworzone przez modele zespolone Usuwa z danych wynikowych wszystkie dodatkowe zmienne wygenerowane przez poszczególne modele zasilające węzeł zespolenia. To pole wyboru należy zaznaczyć w przypadku zainteresowania tylko oceną zespoloną wszystkich modeli wejściowych. Upewnij się, że ta opcja nie jest zaznaczona, jeśli na przykład, chcesz użyć trybu Analiza lub Ewaluacja do porównania dokładności oceny zespolonej z oceną poszczególnych modeli wejściowych.

Węzeł Auto Predykcja

Węzeł Auto Predykcja estymuje i porównuje modele dla wyników ilościowych przedziału liczbowego, korzystając z szeregu różnych metod, co pozwala na wypróbowanie szeregu podejść w jednym przebiegu modelowania. Istnieje możliwość wyboru algorytmów, które mają być używane, oraz eksperymentowania z wieloma kombinacjami opcji. Możliwa jest na przykład predykcja wartości składowych z użyciem sieci neuronowych, regresji liniowej, modelu C&RT lub modelu CHAID w celu stwierdzenia, który z nich sprawdza się najlepiej. Możliwe jest także wypróbowanie różnych kombinacji metod regresji krokowej, postępującej i wstecznej. Węzeł umożliwia eksplorację każdej możliwej kombinacji opcji, rangując każdy model kandydacki w oparciu o określoną przez użytkownika miarę, a następnie zapisuje najlepszy z nich do wykorzystania w ocenie lub do dalszej analizy. Więcej informacji można znaleźć w temacie Rozdział 5, “Zautomatyzowane węzły modelowania”, na stronie 63.

Przykład

Zarząd miasta potrzebuje dokładniejszych oszacowań podatku od nieruchomości oraz możliwości dostosowania wartości dla konkretnych nieruchomości odpowiednio do potrzeb, bez konieczności sprawdzania każdej z nich. Korzystając z węzła Auto Predykcja, analityk może generować i porównywać szereg modeli predykcyjnych wartości nieruchomości w oparciu o rodzaj budynku, sąsiedztwo, wielkość i inne znane czynniki.

Wymagania

Pojedyncza zmienna przewidywana (dla której jako rolę ustawiono **Przewidywana**), oraz co najmniej jedna zmienna wejściowa (dla której jako rolę ustawiono **Dane wejściowe**). Zmienna przewidywana musi być zmienną ilościową (przedziałem liczbowym), taką jak *wiek* czy *przychód*. Zmienne wejściowe mogą być ilościowe lub jakościowe, przy zastrzeżeniu, że niektóre dane wejściowe mogą być nieodpowiednie w przypadku niektórych typów modeli. Na przykład modele C&RT mogą korzystać z jakościowych zmiennych łańcuchowych jako danych wejściowych, podczas gdy modele regresji liniowej nie mogą używać takich zmiennych i będą je ignorować. Wymagania są takie same jak w przypadku korzystania z indywidualnych węzłów modelowania. Na przykład model CHAID działa tak samo niezależnie od tego, czy został wygenerowany na podstawie węzła CHAID, czy na podstawie węzła Auto Predykcja.

Zmienne częstości i ważąca.

Częstość i waga pozwalają nadać niektórym rekordom większe znaczenie niż innym, na przykład wówczas, kiedy użytkownik wie, że wbudowany zbiór danych nie zapewnia właściwej reprezentacji części populacji nadrzędnej (Waga) lub ponieważ jeden rekord reprezentuje pewną liczbę identycznych obserwacji (Częstość). W przypadku zaznaczenia tej opcji zmienna częstości może być wykorzystywana przez algorytmy C&RT oraz CHAID. Zmienna ważąca może być wykorzystywana przez algorytmy C&RT, CHAID, regresję i Modele uogólnione. Inne typy modeli będą ignorować te zmienne, tworząc modele mimo to. Zmienne częstości i ważąca są używane tylko do tworzenia modeli i nie są uwzględniane podczas oceniania modeli. Więcej informacji można znaleźć w temacie “Użycie zmiennych częstości i ważących” na stronie 33.

Prefiksy

W przypadku dołączenia węzła tabeli do modelu użytkowego dla węzła Auto Predykcja tabeli dostępnych jest kilka nowych zmiennych o nazwach rozpoczynających się prefiksem $.

Nazwy zmiennych, które są wygenerowane podczas oceniania, są tworzone na podstawie zmiennej przewidywanej, ale dodawany jest standardowy przedrostek. Różne typy modeli używają różnych zestawów przedrostków.

Na przykład prefiksy $G, $R, $C są używane jako prefiksy dla predykcji generowanych odpowiednio przez modele Uogólniony liniowy, CHAID i C5.0. Przedrostek $X jest zwykle generowany w przypadku użycia zespołów, a przedrostki $XR, $XS i $XF są używane, w przypadku gdy zmienna przewidywana jest odpowiednio zmienną ilościową, jakościową lub zmienną typu flaga.

$..E są używane do predykcji ufności docelowych wartości ilościowych zmiennej; na przykład $XRE stanowi prefiks dla ufności predykcji zespołu Zmienna ilościowa. $GE to prefiks dla pojedynczej predykcji ufności dla uogólnionego modelu liniowego.

Obsługiwane typy modeli

Do obsługiwanych typów modeli należą: Sieci neuronowe, Drzewo C&R, CHAID, Regresja, Modele uogólnione, Najbliższy sąsiad, SVM, Liniowy XGBoost, GLE i XGBoost-AS. Aby uzyskać więcej informacji, patrz “Opcje zaawansowane węzła Auto Predykcja” na stronie 73.

Opcje modelu węzła Auto Predykcja

Karta Model węzła Auto Predykcja umożliwia określenie liczby modeli do zapisania wraz z kryteriami używanymi do porównywania modeli.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Utwórz modele rozdzielone. Tworzy osobny model dla każdej możliwej wartości zmiennych wejściowych, jakie są określone jako zmienne podziału. Więcej informacji można znaleźć w “Budowanie modeli rozdzielonych” na stronie 28.

Ranguj modele według. Określa kryteria używane do porównywania modeli.

- **Korelacja.** Korelacja Pearsona między obserwowaną wartością każdego rekordu a wartością predykowaną przez model. Korelacja stanowi miarę powiązania liniowego między dwiema zmiennymi, przy czym wartości bliższe 1 oznaczają relację silniejszą. (Wartości korelacji wahają się między –1 w przypadku związku idealnie ujemnego a +1 w przypadku związku idealnie dodatniego. Wartość 0 oznacza brak związku liniowego, zaś model o korelacji ujemnej będzie miał rangę najniższą ze wszystkich).
- **Liczba zmiennych.** Liczba zmiennych używanych jako predyktory w modelu. Wybór modeli o mniejszej liczbie zmiennych może uprościć proces przygotowania danych oraz, w niektórych przypadkach, poprawić wydajność.
- **Błąd względny.** Błąd względny stanowi iloraz wariancji obserwowanych wartości z wartości predykowanych przez model względem wariancji wartości obserwowanych ze średniej. Praktycznie rzecz biorąc, porównuje on, jak dobrze model sprawdza się względem modelu **null** lub **wyraz wolny**, zwracającego po prostu średnią wartość zmiennej przewidywanej jako predykcję. W przypadku prawidłowego modelu wartość ta powinna być niższa od 1, co będzie oznaczać, że model jest dokładniejszy od modelu null. Model o błędzie względnym większym niż 1 jest mniej dokładny niż model null i dlatego nie jest przydatny. W przypadku modeli regresji liniowej błąd względny jest równy kwadratowi korelacji i nie niesie ze sobą żadnych nowych informacji. W przypadku modeli nieliniowych błąd względny jest niepowiązany z korelacją i zapewnia dodatkową miarę umożliwiającą ocenę wydajności modelu.

Ranguj modele wykorzystując. Jeśli używany jest podzbiór, możliwe jest określenie, czy rangi bazują na podzbiorze uczenia, czy na podzbiorze testowym. W przypadku dużych zbiorów danych użycie podzbioru do wstępnego monitorowania modeli może znacząco poprawić wydajność.

Liczba modeli do wykorzystania. Określa maksymalną liczbę modeli do wyświetlenia w modelu użytkowym wygenerowanym przez węzeł. Modele rangowane najwyżej są wymienione zgodnie z określonym kryterium rangowania. Zwiększenie tego limitu umożliwi porównanie wyników dla wielu modeli, lecz może zmniejszyć wydajność. Maksymalna dozwolona wartość to 100.

Oblicz ważność predyktora. W przypadku modeli generujących odpowiednią miarę ważności możliwe jest wyświetlenie tabeli wskazującej ważność względną każdego predyktora w procesie estymacji modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Należy zwrócić uwagę, że ważność predyktora może wydłużyć czas potrzebny do obliczenia niektórych modeli i jej użycie nie jest zalecane w przypadku porównywania wielu różnych modeli. Jest to znacznie bardziej użyteczne w przypadku zawężenia analizy do niewielkiej liczby modeli, które mają być analizowane bardziej szczegółowo. Więcej informacji można znaleźć w temacie “Ważność predyktorów” na stronie 43.

Nie zachowuj modeli, jeśli. Określa wartości progów korelacji, błąd względny oraz liczbę używanych zmiennych. Modele niespełniające żadnego z tych kryteriów będą odrzucane i nie będą uwzględniane w raporcie podsumowującym.

- **Korelacja jest mniejsza niż.** Korelacja minimalna (wyrażona wartością bezwzględną) dla modelu do uwzględnienia w raporcie podsumowującym.
- **Liczba zmiennych jest większa niż.** Maksymalna liczba zmiennych, jaka ma być używana przez dowolny uwzględniany model.
- **Względny błąd jest większy niż.** Maksymalny błąd względny dla dowolnego uwzględnianego modelu.

Opcjonalnie można skonfigurować węzeł tak, by wykonywanie było zatrzymywane po wygenerowaniu pierwszego modelu spełniającego wszystkie określone kryteria. Więcej informacji można znaleźć w temacie “Reguły zatrzymujące zautomatyzowanego węzła modelowania” na stronie 64.

Opcje zaawansowane węzła Auto Predykcja

Karta Zaawansowany węzła Auto Predykcja umożliwia wybór algorytmów i opcji, które mają być używane, oraz określenie reguł zatrzymujących.

Wybierz modele. Domyślnie wybierane są wszystkie modele; jednak w przypadku serwera Analytic Server można zdecydować o ograniczeniu modeli tylko do tych, które działają na serwerze Analytic Server, konfigurują je w taki sposób, aby tworzyły modele rozdzielone lub były gotowe na przetwarzanie bardzo dużych zbiorów danych.

Uwaga: Zbudowany lokalnie w węźle Auto Predykcji model Analytic Server nie jest obsługiwany.

Używane modele. Zaznacz pola wyboru w kolumnach po lewej stronie typów modeli (algorytmów), które mają być uwzględnione w porównaniu. Im więcej typów zostanie wybranych, tym więcej zostanie utworzonych modeli i tym dłużej będzie trwało przetwarzanie.

Typ modelu. Lista dostępnych algorytmów (patrz niżej).

Parametry modelu. Dla każdego typu modelu można użyć ustawień domyślnych lub wybrać opcję **Określ** i samodzielnie określić opcje. Opcje są podobne do dostępnych w poszczególnych węzłach modelowania, z tym, że można wybrać dowolną liczbę ustawień obowiązujących w większości przypadków. Na przykład, porównując modele sieci neuronowych możemy, zamiast wybierać jedną z sześciu metod uczenia, wybrać je wszystkie i uczyć sześć modeli w jednym przebiegu.

Liczba modeli. Liczba modeli generowanych przez każdy z algorytmów na podstawie bieżących ustawień. Łączenie opcji może spowodować szybki wzrost liczby modeli, dlatego zdecydowanie zaleca się uważną obserwację tej liczby, zwłaszcza przy pracy z dużymi zbiorami danych.

Ogranicz maksymalny czas na budowanie jednego modelu. (Tylko modele Metoda k-średnich, Sieć Kohonena, Dwustopniowa, SVM, KNN, Sieci Bayesa i Lista decyzyjna) Określa limit czasu tworzenia jednego modelu. Na

przykład, jeśli czas uczenia jednego konkretnego modelu jest nieprzewidywalny ze względu na pewne złożone interakcje, to nie chcemy, by ten model wstrzymywał cały przebieg modelowania.

Obsługiwane algorytmy

Węzeł Sieci neuronowe korzysta z uproszczonego modelu przetwarzania informacji przez ludzki umysł. Polega on na symulowaniu dużej liczby połączonych wzajemnie jednostek prostego przetwarzania, które przypominają abstrakcyjne wersje neuronów. Sieci neuronowe są estymatorami funkcji ogólnych o dużej wydajności, a do uczenia i stosowania ich wymagane jest tylko minimum wiedzy w zakresie statystyki lub matematyki.

Węzeł klasyfikacji i regresji (C&RT) generuje drzewo decyzyjne umożliwiające predykcję lub klasyfikację przyszłych obserwacji. W metodzie tej stosowany jest rekursywny podział rekordów na segmenty przez minimalizację zanieczyszczeń w każdym kroku, przy czym węzeł w drzewie jest uważany za „czysty”, jeśli 100% obserwacji w węźle przypada na konkretną kategorię zmiennej przewidywanej. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi lub jakościowymi (nominalnymi, porządkowymi lub flagami); wszystkie podziały są binarne (tylko dwie podgrupy).

Węzeł CHAID generuje drzewa decyzyjne, korzystając ze statystyk chi-kwadrat w celu identyfikacji optymalnych podziałów. W odróżnieniu od węzłów C&RT i węzłów QUEST, CHAID może generować drzewa niebinarne, co oznacza, że niektóre podziały mają więcej niż dwie gałęzie. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi (ilościowymi) lub jakościowymi. Wyczerpujący CHAID stanowi modyfikację CHAID umożliwiającą dokładniejsze badanie wszystkich możliwych podziałów, lecz obliczenia w jego przypadku zajmują więcej czasu.

Regresja liniowa to typowa technika statystyczna umożliwiająca podsumowanie danych i przewidywanie poprzez dopasowanie do linii prostej lub powierzchni, co powoduje zminimalizowane rozbieżności pomiędzy przewidywanymi a rzeczywistymi wartościami zmiennych wyjściowych.

Węzeł Modele uogólnione rozszerza ogólny model liniowy w taki sposób, że zmienna zależna jest liniowo powiązana z czynnikami i współzmiennymi za pośrednictwem określonej funkcji łączenia. Model pozwala ponadto, aby zmienna zależna nie miała rozkładu normalnego. Obejmuje ona funkcjonalność dużej liczby modeli statystycznych, m.in. regresji liniowej, regresji logistycznej, modeli logarytmiczno-liniowych dla danych o liczebności.

Węzeł KNN (*k*-najbliższego sąsiedztwa) wiąże nową obserwację z kategorią lub wartością *k* (gdzie *k* jest liczbą całkowitą) najbliższych obiektów w przestrzeni predyktora. Podobne obserwacje znajdują się blisko siebie, a niepodobne — daleko.

Węzeł SVM umożliwia szybką klasyfikację danych do jednej lub dwu grup bez przeuczenia. Algorytm SVM działa prawidłowo dla szerokiego zbioru danych, na przykład takiego o bardzo dużej liczbie zmiennych wejściowych.

Modele regresji liniowej przewidują docelową wartość ilościową na podstawie liniowych relacji między docelową wartością ilościową a jednym lub większą liczbą predyktorów.

Węzeł LSVM umożliwia klasyfikację danych do jednej lub dwu grup bez przeuczenia. Algorytm LSVM jest liniowy i działa prawidłowo z szerokimi zbiorami danych, na przykład zbiorami o bardzo dużej liczbie rekordów.

Węzeł Drzewa losowe jest podobny do istniejącego węzła C&RT; jednak węzeł Drzewa losowe jest przeznaczony do przetwarzania dużych zbiorów danych w celu utworzenia pojedynczego drzewa i wyświetla model wynikowy w przeglądarce wyników, która została dodana w programie SPSS Modeler, wersja 17. Węzeł Drzewa losowe generuje drzewo decyzyjne umożliwiające predykcję lub klasyfikację przyszłych obserwacji. W metodzie tej stosowany jest rekursywny podział rekordów na segmenty przez minimalizację zanieczyszczeń w każdym kroku, przy czym węzeł w drzewie jest uważany za *czysty*, jeśli 100% obserwacji w węźle przypada na konkretną kategorię zmiennej przewidywanej. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi lub jakościowymi (nominalnymi, porządkowymi lub flagami); wszystkie podziały są binarne (tylko dwie podgrupy).

Węzeł Drzewo-AS jest podobny do istniejącego węzła CHAID; jednak węzeł Drzewo-AS jest przeznaczony do przetwarzania dużych zbiorów w celu utworzenia pojedynczego drzewa i wyświetla model wynikowy w przeglądarce wyników, która została dodana w programie SPSS Modeler, wersja 17. Węzeł generuje drzewo decyzyjne używając statystyki chi-kwadrat (CHAID), aby określić optymalne podziały. CHAID może generować drzewa niebinarne, co oznacza, że niektóre podziały mają więcej niż dwie gałęzie. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi (ilościowymi) lub jakościowymi. Wyczerpujący CHAID stanowi modyfikację CHAID umożliwiającą dokładniejsze badanie wszystkich możliwych podziałów, lecz obliczenia w jego przypadku zajmują więcej czasu.

XGBoost Linear© to zaawansowana implementacja algorytmu wzmacniania gradientowego, który jako model bazowy wykorzystuje model liniowy. Algorytmy wzmacniania iteracyjnie ucząc się, wyznaczają słabe klasyfikatory i dodają je do ostatecznego silnego klasyfikatora. Węzeł Liniowy XGBoost w programie SPSS Modeler jest zaimplementowany w języku Python.

GLE stanowi wersję modelu liniowego rozszerzoną w taki sposób, że zmienna przewidywana może mieć rozkład inny niż normalny, jest liniowo powiązana z czynnikami i współzmiennymi za pośrednictwem określonej funkcji łączenia, a obserwacje mogą być skorelowane. Uogólnione liniowe modele mieszane obejmują szeroki wachlarz modeli, począwszy od prostych modeli regresji liniowej, aż po złożone wielopoziomowe modele dla danych z obserwacji długofalowych nieposiadających rozkładu normalnego.

XGBoost© to zaawansowana implementacja algorytmu wzmacniania gradientowego. Algorytmy wzmacniania iteracyjnie ucząc się, wyznaczają słabe klasyfikatory i dodają je do ostatecznego silnego klasyfikatora. XGBoost jest algorytmem bardzo elastycznym i oferuje liczne parametry, które mogą być trudne do praktycznego wykorzystania przez większość użytkowników. Dlatego węzeł XGBoost-AS w programie SPSS Modeler eksponuje tylko funkcje podstawowe i najczęściej używane parametry. Węzeł XGBoost-AS jest zaimplementowany w środowisku Spark.

Opcje ustawień węzła Auto Predykcja

Karta Ustawienia węzła Auto Predykcja umożliwia wstępną konfigurację opcji ocena-czas dostępnych w modelu użytkowym.

Odfiltruj zmienne utworzone przez modele zespolone Usuwa z danych wynikowych wszystkie dodatkowe zmienne wygenerowane przez poszczególne modele zasilające węzeł zespolenia. To pole wyboru należy zaznaczyć w przypadku zainteresowania tylko oceną zespoloną wszystkich modeli wejściowych. Upewnij się, że ta opcja nie jest zaznaczona, jeśli na przykład, chcesz użyć trybu Analiza lub Ewaluacja do porównania dokładności oceny zespolonej z oceną poszczególnych modeli wejściowych.

Oblicz błąd standardowy. W przypadku ilościowej zmiennej przewidywanej (zakresu liczbowego) obliczenia błędu standardowego są uruchamiane domyślnie w celu obliczenia różnicy między wartością zmierzoną lub estymowaną a rzeczywistą, a także do prezentacji stopnia dopasowania tych estymacji.

Węzeł Auto Grupowanie

Węzeł Auto Grupowanie szacuje i porównuje modele skupień identyfikujące grupy rekordów o podobnej charakterystyce. Węzeł działa tak samo, jak pozostałe zautomatyzowane węzły modelowania, umożliwiając eksperymentowanie z wieloma kombinacjami opcji w pojedynczym przebiegu modelowania. Modele można porównywać, korzystając z miar bazowych, które pozwalają podejmować próby filtrowania i oceny przydatności modelu skupień oraz udostępniają miary bazujące na istotności poszczególnych zmiennych.

Modele skupień są często używane do identyfikacji grup, które mogą być używane jako dane wejściowe do dalszej analizy. Załóżmy, że chcemy skierować ofertę do grup klientów na podstawie ich cech demograficznych, takich jak dochód, lub na podstawie usług, które kupowali w przeszłości. Możemy to zrobić nie dysponując wcześniej żadną wiedzą o tych grupach i ich cechach — możliwe nawet, że nie wiadomo, ilu grup szukać i jakich predyktorów używać do ich definiowania. Modele skupień nazywane są często modelami uczenia nienadzorowanego, ponieważ nie korzystają ze zmiennej przewidywanej i nie zwracają konkretnej predykcji o wartości prawda albo fałsz. Wartość modelu skupień określana jest przez jego zdolność do wykrywania interesujących skupień w danych i oferowania użytecznych opisów tych skupień. Więcej informacji można znaleźć w Rozdział 11, “Modele skupień”, na stronie 235.

Wymagania. Jedna lub więcej zmiennych definiujących interesujące nas cechy. W modelach skupień nie są używane zmienne przewidywane, tak jak w innych modelach, ponieważ modele skupień nie generują konkretnych predykcji o wartości prawda albo fałsz. Służą one natomiast do wykrywania grup obserwacji, które mogą być ze sobą powiązane. Na przykład nie można użyć modelu skupień do przewidzenia, czy dany klient odejdzie lub odpowie na ofertę. Ale można użyć takiego modelu do przypisywania klientów do grup na podstawie ich tendencji do podejmowania takich decyzji. Zmienne wagi i częstości nie są używane.

Zmienne ewaluacyjne Mimo że nie jest używana zmienna przewidywana, można opcjonalnie określić jedną lub wiele zmiennych ewaluacyjnych, które posłużą do porównywania modeli. Użyteczność modelu skupień można ocenić, mierząc, na ile dobrze (lub źle) skupienia różnicują te zmienne.

Obsługiwane typy modeli

Do obsługiwanych typów modeli należą: Dwustopniowa, K-średnie, Kohonen, SVM z jedną klasą i K-średnie-AS.

Opcje modelu węzła Auto Grupowanie

Karta Model węzła Auto Grupowanie umożliwia określenie liczby modeli do zapisania wraz z kryteriami używanymi do porównywania modeli.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Ranguj modele według. Określa kryteria używane do porównywania i rangowania modeli.

- **Silhouette.** Wskaźnik spójności i odrębności modelu. Więcej informacji zawiera sekcja *Miara rangowania Silhouette* (poniżej).
- **Liczba grup.** Liczba grup w modelu.
- **Wielkość najmniejszej grupy.** Wielkość najmniejszej grupy.
- **Wielkość największej grupy.** Wielkość największej grupy.
- **Najmniejsza/największa grupa.** Iloraz wielkości najmniejszej grupy i największej grupy.

- **Ważność.** Ważność zmiennej **Ewaluacja** na karcie **Zmienne**. Należy zauważyć, że ważność można obliczyć tylko jeśli określono zmienną **Ewaluacja**.

Ranguj modele wykorzystując. Jeśli używany jest podzbiór, możliwe jest określenie, czy rangi bazują na zbiorze danych uczących, czy na zbiorze testowym. W przypadku dużych zbiorów danych użycie podzbioru do wstępnego monitorowania modeli może znacząco poprawić wydajność.

Liczba modeli do zachowania. Określa maksymalną liczbę modeli do wyświetlenia w modelu użytkowym wygenerowanym przez węzeł. Modele rangowane najwyżej są wymienione zgodnie z określonym kryterium rangowania. Należy pamiętać, że zwiększenie tego limitu może spowodować spowolnienie działania. Maksymalna dozwolona wartość to 100.

Miara rangowania Silhouette

Domyślna miara rangowania, Silhouette, ma wartość domyślną 0, ponieważ wartość mniejsza od 0 (tj. ujemna) wskazuje, że średnia odległość między obserwacją a punktami w przypisanej jej grupie jest większa od minimalnej średniej odległości od punktów w innej grupie. Dlatego modele z ujemną miarą Silhouette można bezpiecznie odrzucić.

Ta miara rangowania jest faktycznie zmodyfikowanym współczynnikiem Silhouette, który łączy pojęcie spójności grupy (promujące modele zawierające blisko położone grupy) i odseparowania grupy (promujące modele zawierające odległe od siebie grupy). Średni współczynnik Silhouette jest po prostu wyciągniętą ze wszystkich obserwacji średnią wartości obliczonych (dla każdej obserwacji z osobna) wg wzoru:

$(B - A) / \max(A, B)$

gdzie A jest odległością obserwacji od środka ciężkości grupy, do którego obserwacja należy; a B jest minimalną odległością między obserwacją a środkami ciężkości pozostałych grup.

Współczynnik Silhouette (i jego średnia) przyjmuje wartości z przedziału od -1 (bardzo słaby model) do 1 (doskonały model). Średnią można obliczać na poziomie sumy obserwacji (uzyskując łączną miarę Silhouette) lub na poziomie grup (uzyskując miarę Silhouette grupy). Odległości można obliczać jako euklidesowe.

Opcje zaawansowane węzła Auto Grupowanie

Karta Zaawansowane węzła Auto Grupowanie umożliwia zastosowanie podzbioru (jeśli jest dostępny), wybór algorytmów i opcji, które mają być używane, oraz określenie reguł zatrzymujących.

Wybierz modele. Domyślnie wybierane są wszystkie modele; jednak w przypadku serwera Analytic Server można zdecydować o ograniczeniu modeli tylko do tych, które działają na serwerze Analytic Server, konfigurują je w taki sposób, aby tworzyły modele rozdzielone lub były gotowe na przetwarzanie bardzo dużych zbiorów danych.

Uwaga: Lokalne tworzenie modeli Analytic Server w węźle Auto Grupowanie nie jest obsługiwane.

Używane modele. Zaznacz pola wyboru w kolumnach po lewej stronie typów modeli (algorytmów), które mają być uwzględnione w porównaniu. Im więcej typów zostanie wybranych, tym więcej zostanie utworzonych modeli i tym dłużej będzie trwało przetwarzanie.

Typ modelu. Lista dostępnych algorytmów (patrz niżej).

Parametry modelu. Dla każdego typu modelu można użyć ustawień domyślnych lub wybrać opcję **Określ** i samodzielnie określić opcje. Opcje są podobne do dostępnych w poszczególnych węzłach modelowania, z tym, że można wybrać dowolną liczbę ustawień obowiązujących w większości przypadków. Na przykład, porównując modele sieci neuronowych możemy, zamiast wybierać jedną z sześciu metod uczenia, wybrać je wszystkie i uczyć sześć modeli w jednym przebiegu.

Liczba modeli. Liczba modeli generowanych przez każdy z algorytmów na podstawie bieżących ustawień. Łączenie opcji może spowodować szybki wzrost liczby modeli, dlatego zdecydowanie zaleca się uważną obserwację tej liczby, zwłaszcza przy pracy z dużymi zbiorami danych.

Ogranicz maksymalny czas na budowanie jednego modelu. (Tylko modele Metoda k-średnich, Sieć Kohonena, Dwustopniowa, SVM, KNN, Sieci Bayesa i Lista decyzyjna) Określa limit czasu tworzenia jednego modelu. Na przykład, jeśli czas uczenia jednego konkretnego modelu jest nieprzewidywalny ze względu na pewne złożone interakcje, to nie chcemy, by ten model wstrzymywał cały przebieg modelowania.

Obsługiwane algorytmy

Węzeł Metoda k-średnich grupuje zbiór danych w osobne grupy (lub skupienia). Metoda ta definiuje stałą liczbę skupień, w sposób iteracyjny przypisuje rekordy do skupień i dopasowuje centra skupień do chwili, gdy dalsze pokrycie nie będzie miało wpływu na ulepszenie modelu. Zamiast prób predykcji danych wynikowych *k*-średnia korzysta z procesu znanego jako nienadzorowane uczenie w celu ujawnienia wzorców w zbiorze zmiennych wejściowych.

Węzeł Sieć Kohonena generuje typ sieci neuronowej, którą można wykorzystać do grupowania zbioru danych w osobne grupy. Po pełnym przeszkoleniu sieci rekordy podobne do siebie powinny znajdować się blisko siebie na mapie wyników, podczas gdy rekordy różne od siebie powinny znajdować się daleko od siebie. Na podstawie liczby obserwacji przechwyconych przez każdą jednostkę w modelu użytkowym można rozpoznać silne jednostki. Może to dać pojęcie o odpowiedniej liczbie skupień.

Węzeł Dwustopniowa korzysta z dwustopniowej metody grupowania. Pierwszy krok stanowi pojedynczy przebieg danych z myślą o kompresji surowych danych wejściowych w łatwy w zarządzaniu zestaw podgrup. Drugi krok korzysta z hierarchicznej metody grupowania w celu progresywnego scalania podgrup w coraz większe grupy. Metoda Dwustopniowa oferuje korzyści wynikające z automatycznego szacowania optymalnej liczby grup na potrzeby danych szkoleniowych. Pozwala ona skutecznie obsługiwać mieszane typy zmiennych i duże zbiory danych.

Węzeł SVM z jedną klasą korzysta z algorytmu uczenia nienadzorowanego. Węzeł ten można wykorzystać do wykrywania nowości. Wykryje on miękką granicę danego zbioru próbek, a następnie sklasyfikuje nowe punkty jako należące do tego zbioru albo do niego nienależące. Węzeł modelowania SVM z jedną klasą w programie SPSS Modeler został zaimplementowany w języku Python i wymaga biblioteki Python scikit-learn©.

K-średnie to jeden z najpowszechniej używanych algorytmów grupowania. Grupuje on punkty danych w określoną z góry liczbę skupień. Węzeł K-średnie-AS w programie SPSS Modeler jest zaimplementowany w środowisku Spark. Aby uzyskać szczegółowe informacje na temat algorytmów K-średnich, patrz https://spark.apache.org/docs/2.2.0/ml-clustering.html. Należy zwrócić uwagę, że węzeł K-średnie-AS automatycznie wykonuje kodowanie one-hot (kodowanie z gorącą jedynką) dla zmiennych kategorialnych.

Opcje odrzucania węzła Auto Grupowanie

Karta Odrzuć węzła Auto Grupowanie umożliwia automatyczne odrzucanie modeli, które nie spełniają określonych kryteriów. Modele te nie będą wymienione w modelu użytkowym.

Można określić minimalną wartość Silhouette, liczby grup, wielkości grup i ważność zmiennej ewaluacyjnej używanej w modelu. Miara Silhouette i liczba oraz wielkość grup określane są tak, jak podano to w węźle modelowania. Więcej informacji można znaleźć w temacie “Opcje modelu węzła Auto Grupowanie” na stronie 76.

Opcjonalnie można skonfigurować węzeł tak, by wykonywanie było zatrzymywane po wygenerowaniu pierwszego modelu spełniającego wszystkie określone kryteria. Więcej informacji można znaleźć w temacie “Reguły zatrzymujące zautomatyzowanego węzła modelowania” na stronie 64.

Zautomatyzowane modele użytkowe

Po wykonaniu zautomatyzowanego węzła modelowania węzeł ocenia modele kandydackie pod kątem każdej możliwej kombinacji opcji, przydziela rangę każdemu modelowi kandydackiemu na podstawie miary określonej przez użytkownika i zapisuje najlepsze modele w złożonym zautomatyzowanym modelu użytkowym. Ten model użytkowy w rzeczywistości zawiera zestaw jednego lub większej liczby modeli wygenerowanych przez węzeł. Te modele można pojedynczo przeglądać i wybierać w celu oceniania. Typ modelu i czas utworzenia są zawarte na liście dla każdego modelu, podobnie jak liczba innych miar odpowiednich dla typu modelu. Tabelę można posortować według dowolnej z tych kolumn, aby szybko zidentyfikować najbardziej interesujące modele.

- W celu przeglądania dowolnego modelu użytkowego kliknij dwukrotnie ikonę takiego modelu użytkowego. Następnie wygeneruj węzeł modelowania dla tego modelu do obszaru roboczego strumienia albo skopiuj model użytkowy do palety modeli.
- Wykresy miniaturowe umożliwiają szybką wzrokową ocenę poszczególnych typów modeli, co zostało podsumowane poniżej. W celu wygenerowania wykresu w pełnym wymiarze można kliknąć dwukrotnie ikonę. Wykres w pełnym wymiarze przedstawia nawet 1000 punktów, a jeśli zbiór danych zawiera więcej punktów, wówczas wykres jest oparty na próbie. (Wykresy rozrzutu są generowane ponownie każdorazowo w momencie wyświetlenia, dlatego zmiany w danych źródłowych dla wykresu — na przykład aktualizacja próby losowej lub podzbioru, gdy nie jest zaznaczona opcja **Ustaw wartość początkową generatora liczb losowych** — może zostać odzwierciedlona każdorazowo w przypadku ponownego wyświetlenia wykresu rozrzutu).
- Użyj paska narzędzi, aby pokazać lub ukryć konkretne kolumny na karcie Model, albo zmienić kolumnę używaną do sortowania tabeli. (Sortowanie można również zmienić, klikając nagłówki kolumn).
- W celu usunięcia na stałe nieużywanych modeli użyj przycisku Usuń.
- Aby zmienić kolejność kolumn, kliknij nagłówek kolumny i przeciągnij ją w żądane miejsce.
- Jeśli używany jest podzbiór, możesz wyświetlić wyniki dla podzbioru uczącego lub testującego.

Konkretne kolumny są zależne od typu porównywanych modeli, a szczegóły zostały opisane poniżej.

Binarne zmienne przewidywane

- W przypadku modeli binarnych wykres miniaturowy przedstawia rozkład rzeczywistych wartości nałożonych na wartości przewidywane, co zapewnia graficzne wskazanie tego, ile rekordów zostało poprawnie przewidzianych w poszczególnych kategoriach.
- Kryteria rangowania są zgodne z opcjami wybranymi w węźle modelowania Auto Klasyfikacja. Więcej informacji można znaleźć w temacie “Opcje modelu węzła Auto Klasyfikacja” na stronie 65.
- W celu uzyskania maksimum korzyści zgłaszany jest również percentyl, w którym występuje maksimum.
- W przypadku wzrostu skumulowanego można zmienić wybrany percentyl, używając paska narzędzi.

Nominalne zmienne przewidywane

- W przypadku modeli nominalnych wykres miniaturowy przedstawia rozkład rzeczywistych wartości nałożonych na wartości przewidywane, co zapewnia graficzne wskazanie tego, ile rekordów zostało poprawnie przewidzianych w poszczególnych kategoriach.
- Kryteria rangowania są zgodne z opcjami wybranymi w węźle modelowania Auto Klasyfikacja. Więcej informacji można znaleźć w temacie “Opcje modelu węzła Auto Klasyfikacja” na stronie 65.

Przewidywane zmienne ilościowe

- W przypadku modeli ilościowych (zakres wartości numerycznych) wykres przedstawia wartości przewidywane i obserwowane dla każdego modelu, co zapewnia wizualne wskazanie korelacji między tymi wartościami. W celu uzyskania dobrego modelu punkty powinny wykazywać tendencje do skupiania się wzdłuż przekątnej, a nie powinny być losowo rozrzucone po wykresie.
- Kryteria rangowania są zgodne z opcjami wybranymi w węźle modelowania Auto Predykcja. Więcej informacji można znaleźć w temacie “Opcje modelu węzła Auto Predykcja” na stronie 72.

Przewidywane zmienne grupowania

- W przypadku modeli skupień wykres przedstawia zliczenia wobec grup dla każdego modelu, co zapewnia graficzne wskazanie rozkładu grup.
- Kryteria rangowania są zgodne z opcjami wybranymi w węźle modelowania Auto Grupowanie. Więcej informacji można znaleźć w temacie “Opcje modelu węzła Auto Grupowanie” na stronie 76.

Wybór modeli do oceniania

Kolumna **Użyć?** umożliwia wybieranie modeli do oceniania.

- W przypadku binarnych, nominalnych i liczbowych zmiennych przewidywanych można wybrać wiele modeli do oceniania i połączyć wyniki w jeden, zespolony model użytkowy. Połączenie predykcji z wielu modeli umożliwia obejście ograniczeń w poszczególnych modelach, co często powoduje wyższą ogólną dokładność niż dokładność, jaką można uzyskać z dowolnego modelu.
- W przypadku modeli skupień w danym momencie można wybrać tylko jeden model do oceniania. Domyślnie jako pierwszy wybierany jest model posiadający najwyższą rangę.

Generowanie węzłów i modeli

Istnieje możliwość wygenerowania kopii złożonego zautomatyzowanego modelu użytkowego albo zautomatyzowanego węzła modelowania, z którego model został zbudowany. Na przykład może to być użyteczne, jeśli użytkownik nie posiada oryginalnego strumienia, z którego został zbudowany zautomatyzowany model użytkowy. Można również wygenerować model użytkowy lub węzeł modelowania dla dowolnych modeli określonych na liście zautomatyzowanego modelu użytkowego.

Model użytkowy zautomatyzowanego modelowania

Z menu Utwórz wybierz opcję **Model do palety**, aby dodać zautomatyzowany model użytkowy do palety Modele. Wygenerowany model można zapisać lub użyć bez konieczności ponownego uruchamiania strumienia.

Można również wybrać opcję **Utwórz węzeł modelowania** z menu Utwórz, aby dodać węzeł modelowania do obszaru roboczego strumienia. Ten węzeł można użyć w celu ponownego oszacowania wybranych modeli bez konieczności powtarzania całego przebiegu modelowania.

Pojedynczy model użytkowy modelowania

1. W menu **Model** kliknij dwukrotnie pojedynczy wymagany model użytkowy. Kopia tego modelu użytkowego zostanie otwarta w nowym oknie dialogowym.
2. Z menu Utwórz w nowym oknie dialogowym wybierz opcję **Model do palety**, aby dodać pojedynczy model użytkowy modelowania do palety Modele.
3. Można również wybrać opcję **Utwórz węzeł modelowania** z menu Utwórz w nowym oknie dialogowym, aby dodać pojedynczy węzeł modelowania do obszaru roboczego strumienia.

Generowanie wykresów ewaluacyjnych

W przypadku modeli binarnych (i tylko takich) można generować wykresy ewaluacyjne, które umożliwiają wizualną ocenę i porównywanie wydajności poszczególnych modeli. Wykresy ewaluacyjne są niedostępne w przypadku modeli wygenerowanych przez węzły automatycznej predykcji ani węzły automatycznego grupowania.

1. Pod kolumną *Użyć?* w modelu użytkowym automatycznego grupowania wybierz modele do ewaluacji.
2. W menu Utwórz wybierz opcję **Wykresy ewaluacyjne**. Zostanie wyświetlone okno dialogowe Wykres ewaluacyjny.
3. Wybierz typ wykresu i inne opcje.

Wykresy ewaluacyjne

Na karcie Model w zautomatyzowanym modelu użytkowym można przejść w dół, aby wyświetlić pojedyncze wykresy dla każdego z przedstawionych modeli. W przypadku modeli użytkowych z automatyczną klasyfikacją i automatyczną

predykcją karta Wykres przedstawia wykres oraz ważność predyktorów, które odzwierciedlają wyniki wszystkich modeli łącznie. Więcej informacji można znaleźć w temacie “Ważność predyktorów” na stronie 43.

W przypadku automatycznej klasyfikacji wyświetlany jest wykres, natomiast w przypadku automatycznej predykcji wyświetlany jest wykres multiplot (zwany również wykresem rozrzutu).

Rozdział 6. Drzewa decyzyjne

Modele drzew decyzyjnych

Modele drzew decyzyjnych umożliwiają tworzenie systemów klasyfikacji, które przewidują lub klasyfikują przyszłe obserwacje na podstawie zestawu reguł decyzyjnych. Mając dane podzielone na interesujące nas klasy (np. kredyty o wysokim ryzyku kontra kredyty o niskim ryzyku, abonenci kontra użytkownicy prepaid, głosujący kontra niegłosujący, typy bakterii), możemy wykorzystać te dane do budowania reguł klasyfikujących stare lub nowe obserwacje z maksymalną dokładnością. Na przykład możemy zbudować drzewo klasyfikujące ryzyko kredytowe lub zamiar zakupu na podstawie wieku i innych czynników.

Ta strategia, nazywana czasem *wywodzeniem reguł*, ma kilka zalet. Po pierwsze, proces wnioskowania będący zapleczem modelu jest oczywisty dla osoby przeglądającej drzewo. W przypadku technik modelowania typu *czarna skrzynka* czasem trudno jest ustalić, jaką wewnętrzną logiką kieruje się algorytm.

Po drugie, proces autonomicznie uwzględnia w regułach tylko te atrybuty, które są naprawdę istotne przy podejmowaniu decyzji. Atrybuty, które nie zwiększają dokładności drzewa, są ignorowane. Takie rozwiązanie może dostarczyć bardzo użytecznych informacji o danych i umożliwia wybranie tylko istotnych zmiennych przed rozpoczęciem uczenia innego modelu, np. sieci neuronowej.

Modele użytkowe drzew decyzyjnych można przekształcać w zbiory reguł co-jeśli (*zestawy reguł*), które w wielu przypadkach przedstawiają informacje w bardziej zrozumiałej postaci. Prezentacja w formie drzewa decyzyjnego jest użyteczna, gdy chcemy sprawdzić, w jaki sposób atrybuty w danych mogą *dzielić* populację na podzbiory istotne dla naszego problemu. Wyniki węzła Drzewo - AS różnią się od wyników innych węzłów drzew decyzyjnych, ponieważ model użytkowy od razu zawiera listę reguł i nie wymaga tworzenia zestawu reguł. Prezentacja w formie zestawu reguł jest użyteczna, gdy chcemy dowiedzieć się, jaki związek mają poszczególne grupy elementów z konkretnym wnioskiem. Na przykład następująca reguła tworzy *profil* grupy samochodów wartych kupienia (sprawdzonych i z niskim przebiegiem):

```
IF tested = 'yes'
AND mileage = 'low'
THEN -> 'BUY'.
```

Algorytmy budowania drzewa

Dostępnych jest kilka algorytmów służących do klasyfikacji i analizy segmentacji. Wszystkie te algorytmy zasadniczo realizują to samo zadanie: dzieląc dane na kolejne podgrupy, analizują wszystkie zmienne w zbiorze danych, by znaleźć zmienną zapewniającą najlepszą klasyfikację lub predykcję. Proces jest rekursywny, a grupy są dzielone na coraz mniejsze jednostki aż do ukończenia drzewa (zgodnie z określonym kryterium zatrzymania). Zmienne przewidywane i wejściowe używane do budowania drzewa mogą być ilościowe (przedział liczbowy) lub jakościowe, w zależności od algorytmu. Jeśli zmienna przewidywana jest ilościowa, generowane jest drzewo regresji; jeśli zmienna przewidywana jest jakościowa, generowane jest drzewo klasyfikacji.

Węzeł klasyfikacji i regresji (C&RT) generuje drzewo decyzyjne umożliwiające predykcję lub klasyfikację przyszłych obserwacji. W metodzie tej stosowany jest rekursywny podział rekordów na segmenty przez minimalizację zanieczyszczeń w każdym kroku, przy czym węzeł w drzewie jest uważany za „czysty”, jeśli 100% obserwacji w węźle przypada na konkretną kategorię zmiennej przewidywanej. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi lub jakościowymi (nominalnymi, porządkowymi lub flagami); wszystkie podziały są binarne (tylko dwie podgrupy).

Węzeł CHAID generuje drzewa decyzyjne, korzystając ze statystyk chi-kwadrat w celu identyfikacji optymalnych podziałów. W odróżnieniu od węzłów C&RT i węzłów QUEST, CHAID może generować drzewa niebinarne, co oznacza, że niektóre podziały mają więcej niż dwie gałęzie. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi (ilościowymi) lub jakościowymi. Wyczerpujący CHAID stanowi modyfikację CHAID umożliwiającą dokładniejsze badanie wszystkich możliwych podziałów, lecz obliczenia w jego przypadku zajmują więcej czasu.

Węzeł QUEST oferuje metodę klasyfikacji binarnej służącą do budowania drzew decyzyjnych, zaprojektowaną w celu redukcji czasu przetwarzania analiz dużych drzew decyzyjnych C&R, a jednocześnie w celu redukcji tendencji obecnej w metodach drzew klasyfikacji do preferowania danych wejściowych dopuszczających więcej podziałów. Zmienne wejściowe mogą być zakresami liczbowymi (ciągłymi), lecz zmienna przewidywana musi być jakościowa. Wszystkie podziały są binarne.

Węzeł C5.0 tworzy drzewo decyzyjne lub zestaw reguł. Model działa w oparciu o podział próby na podstawie zmiennej oferującej maksimum korzyści z informacji na każdym z poziomów. Zmienna przewidywana musi być jakościowa. Dozwolonych jest wiele podziałów na więcej niż dwie podgrupy.

Węzeł Drzewo-AS jest podobny do istniejącego węzła CHAID; jednak węzeł Drzewo-AS jest przeznaczony do przetwarzania dużych zbiorów w celu utworzenia pojedynczego drzewa i wyświetla model wynikowy w przeglądarce wyników, która została dodana w programie SPSS Modeler, wersja 17. Węzeł generuje drzewo decyzyjne używając statystyki chi-kwadrat (CHAID), aby określić optymalne podziały. CHAID może generować drzewa niebinarne, co oznacza, że niektóre podziały mają więcej niż dwie gałęzie. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi (ilościowymi) lub jakościowymi. Wyczerpujący CHAID stanowi modyfikację CHAID umożliwiającą dokładniejsze badanie wszystkich możliwych podziałów, lecz obliczenia w jego przypadku zajmują więcej czasu.

Węzeł Drzewa losowe jest podobny do istniejącego węzła C&RT; jednak węzeł Drzewa losowe jest przeznaczony do przetwarzania dużych zbiorów danych w celu utworzenia pojedynczego drzewa i wyświetla model wynikowy w przeglądarce wyników, która została dodana w programie SPSS Modeler, wersja 17. Węzeł Drzewa losowe generuje drzewo decyzyjne umożliwiające predykcję lub klasyfikację przyszłych obserwacji. W metodzie tej stosowany jest rekursywny podział rekordów na segmenty przez minimalizację zanieczyszczeń w każdym kroku, przy czym węzeł w drzewie jest uważany za *czysty*, jeśli 100% obserwacji w węźle przypada na konkretną kategorię zmiennej przewidywanej. Zmienne przewidywana i wejściowa mogą być zakresami liczbowymi lub jakościowymi (nominalnymi, porządkowymi lub flagami); wszystkie podziały są binarne (tylko dwie podgrupy).

Ogólne zastosowania analizy w oparciu o drzewo

Poniżej przedstawiono niektóre ogólne zastosowania analizy w oparciu o drzewo:

Segmentacja: określenie, które osoby prawdopodobnie należą do konkretnej klasy.

Podział na warstwy: przypisywanie każdej obserwacji do jednej z kilku kategorii, np. grupy wysokiego, średniego i niskiego ryzyka.

Predykcja: tworzenie reguł i wykorzystanie ich do przewidywania przyszłych zdarzeń. Predykcja może być także rozumiana jako próby powiązania atrybutów predykcyjnych z wartościami zmiennej ilościowej.

Redukcja danych i monitorowanie zmiennych: wybór użytecznych podzbiorów predyktorów z dużego zbioru zmiennych do wykorzystania przy budowaniu formalnego modelu parametrycznego.

Identyfikacja interakcji: identyfikacja relacji dotyczących tylko konkretnych podgrup i użycie ich w formalnym modelu parametrycznym.

Scalanie kategorii i kategoryzowanie zmiennych ilościowych: przekodowywanie kategorii predyktorów grup i zmiennych ilościowych w sposób minimalizujący straty informacji.

Interaktywny konstruktor drzewa

Model drzewa można wygenerować automatycznie, tam gdzie algorytm decyduje o najlepszym podziale na każdym poziomie, lub można przejąć kontrolę za pomocą interaktywnego konstruktora drzewa, wykorzystując wiedzę biznesową do zawężenia lub uproszczenia drzewa przed zapisaniem modelu użytkowego.

1. Należy utworzyć strumień i dodać jeden z węzłów drzew decyzyjnych: C&RT, CHAID lub QUEST.

   Uwaga: Budowanie drzew interaktywnych nie jest obsługiwane w przypadku drzew C5.0 ani Drzewo-AS.
2. Otwórz węzeł i na karcie Zmienne wybierz zmienne przewidywane oraz predykcyjne, a następnie wskaż dodatkowe opcje modelu odpowiednio do potrzeb. Szczegółowe instrukcje zawiera dokumentacja dla każdego węzła budowania drzewa.
3. Na panelu Cele karty Opcje budowania wybierz opcję **Drzewo interakcyjne**.
4. Kliknij przycisk **Uruchom**, aby uruchomić konstruktor drzewa.

Zostanie wyświetlone bieżące drzewo, począwszy od węzła głównego. Drzewo można edytować i przycinać poziom po poziomie; można także przed wygenerowaniem jednego lub większej liczby modeli uzyskać dostęp do korzyści, ryzyk i powiązanych informacji.

Komentarze

- W przypadku węzłów C&RT, CHAID i QUEST wszelkie zmienne porządkowe użyte w modelu muszą charakteryzować się składowaniem typu numerycznego (nie łańcuchowego). W razie potrzeby do ich przekształcenia można użyć węzła rekodowania.
- Opcjonalnie można użyć zmiennej dzielącej na podzbiory do podzielenia danych na próby do uczenia i testowe.
- Alternatywnie wobec skorzystania z konstruktora drzewa, można także wygenerować model bezpośrednio z węzła modelowania — podobnie jak w przypadku innych modeli IBM SPSS Modeler. Więcej informacji można znaleźć w temacie “Budowanie modelu drzewa bezpośrednio” na stronie 95.

Rozwijanie i przycinanie drzewa

Karta Przegląd w konstruktorze drzewa umożliwia wyświetlanie bieżącego drzewa, począwszy od węzła głównego.

1. Aby rozwinąć drzewo, z menu wybierz:

   Drzewo > Rozwiń drzewo

   System buduje drzewo, rekurencyjnie dzieląc każdą gałąź aż do spełnienia co najmniej jednego kryterium zatrzymania. Dla każdego podziału na podzbiory automatycznie wybierany jest na podstawie metody modelowania najlepszy predyktor.
2. Można również wybrać opcję **Rozwiń drzewo o jeden poziom**, aby dodać jeden poziom.
3. Aby dodać gałąź poniżej określonego węzła, wybierz węzeł, a następnie wybierz opcję **Rozwiń gałąź**.
4. Aby wybrać predyktor używany do podziału na podzbiory, wybierz żądany węzeł, a następnie wybierz opcję **Rozwiń gałąź według podziału**. Więcej informacji można znaleźć w temacie “Definiowanie podziałów niestandardowych” na stronie 86.
5. Aby przyciąć gałąź, wybierz węzeł, a następnie wybierz opcję **Usuń gałąź**, aby skasować wybrany węzeł.
6. Aby usunąć dolny poziom z drzewa, wybierz opcję **Usuń jeden poziom**.
7. W przypadku drzew K&R i QUEST wybierz opcję **Rozwiń drzewo i przytnij**, aby przyciąć drzewo w oparciu o algorytm złożoności kosztów dostosowujący oszacowanie ryzyka w oparciu o liczbę węzłów końcowych, co zwykle skutkuje prostszym drzewem. Więcej informacji można znaleźć w temacie “Węzeł C&RT” na stronie 97.

Odczytywanie reguł podziału na karcie Przegląd

Podczas wyświetlania reguł podziału na karcie Przegląd nawiasy kwadratowe oznaczają, że sąsiednia wartość jest uwzględniona w zakresie, zaś cudzysłów oznacza, że sąsiednia wartość jest wykluczona z zakresu. Wyrażenie (23,37] oznacza więc: od 23 wyłącznie do 37 włącznie, innymi słowy: od ponad 23 do 37. Na karcie Model ten sam warunek będzie wyświetlany następująco:

```
Age > 23 i Age <= 37
```

Przerywanie rozwijania drzewa. Aby przerwać operację rozwijania drzewa (na przykład w sytuacji, gdy zajmuje ona znacznie więcej czasu, niż oczekiwano), kliknij przycisk Zatrzymaj wykonywanie na pasku narzędzi.

Rysunek 28. Przycisk Zatrzymaj wykonywanie

Przycisk jest aktywny tylko podczas rozwijania drzewa. Zatrzymuje on bieżącą operację rozwijania w aktualnym punkcie, pozostawiając wszelkie dodane już węzły, bez zapisywania zmian i zamykania okna. Konstruktor drzewa pozostaje otwarty, co pozwala stosownie do potrzeb wygenerować model, zaktualizować dyrektywy lub wyeksportować wyniki w odpowiednim formacie.

Definiowanie podziałów niestandardowych

Okno dialogowe Definiuj podział umożliwia wybranie predyktora i określenie warunków dla każdego podziału.

1. W konstruktorze drzewa wybierz węzeł na karcie Przegląd, a następnie z menu wybierz:

   Drzewo > Rozwiń gałąź według podziału
2. Wybierz żądany predyktor z listy rozwijanej lub kliknij przycisk **Predyktory**, aby wyświetlić szczegóły każdego z predyktorów. Więcej informacji można znaleźć w temacie “Wyświetlanie szczegółów predyktora”.
3. Można zaakceptować warunki domyślne dla każdego podziału lub wybrać opcję **Użytkownika** w celu określenia warunków dla podziału odpowiednio do potrzeb.
   - W przypadku predyktorów ilościowych (przedziałów liczbowych) można użyć zmiennych **Edytuj wartości przedziału** w celu określenia rozstępu wartości przypadających na każdy nowy węzeł.
   - W przypadku predyktorów jakościowych można użyć opcji **Edytuj wartości nominalne** lub **Edytuj wartości porządkowe** w celu określenia konkretnych wartości (lub przedziału wartości w przypadku predyktora porządkowego) mapowanych na każdy nowy węzeł.
4. Wybierz opcję **Rozwiń**, aby ponownie rozwinąć gałąź z użyciem wybranego predyktora.

Drzewo można w ogólnym przypadku podzielić z użyciem dowolnego predyktora, niezależnie od reguł zatrzymujących. Jedyne wyjątki dotyczą sytuacji, w których węzeł jest węzłem czystym (to znaczy 100% obserwacji przypada na tę samą klasę przewidywaną, co oznacza brak pozostałych danych do podziału) lub gdy wybrany predyktor jest stałą (brak podziału).

Brakujące wartości w. W przypadku drzew CHAID, jeśli dla danego predyktora występują braki danych, wówczas możesz zdefiniować niestandardowy podział umożliwiający przypisanie ich do konkretnego węzła podrzędnego. (W przypadku drzew K&R i QUEST braki danych są obsługiwane z użyciem substytutów zdefiniowanych w algorytmie. Więcej informacji można znaleźć w temacie “Substytuty i szczegóły podziału” na stronie 87.)

Wyświetlanie szczegółów predyktora

W oknie dialogowym Wybierz predyktor wyświetlane są statystyki na temat dostępnych predyktorów (lub „konkurentów”, jak się je niekiedy nazywa), które mogą być używane przy bieżącym podziale.

- W przypadku CHAID i wyczerpującego CHAID dla każdego predyktora jakościowego wyświetlana jest statystyka chi-kwadrat; jeśli predyktor charakteryzuje się przedziałem liczbowym, wyświetlana jest statystyka *F*. Statystyka chi-kwadrat stanowi miarę niezależności zmiennej przewidywanej od zmiennej podziału. Wyższa statystyka chi-kwadrat generalnie wiąże się z niższym prawdopodobieństwem, co oznacza mniejsze szanse na niezależność dwu zmiennych — a to z kolei wskazuje, że podział jest dobry. Uwzględniane są także stopnie swobody — pozwala to bowiem uwzględnić fakt, że łatwiej jest uzyskać dużą statystykę i niewielkie prawdopodobieństwo w przypadku podziału trójwymiarowego, niż w przypadku podziału dwuwymiarowego.

- W przypadku Drzewa K&R i QUEST wyświetlana jest poprawa dla każdego predyktora. Im większa poprawa, tym większa redukcja zanieczyszczeń między węzłami nadrzędnymi i podrzędnymi, o ile ten predyktor jest używany. (Węzeł czysty to taki, w którym wszystkie obserwacje przypadają do jednej kategorii zmiennych przewidywanych; im mniejsza wartość zanieczyszczenia w drzewie, tym lepsze dopasowanie modelu do danych). Innymi słowy, wyższa wartość zanieczyszczenia generalnie wskazuje przydatność podziału dla drzewa danego typu. Używana miara zanieczyszczenia jest określona w węźle budowania drzewa.

Substytuty i szczegóły podziału

Wybierając dowolny węzeł na karcie Przegląd, a następnie wybierając przycisk informacji o podziale na karcie Przegląd po prawej stronie paska narzędzi, można wyświetlić szczegółowe informacje dotyczące podziału dla tego węzła. Wyświetlana jest używana reguła podziału wraz z odpowiednimi statystykami. W przypadku drzew jakościowych C&RT wyświetlane są poprawa i związek. Związek stanowi miarę korespondencji między substytutem a podstawową zmienną podziału, przy czym „najlepszy" substytut jest zwykle tym, który najlepiej naśladuje zmienną podziału. W przypadku drzew K&R i QUEST wyświetlane są również wszelkie substytuty użyte zamiast predyktora podstawowego.

Aby edytować podział dla wybranego węzła, można kliknąć ikonę po lewej stronie panelu substytutów, aby otworzyć okno dialogowe Definiuj podział. (Można także wybrać substytut z listy przed kliknięciem ikony w celu wybrania go jako podstawowej zmiennej podziału).

Substytuty. O ile ma to zastosowanie, wyświetlane są wszelkie substytuty głównej zmiennej podziału dla wybranego węzła. Substytuty to zmienne alternatywne, używane w przypadku, gdy brakuje wartości głównego predyktora dla danego rekordu. Maksymalna dozwolona liczba substytutów dla danego podziału jest określona w węźle budowania drzewa, ale liczba rzeczywista zależy od danych uczących. Ogólnie, im więcej danych brakuje, tym więcej substytutów można użyć. Dla innych modeli drzewa decyzyjnego ta karta jest pusta.

Uwaga: Aby substytuty były uwzględnione w modelu, muszą zostać zidentyfikowane podczas fazy uczenia. Jeśli w próbie uczącej nie brakuje wartości, wówczas nie zostanie określony żaden substytut, a wszystkie rekordy z brakami danych napotkane podczas testowania lub oceniania zostaną automatycznie przeniesione do węzła podrzędnego z największą liczbą rekordów. Jeśli podczas testowania lub oceniania spodziewane są braki danych, należy upewnić się, że wartości tych nie ma również w próbie uczącej. Substytuty są niedostępne dla drzew CHAID.

Mimo że substytuty nie są używane w przypadku drzew CHAID, podczas definiowania podziału użytkownika dostępna jest opcja przypisywania ich do określonego węzła podrzędnego. Więcej informacji można znaleźć w temacie "Definiowanie podziałów niestandardowych" na stronie 86.

Dostosowywanie widoku drzewa

Na karcie Przegląd w konstruktorze drzewa wyświetlane jest bieżące drzewo. Domyślnie wszystkie gałęzie drzewa są rozwinięte, można jednak rozwijać je i zwijać odpowiednio do potrzeb, a także dostosowywać inne ustawienia.

- Kliknij symbol minus (**–**) w dolnym prawym rogu węzła nadrzędnego, aby ukryć wszystkie jego węzły podrzędne. Kliknij symbol plus (**+**) w dolnym prawym rogu węzła nadrzędnego, aby wyświetlić wszystkie jego węzły podrzędne.
- Użyj menu Widok lub paska narzędzi, aby zmienić orientację drzewa (z góry na dół, od lewej do prawej lub od prawej do lewej strony).
- Kliknij przycisk „Wyświetl etykiety zmiennej i wartości" na głównym pasku narzędzi, aby wyświetlać i ukrywać zmienne i etykiety wartości.
- Użyj przycisku szkła powiększającego w celu powiększenia lub pomniejszenia obrazu, lub kliknij przycisk mapy drzewa po prawej stronie paska narzędzi, aby wyświetlić diagram kompletnego drzewa.
- Jeśli używana jest zmienna dzieląca na podzbiory, można zamienić widok drzewa miedzy podzbiorami uczącym a testowym (**Widok > Partycja**). Jeśli wyświetlany jest podzbiór testujący, drzewo można wyświetlić, ale nie można go edytować. (Bieżący podzbiór jest wyświetlany na pasku stanu w prawym dolnym rogu okna).
- Kliknij przycisk informacji o podziale (przycisk „i" na skraju paska narzędzi po prawej stronie), aby wyświetlić szczegóły bieżącego podziału. Więcej informacji można znaleźć w temacie "Substytuty i szczegóły podziału".

- Wyświetl dane w postaci statystyk, wykresów lub w obu tych postaciach w ramach każdego węzła (patrz poniżej).

Wyświetlanie statystyk i wykresów

Statystyka dla węzła. W przypadku przewidywanej zmiennej jakościowej tabela w każdym węźle przedstawia liczbę i wartość procentową rekordów w każdej kategorii oraz wartość procentową względem całej próby, jaką reprezentuje węzeł. W przypadku ilościowej zmiennej przewidywanej (przedział liczbowy) w tabeli wyświetlane są średnia, odchylenie standardowe, liczba rekordów oraz wartość predykcyjna zmiennej przewidywanej.

Wykresy dla węzła. W przypadku jakościowej zmiennej przewidywanej wykres jest wykresem słupkowym i przedstawia wartości procentowe dla każdej kategorii zmiennej przewidywanej. Każdy wiersz w tabeli poprzedza próbka koloru zgodnego z kolorem kategorii zmiennej przewidywanej na wykresach dla tego węzła. W przypadku ilościowej zmiennej przewidywanej (przedział liczbowy) wykres zawiera histogram zmiennej przewidywanej dla rekordów w węźle.

Korzyści

Na karcie Korzyści wyświetlane są statystyki dla wszystkich węzłów końcowych drzewa. Korzyści stanowią miarę tego, jak bardzo średnia lub proporcja dla danego węzła różnią się względem ogólnej średniej. Ogólnie mówiąc, im większa ta różnica, tym bardziej użyteczne jest drzewo jako narzędzie podejmowania decyzji. Na przykład indeks lub wartość przyrostu dla węzła wynosząca 148% oznacza, że rekordy w tym węźle z około półtorakrotnie większym prawdopodobieństwem, niż cały zbiór danych, należą do kategorii zmiennych przewidywanych.

W przypadku węzłów C&RT i QUEST, w przypadku których określono zbiór zabezpieczający przed przeuczeniem wyświetlane są dwa zbiory statystyk:

- zbiór rozwijania drzewa — podzbiór uczący z usuniętym zbiorem zabezpieczającym przed przeuczeniem
- zbiór zabezpieczający przed przeuczeniem

W przypadku innych drzew interaktywnych C&RT i QUEST, a także w przypadku wszystkich drzew interaktywnych CHAID wyświetlane są tylko statystyki zbioru rozwijania drzewa.

Karta Korzyści umożliwia:

- Wyświetlanie statystyk węzeł po węźle, skumulowanych lub kwantyli.
- Wyświetlanie korzyści lub zysków.
- Przełączanie między widokami tabel i wykresów.
- Wybieranie kategorii zmiennych przewidywanych (tylko przewidywane zmienne jakościowe).
- Sortowanie tabeli w porządku rosnącym lub malejącym, w oparciu o wartość procentową indeksu. Jeśli wyświetlane są statystyki dla wielu podzbiorów, sortowanie odbywa się zawsze dla podzbioru uczącego, nie zaś dla podzbioru testującego.

W ogólnym przypadku wybory dokonywane w tabeli korzyści są aktualizowane w widoku drzewa i odwrotnie. Na przykład w przypadku wyboru wiersza w tabeli w drzewie zostanie wybrany odpowiadający mu węzeł.

Korzyści dla klasyfikacji

W przypadku drzew klasyfikacji (tych z przewidywanymi zmiennymi jakościowymi) wartość procentowa indeksu korzyści mówi o tym, o ile bardziej różnią się proporcje dla danej kategorii zmiennych przewidywanych w każdym węźle od proporcji ogólnych.

Statystyka węzeł po węźle

W tym widoku w tabeli wyświetlany jest jeden wiersz dla każdego węzła końcowego. Na przykład, jeśli ogólna odpowiedź na kampanię mailingową wyniosła 10%, a odpowiedź pozytywną uzyskano w przypadku 20% rekordów należących do węzła X, wartość procentowa indeksu dla tego węzła będzie wynosić 200%, co oznacza, że respondenci w tej grupie zdecydują się na zakup z dwukrotnie większym prawdopodobieństwem w porównaniu z całą populacją.

W przypadku węzłów C&RT i QUEST, w przypadku których określono zbiór zabezpieczający przed przeuczeniem wyświetlane są dwa zbiory statystyk:

- zbiór rozwijania drzewa — podzbiór uczący z usuniętym zbiorem zabezpieczającym przed przeuczeniem
- zbiór zabezpieczający przed przeuczeniem

W przypadku innych drzew interaktywnych C&RT i QUEST, a także w przypadku wszystkich drzew interaktywnych CHAID wyświetlane są tylko statystyki zbioru rozwijania drzewa.

Węzły. Identyfikator bieżącego węzła (wyświetlany na karcie Przegląd).

Węzeł: n. Łączna liczba rekordów w tym węźle.

Węzeł (%). Wartość procentowa wszystkich rekordów w tym zbiorze danych, przypadających na ten węzeł.

Korzyść: n. Liczba rekordów wybranej kategorii zmiennych przewidywanych przypadających na ten węzeł. Innymi słowy: ile spośród wszystkich rekordów ze zbioru danych należących do kategorii zmiennych przewidywanych znajduje się w tym węźle?

Korzyść (%). Wartość procentowa wszystkich rekordów w kategorii zmiennych przewidywanych w całym zbiorze danych, przypadających na ten węzeł.

Odpowiedź (%). Wartość procentowa rekordów w bieżącym węźle należących do kategorii zmiennych przewidywanych. Odpowiedzi w tym kontekście są niekiedy nazywane „trafieniami”.

Indeks (%). Wartość procentowa dla tego węzła wyrażona jako procent wartości procentowej odpowiedzi w całym zbiorze danych. Na przykład indeks wynoszący 300% oznacza, że rekordy w tym węźle z około trzykrotnie większym prawdopodobieństwem, niż cały zbiór danych, należą do kategorii zmiennych przewidywanych.

Statystyka skumulowana

W widoku skumulowanym w tabeli wyświetlany jest jeden węzeł na każdy wiersz, lecz statystyki są skumulowane i posortowane rosnąco lub malejąco, wg wartości procentowej indeksu. Na przykład w przypadku zastosowania sortowania malejącego węzeł o najwyższej wartości procentowej indeksu jest wymieniony jako pierwszy, zaś statystyki prezentowane w kolejnych wierszach są skumulowane dla tego wiersza i powyższych.

Skumulowana wartość procentowa indeksu zmniejsza się wiersz po wierszu w miarę, jak dodawane są wiersze o coraz mniejszej i mniejszej wartości procentowej odpowiedzi. Wartość skumulowana indeksu dla wiersza końcowego wynosi zawsze 100%, ponieważ na tym etapie uwzględniany jest cały zbiór danych.

Kwantyle

W tym widoku każdy wiersz w tabeli reprezentuje kwantyl, nie zaś węzeł. Kwantyle to kwartyle, kwintyle (części piąte), decyle (części dziesiętne), vingtyle (części dwudzieste) albo percentyle (części setne). W jednym kwantylu można wyświetlić wiele węzłów, o ile uzyskanie danej wartości procentowej wymaga więcej niż jednego węzła (na przykład, jeśli wyświetlane są kwartyle, lecz dwa pierwsze węzły zawierają mniej niż 50% wszystkich obserwacji). Pozostałe wartości w tabeli są skumulowane i mogą być interpretowane w ten sam sposób, co widok skumulowany.

Zyski dla klasyfikacji i ROI

W przypadku drzew klasyfikacji statystyka korzyści może być również wyświetlana w postaci zysku i zwrotu z inwestycji. Okno dialogowe Definicja zysków umożliwia określenie przychodów i wydatków dla każdej kategorii.

1. Na pasku narzędzi na karcie Korzyści kliknij przycisk Zysk (oznaczony jako $/$), aby przejść do okna dialogowego.
2. Wprowadź wartości przychodów i wydatków dla każdej kategorii zmiennej przewidywanej.

Na przykład, jeśli wysłanie oferty każdemu klientowi to koszt 0,48 USD, zaś przychód z odpowiedzi pozytywnej to 9,95 USD w przypadku subskrypcji trzymiesięcznej, wówczas każda odpowiedź *nie* oznacza koszt w wysokości 0,48 USD, zaś każda odpowiedź *tak* oznacza przychód w wysokości 9,47 USD (wynik odejmowania: 9,95–0,48).

W tabeli korzyści **zysk** jest obliczany jako suma przychodów minus wydatki na każdy z rekordów w węźle końcowym. **ROI (zwrot z inwestycji)** jest wówczas łącznym zyskiem podzielonym przez łączny wydatek dla węzła.

Komentarze

- Wartości zysku wpływają tylko na wartości zysku średniego i ROI wyświetlane w tabeli korzyści, która prezentuje statystyki w sposób znacznie bardziej odpowiadający charakterowi osiąganych zysków. Wartości te nie mają wpływu na strukturę podstawowego modelu drzewa. Zysków nie należy mylić z kosztami błędnej klasyfikacji, które są określane w węźle budowania drzewa i są faktoryzowane w modelu w celu zabezpieczenia przed kosztownymi błędami.
- Specyfikacje zysku są nietrwałe między jedną interaktywną sesją budowy drzewa a drugą.

Korzyści dla regresji

W przypadku drzew regresji można wybrać widok węzeł po węźle, widok skumulowany węzeł po węźle lub jeden z widoków kwantyli. Wartości średnie wyświetlono w tabeli. Wykresy są dostępne tylko dla kwantyli.

Wykresy korzyści

Wykresy mogą być wyświetlane na karcie Korzyści, jako alternatywa dla tabel.

1. Na karcie Korzyści wybierz ikonę Kwantyle (trzecia od lewej na pasku narzędzi). (Wykresy nie są dostępne w przypadku statystyk węzeł po węźle ani skumulowanej.)
2. Wybierz ikonę Wykresy.
3. Wybierz jednostki wyświetlania (percentyle, decyle itd.) odpowiednio do potrzeb z listy rozwijanej.
4. Wybierz opcje **Korzyści**, **Odpowiedź** lub **Wzrost**, aby zmienić wyświetlaną miarę.

Wykres korzyści

Wykres korzyści przedstawia wartości z kolumny *Korzyści (%)* w tabeli. Korzyści są definiowane jako proporcja trafień w każdym przyroście względem łącznej liczby trafień w drzewie, według równania:

(trafień w przyroście / łączna liczba trafień) x 100%

Wykres obrazuje w istocie, jak szeroki zakres danych należy objąć, aby uzyskać daną wartość procentową wszystkich trafień w drzewie. Przekątna przedstawia oczekiwaną odpowiedź dla całej próby w przypadku, gdyby model nie został użyty. W takim przypadku wskaźnik odpowiedzi byłby stały, ponieważ prawdopodobieństwo udzielenia odpowiedzi przez poszczególne osoby jest jednakowe. Podwojenie wyniku wymagałoby przepytania dwukrotnie większej liczby osób. Krzywa wskazuje, o ile można poprawić odpowiedź, uwzględniając tylko osoby, w przypadku których korzyści mieszczą się w górnych percentylach. Na przykład uwzględnienie pierwszych 50% może dać więcej niż 70% odpowiedzi pozytywnych. Im większe nachylenie krzywej, tym wyższa korzyść.

Wykres przyrostu

Wykres przyrostu przedstawia wartości z kolumny *Indeks (%)* w tabeli. Wykres ten umożliwia porównanie wartości procentowej rekordów w każdym przyroście będących trafieniami z łączną wartością procentową trafień w tym zbiorze danych uczących, według równania:

(trafienia w przyroście / rekordy w przyroście) / (łączna liczba trafień / łączna liczba rekordów)

Wykres odpowiedzi

Wykres odpowiedzi przedstawia wartości z kolumny *Odpowiedź (%)* w tabeli. Wartość odpowiedzi to wartość procentowa rekordów w przyroście będących trafieniami, według równania:

(odpowiedzi w przyroście / rekordy w przyroście) x 100%

Wybór w oparciu o korzyść

Okno dialogowe Wybór korzyści umożliwia automatyczny wybór węzłów końcowych o najlepszych (lub najgorszych) korzyściach, na podstawie podanej reguły lub progu. Na podstawie tego wyboru można następnie wygenerować węzeł Selekcja.

1. Na karcie Korzyści wybierz widok węzeł po węźle lub skumulowany, a następnie wybierz kategorię zmiennych przewidywanych, w oparciu o którą ma zostać dokonany wybór. (Wybory bazują na aktualnie wyświetlanej zawartości tabeli i nie są dostępne dla kwantyli).
2. Na karcie Korzyści z menu wybierz:

   Edycja > Wybierz węzły końcowe > Wybór w oparciu o korzyść

   Wybierz tylko. Istnieje możliwość wyboru węzłów dopasowanych *lub* węzłów niedopasowanych — na przykład można wybrać *wszystkie z wyjątkiem* pierwszych 100 rekordów.

   Dopasuj według informacji o korzyściach. Dopasowuje węzły w oparciu o statystyki korzyści dla bieżącej kategorii zmiennych przewidywanych, w tym:

   - Węzły, których korzyść, odpowiedź lub przyrost (indeks) spełnia zadany próg — na przykład odpowiedź większa lub równa 50%.
   - Pierwszych *n* węzłów w oparciu o korzyść dla kategorii zmiennych przewidywanych.
   - Pierwsze z węzłów, aż do uzyskania zadanej liczby rekordów.
   - Pierwsze z węzłów, aż do uzyskania zadanej liczby danych uczących.
3. Kliknij przycisk **OK**, aby zaktualizować wybór na karcie Przegląd.
4. Aby utworzyć nowy węzeł Selekcja w oparciu o bieżący wybór na karcie Przegląd, wybierz opcję **Wybierz węzeł** z menu Utwórz. Więcej informacji można znaleźć w temacie “Generowanie węzłów filtrowania i selekcji” na stronie 94.

Uwaga: Z uwagi na fakt wyboru węzłów, nie zaś rekordów czy wartości procentowych, nie zawsze możliwe będzie osiągnięcie idealnego dopasowania do kryteriów wyboru. System wybiera kompletne węzły *aż do* podanego poziomu. Na przykład w przypadku wyboru pierwszych 12 obserwacji, z których 10 znajduje się w pierwszym węźle, zaś 2 w drugim, zostanie wybrany tylko pierwszy węzeł.

Ryzyka

Ryzyka informują o prawdopodobieństwie błędnej klasyfikacji na dowolnym poziomie. Na karcie Ryzyka wyświetlane jest punktowe oszacowanie ryzyka oraz (w przypadku wyników jakościowych) tabela błędnych klasyfikacji.

- W przypadku predykcji numerycznych ryzyko to sumaryczne oszacowanie wariancji w każdym z węzłów końcowych.
- W przypadku predykcji jakościowych ryzyko stanowi proporcję obserwacji nieprawidłowo sklasyfikowanych, dopasowane pod kątem ewentualnych prawdopodobieństw wstępnych lub kosztów błędnej klasyfikacji.

Zapisywanie modeli drzew i wyników

Wyniki interaktywnych sesji budowy drzewa można zapisać lub wyeksportować na szereg różnych sposobów; można między innymi:

- Wygenerować model w oparciu o bieżące drzewo (**Utwórz > Model**).
- Zapisać dyrektywy służące do rozwijania bieżącego drzewa. Przy kolejnym wykonaniu węzła budowania drzewa bieżące drzewo zostanie automatycznie ponownie rozwinięte, z uwzględnieniem ewentualnych, zdefiniowanych przez użytkownika podziałów na podzbiory.
- Wyeksportować informacje o modelu, korzyściach i ryzyku. Więcej informacji można znaleźć w temacie “Eksportowanie informacji o modelu, korzyściach i ryzyku” na stronie 94.

Korzystając z konstruktora drzew lub modelu użytkowego drzewa, można:

- Wygenerować węzeł Filtrowanie lub Selekcja w oparciu o bieżące drzewo. Więcej informacji można znaleźć w “Generowanie węzłów filtrowania i selekcji” na stronie 94.

- Wygenerować model użytkowy Zestaw reguł reprezentujący strukturę drzewa jako zestaw reguł definiujących gałęzie końcowe drzewa. Więcej informacji można znaleźć w “Generowanie zestawu reguł z drzewa decyzyjnego” na stronie 95.
- Ponadto model można wyeksportować w formacie PMML (dot. tylko modeli użytkowych drzewa). Więcej informacji można znaleźć w “Paleta modeli” na stronie 40. Jeśli model zawiera niestandardowe podziały, taka informacja nie jest zachowywana w wyeksportowanym pliku PMML. (Podział jest zachowywany, jednak fakt, że jest on niestandardowy, a nie wybrany przez algorytm, już nie).
- Wygenerować wykres na podstawie wybranej części bieżącego drzewa. Ma to zastosowanie tylko w przypadku węzła użytkowego, jeśli jest on dołączony do innego węzła w strumieniu. Więcej informacji można znaleźć w “Tworzenie wykresów” na stronie 124.

Uwaga: Samego drzewa interaktywnego nie można zapisać. Aby uniknąć utraty efektów swojej pracy, należy wygenerować model i/lub zaktualizować dyrektywy drzewa przed zamknięciem okna konstruktora drzewa.

Generowanie modelu za pomocą konstruktora drzewa

Aby wygenerować model w oparciu o bieżące drzewo, z menu konstruktora drzewa wybierz:

Utwórz > Model

W oknie dialogowym Generuj nowy model można wybrać spośród następujących opcji:

Nazwa modelu. Można podać nazwę użytkownika lub wygenerować nazwę automatycznie na podstawie nazwy węzła modelowania.

Utwórz węzeł na. Można wybrać jedną z następujących opcji dotyczących lokalizacji dodawanego węzła: **Obszar roboczy**, **Paleta modeli** lub **Łącznie**.

Dołącz dyrektywy drzewa. To pole wyboru należy zaznaczyć, aby dołączyć dyrektywy z bieżącego drzewa do generowanego modelu. Umożliwia to ponowne utworzenie drzewa, gdy zajdzie taka potrzeba. Więcej informacji można znaleźć w temacie “Dyrektywy rozwijania drzewa”.

Dyrektywy rozwijania drzewa

W przypadku modeli C&RT, CHAID i QUEST dyrektywy drzewa określają warunki rozwijania drzewa, jeden poziom naraz. Dyrektywy są stosowane przy każdym uruchomieniu interaktywnego konstruktora drzewa z węzła.

- Najbezpieczniejsze zastosowanie dyrektyw to ponowne generowanie drzew tworzonych w trakcie poprzedniej sesji interaktywnej. Więcej informacji można znaleźć w temacie “Aktualizowanie dyrektyw drzewa” na stronie 94. Dyrektywy można także edytować ręcznie, lecz należy przy tym zachować ostrożność.
- Dyrektywy wykazują wysoką swoistość względem struktury drzewa, które opisują. Dlatego każda zmiana danych bazowych lub opcji modelowania może spowodować, że działający wcześniej zestaw dyrektyw przestanie działać. Na przykład, jeśli algorytm CHAID zmieni podział dwuwymiarowy na podział trójwymiarowy na podstawie zaktualizowanych danych, dyrektywy oparte na poprzednim podziale (dwuwymiarowym) przestaną działać.

Uwaga: W przypadku wygenerowania modelu bezpośrednio (bez użycia konstruktora drzew) wszelkie dyrektywy drzewa są ignorowane.

Edycja dyrektyw

1. Aby wyświetlić lub edytować zapisane dyrektywy, otwórz węzeł budowania drzewa i wybierz panel Cele na karcie Opcje budowania.
2. Wybierz opcję **Drzewo interakcyjne**, aby aktywować elementy sterujące, wybierz opcję **Stosuj dyrektywy drzewa**, a następnie kliknij opcję **Dyrektywy**.

Składnia dyrektywy

Dyrektywy określają warunki wzrostu drzewa, począwszy od węzła głównego. Na przykład, aby rozwinąć drzewo o jeden poziom:

```
Grow Node Index 0 Children 1 2
```

Ponieważ nie określono predyktora, algorytm wybiera najlepszy podział.

Należy zwrócić uwagę, że pierwszy podział musi zawsze odbywać się na węźle głównym (Index 0) i muszą zostać określone wartości indeksu dla obu węzłów podrzędnych (w tym przypadku 1 i 2). Podanie instrukcji Grow Node Index 2 Children 3 4 jest niepoprawne, o ile wcześniej nie rozwinięto węzła głównego, który utworzył węzeł 2.

Aby rozwinąć drzewo:

```
Grow Tree
```

Aby rozwinąć i przyciąć drzewo (dotyczy tylko drzewa C&RT):

```
Grow_And_Prune Tree
```

Aby określić niestandardowy podział dla predyktora ilościowego:

```
Grow Node Index 0 Children 1 2 Spliton
  ( "EDUCATE", Interval ( NegativeInfinity, 12.5)
     Interval ( 12.5, Infinity ))
```

Aby określić podział dla predyktora nominalnego z dwiema wartościami:

```
Grow Node Index 2 Children 3 4 Spliton
  ( "GENDER", Group( "0.0" )Group( "1.0" ))
```

Aby określić podział dla predyktora nominalnego z wieloma wartościami:

```
Grow Node Index 6 Children 7 8 Spliton
   ( "ORGS", Group( "2.0","4.0" )
             Group( "0.0","1.0","3.0","6.0" ))
```

Aby określić podział dla predyktora porządkowego:

```
Grow Node Index 4 Children 5 6 Spliton
  ( "CHILDS", Interval ( NegativeInfinity, 1.0)
              Interval ( 1.0, Infinity ))
```

Uwaga: W przypadku określania podziałów niestandardowych w nazwach zmiennych i wartości (EDUCATE, GENDER, CHILDS itp.) rozróżniana jest wielkość liter.

Dyrektywy dla drzew CHAID

Dyrektywy dla drzew CHAID są szczególnie czułe na zmiany w danych lub w modelu, ponieważ — inaczej niż w przypadku drzew C&RT i QUEST — nie obowiązują w ich przypadku ograniczenia stosowania tylko podziałów binarnych. Na przykład poniższa składnia wygląda poprawnie, lecz jej wykonanie nie powiedzie się, jeśli algorytm podzieli węzeł główny na więcej niż dwa węzły podrzędne:

```
Grow Node Index 0 Children 1 2
Grow Node Index 1 Children 3 4
```

W przypadku węzła CHAID możliwe jest, że węzeł 0 będzie miał 3 lub 4 węzły podrzędne, co powoduje błąd przy wykonywaniu drugiego wiersza składni.

Używanie dyrektyw w skryptach

Dyrektywy można także osadzać w skryptach, korzystając z potrójnych cudzysłowów.

Aktualizowanie dyrektyw drzewa

Aby zabezpieczyć efekty swojej pracy na czas interaktywnej sesji budowania drzewa, można zapisać dyrektywy, których użyto do wygenerowania bieżącego drzewa. Inaczej niż w przypadku zapisywania modelu użytkowego, którego nie można już później edytować, opcja zapisania dyrektyw pozwala ponownie utworzyć drzewo w jego aktualnym stanie, do dalszej edycji.

Aby zaktualizować dyrektywy, z menu konstruktora drzewa wybierz:

Plik > Aktualizuj dyrektywy

Dyrektywy są zapisywane w węźle modelowania służącym do tworzenia drzewa (takim jak C&RT, QUEST czy CHAID) i mogą być używane do ponownego tworzenia bieżącego drzewa. Więcej informacji można znaleźć w temacie “Dyrektywy rozwijania drzewa” na stronie 92.

Eksportowanie informacji o modelu, korzyściach i ryzyku

Z konstruktora drzewa można wyeksportować model, korzyść i statystykę ryzyk do formatu tekstowego, HTML lub obrazu.

1. W oknie konstruktora drzewa wybierz kartę lub widok, który chcesz wyeksportować.
2. Z menu wybierz:

   Plik > Eksportuj
3. Wybierz odpowiedni format: **Tekst**, **HTML** lub **Wykres**, a następnie wybierz konkretne elementy, które chcesz wyeksportować z menu podrzędnego.

Eksport bazuje na aktualnie wybranych opcjach, o ile mają one zastosowanie.

Eksport do formatu tekstowego lub HTML. Istnieje możliwość wyeksportowania korzyści lub statystyk ryzyka dla podzbioru uczącego lub testującego (o ile taki został zdefiniowany). Eksport bazuje na bieżących wyborach dokonanych na karcie Korzyści — można na przykład wybrać statystykę węzeł po węźle, skumulowaną lub kwantyl.

Eksport do formatu graficznego. Bieżące drzewo można wyeksportować w postaci takiej, w jakiej jest wyświetlane na karcie Przegląd, lub można wyeksportować wykresy korzyści dla podzbioru uczącego lub testującego (o ile taki został zdefiniowany). Dostępne formaty to: *.JPEG*, *.PNG* i *.BMP*. W przypadku korzyści eksport bazuje na bieżących wyborach dokonywanych na karcie Korzyści (dostępnych wyłącznie podczas wyświetlania wykresu).

Generowanie węzłów filtrowania i selekcji

W oknie konstruktora drzew lub podczas przeglądania modelu użytkowego drzewa decyzyjnego z menu należy wybrać następujące opcje:

Utwórz > Węzeł filtrowania

lub

> Węzeł selekcji

Węzeł filtra. Generuje węzeł filtrujący wszelkie zmienne nieużywane przez bieżące drzewo. Jest to szybki sposób na zredukowanie zbioru danych tak, aby uwzględniał tylko zmienne wybrane jako istotne przez algorytm. Jeśli nad tym węzłem drzewa decyzyjnego znajduje się węzeł Typ, wówczas wszystkie zmienne w roli *Zmienna przewidywana* są przekazywane przez model użytkowy Filtr.

Węzeł wyboru. Generuje węzeł dokonujący selekcji wszystkich rekordów przypadających na ten węzeł. Ta opcja wymaga zaznaczenia na karcie Przegląd co najmniej jednej gałęzi drzewa.

Model użytkowy jest umieszczany w obszarze roboczym strumienia.

Generowanie zestawu reguł z drzewa decyzyjnego

Można wygenerować model użytkowy Zestaw reguł reprezentujący strukturę drzewa jako zestaw reguł definiujących gałęzie końcowe drzewa. Zestawy reguł często zachowują większość istotnych informacji z całego drzewa decyzyjnego, ale w postaci mniej złożonego modelu. Najważniejszą różnicą dotyczącą zestawu reguł jest to, że do dowolnego rekordu można zastosować więcej niż jedną regułę lub można nie stosować żadnej reguły. Można na przykład wyświetlić wszystkie reguły, które przewidują wynik *no*, po których następują wszystkie reguły, które przewidują wynik *yes*. Jeśli zastosowanie ma wiele reguł, każda z nich otrzymuje ważony „głos" w oparciu o ufność powiązaną z tą regułą; ostateczna decyzja dotycząca predykcji jest podejmowana poprzez połączenie ważonych głosów ze wszystkich reguł mających zastosowanie do danego rekordu. Jeśli żadna reguła nie ma zastosowania, do rekordu przypisywana jest domyślna predykcja.

Uwaga: Podczas oceniania zestawu reguł można zauważyć różnice w ocenianiu w porównaniu do oceny za pomocą drzewa; wynika to z faktu, że każda gałąź końcowa w drzewie jest oceniana niezależnie. Różnice mogą być zauważalne tylko w przypadku gdy w danych występują brakujące wartości.

Zestawy reguł można generować tylko za pośrednictwem drzew z przewidywanymi zmiennymi ilościowymi (nie jest to możliwe w przypadku drzew regresji).

W oknie konstruktora drzew lub podczas przeglądania modelu użytkowego drzewa decyzyjnego z menu należy wybrać następujące opcje:

Utwórz > Zestaw reguł

Nazwa zestawu reguł Umożliwia określenie nazwy nowego modelu użytkowego zestawu reguł.

Utwórz węzeł na Decyduje o lokalizacji nowego modelu użytkowego zestawu reguł. Można wybrać **Obszar roboczy**, **Paleta modeli** lub **Łącznie**.

Minimum wystąpień Umożliwia określenie minimalnej liczby wystąpień (liczby rekordów, do których reguła ma zastosowanie), aby reguła była zachowywana w modelu użytkowym zestawu reguł. Reguły, które będą obsługiwały mniej wystąpień od określonej wartości, nie będą uwzględniane w nowym zestawie reguł.

Minimalna ufność Umożliwia określenie minimalnej ufności dla reguł, aby były zachowywane w modelu użytkowym zestawu reguł. Reguły, które będą miały niższą ufność od określonej wartości, nie będą uwzględniane w nowym zestawie reguł.

Budowanie modelu drzewa bezpośrednio

Alternatywą wobec użycia interaktywnego konstruktora drzewa jest budowa modelu drzewa decyzyjnego bezpośrednio z węzła po uruchomieniu strumienia. Pozostaje to w zgodzie z większością pozostałych węzłów budowania modelu. W przypadku modeli drzewa C5.0 i Drzewo-AS, które nie są obsługiwane przez interaktywnego konstruktora drzewa, jest to jedyna dostępna metoda.

1. Utwórz strumień i dodaj jeden z węzłów drzew decyzyjnych — C&RT, CHAID, QUEST, C5.0 lub Drzewo-AS.
2. W przypadku węzłów C&RT, QUEST lub CHAID na panelu Cele na karcie Opcje budowania wybierz jeden z głównych celów. W przypadku wybrania opcji **Zbudować pojedyncze drzewo** upewnij się, że dla opcji **Tryb** wybrano wartość **Model**.

   W przypadku drzewa C5.0 na karcie Model ustaw dla opcji **Typ wyjściowy** wartość **Drzewo decyzyjne**.

   W przypadku drzewa Drzewo-AS na panelu Podstawowe wybierz typ **Algorytm wzrostu drzewa**.
3. Wybierz zmienne przewidywane oraz predykcyjne i wskaż dodatkowe opcje modelu, odpowiednio do potrzeb. Szczegółowe instrukcje zawiera dokumentacja dla każdego węzła budowania drzewa.
4. Uruchom strumień, aby wygenerować model.

Uwagi dotyczące budowania drzewa

- W przypadku budowania drzew tą metodą dyrektywy rozwijania drzewa są ignorowane.

- Obie metody tworzenia drzew decyzyjnych — interaktywna i bezpośrednia — generują ostatecznie podobne modele. Różnica polega jedynie na zakresie kontroli, jaką zachowuje użytkownik w trakcie procesu.

Węzły drzew decyzyjnych

Węzły Drzewo decyzyjne w programie IBM SPSS Modeler umożliwiają dostęp do następujących algorytmów budowania drzewa:

- C&RT
- QUEST
- CHAID
- C5.0
- Drzewo-AS
- Drzewa losowe

Więcej informacji można znaleźć w temacie “Modele drzew decyzyjnych” na stronie 83.

Algorytmy są do siebie podobne, jeśli chodzi o możliwość konstruowania drzew decyzyjnych przez rekursywny podział danych na mniejsze i jeszcze mniejsze podgrupy. Występują jednak między nimi także pewne istotne różnice.

Zmienne wejściowe. Zmienne wejściowe (predyktory) mogą być zmiennymi jednego z następujących typów (poziomu pomiaru): ilościowe, jakościowe, nominalne, porządkowe lub typu flaga.

Zmienne przewidywane. Możliwe jest wskazanie tylko jednej zmiennej przewidywanej. W przypadku algorytmów C&RT, CHAID, Drzewo-AS i Drzewa losowe zmienna przewidywana może być ilościowa, jakościowa, nominalna, porządkowa lub typu flaga. W przypadku algorytmu QUEST może być ona jakościowa, nominalna lub typu flaga. W przypadku algorytmu C5.0 zmienna przewidywana może być nominalna, porządkowa lub typu flaga.

Typ podziału. Algorytmy C&RT, QUEST i Drzewa losowe obsługują tylko podziały binarne (każdy węzeł w drzewie może zostać podzielony na nie więcej niż dwie gałęzie). Dla odmiany algorytmy CHAID, C5.0 i Drzewo-AS obsługują podział na więcej niż dwie gałęzie naraz.

Metoda podziału. Algorytmy różnią się między sobą, jeśli chodzi o kryteria decydujące o podziale. W przypadku predykcji wyniku jakościowego przez algorytm C&RT używana jest miara rozproszenia (domyślnie współczynnik Gini, można to jednak zmienić). W przypadku ilościowych zmiennych przewidywanych używana jest metoda odchylenia najmniejszych kwadratów. W przypadku algorytmów CHAID i Drzewo-AS stosowany jest test chi-kwadrat; w przypadku algorytmu QUEST dla predyktorów jakościowych stosowany jest test chi-kwadrat, zaś dla ilościowych danych wejściowych — analiza wariancji. W przypadku algorytmu C5.0 stosowana jest miara teoretyczna informacji, iloraz korzyści dla informacji.

Traktowanie braków danych. Wszystkie algorytmy dopuszczają braki danych dla zmiennych predykcyjnych, choć różnią się co do metody ich obsługi. W przypadku algorytmów C&RT i QUEST stosowane są odpowiednio do potrzeb substytuty zmiennych predykcyjnych, umożliwiające przeprowadzenie rekordu z brakami danych przez drzewo podczas uczenia. Algorytm CHAID tworzy z braków danych osobną kategorię i umożliwia użycie ich do budowy drzewa. Algorytm C5.0 stosuje metodę frakcjonowania, która przekazuje część ułamkową rekordu w dół każdej gałęzi drzewa od węzła, którego dotyczy podział oparty na zmiennej z brakiem danych.

Przycinanie. Algorytmy C&RT, QUEST i C5.0 oferują możliwość wzrostu drzewa w pełni, a następnie przycięcie go przez usunięcie podziałów dolnego poziomu, niewpływających znacząco na dokładność drzewa. Wszystkie algorytmy drzew decyzyjnych oferują jednak możliwość kontroli minimalnej wielkości podgrupy, co pomaga uniknąć gałęzi z niewielką liczbą rekordów danych.

Budowanie drzewa interaktywnego. Algorytmy C&RT, QUEST i CHAID oferują możliwość uruchomienia sesji interaktywnej. Pozwala to budować drzewo po jednym poziomie naraz, edytować podziały i przycinać drzewo przed utworzeniem modelu. Algorytmy C5.0, Drzewo-AS i Drzewa losowe nie oferują opcji wyników interaktywnych.

Prawdopodobieństwa a priori. Algorytmy C&RT i QUEST oferują możliwość określania prawdopodobieństw a priori dla kategorii podczas przewidywania zmiennej jakościowej. Prawdopodobieństwa a priori to oszacowania ogólnej względnej częstości dla każdej kategorii zmiennych przewidywanych w populacji, z której pochodzą dane uczące. Innymi słowy, są to oszacowania prawdopodobieństwa dla każdej możliwej przewidywanej wartości przed uzyskaniem jakichkolwiek informacji na temat wartości predykcyjnych. Algorytmy CHAID, C5.0, Drzewo-AS i Drzewa losowe nie oferują możliwości określania prawdopodobieństw a priori.

Zestaw reguł. Ta opcja jest niedostępna w przypadku algorytmu Drzewo-AS i Drzewa losowe. W przypadku modeli o przewidywanych zmiennych jakościowych węzły drzew decyzyjnych oferują możliwość utworzenia modelu w postaci zestawu reguł, który bywa niekiedy łatwiejszy w interpretacji, niż złożone drzewo decyzyjne. W przypadku algorytmów C&RT, QUEST i CHAID zestaw reguł można wygenerować w sesji interaktywnej; w przypadku algorytmu C5.0 opcję tę można wybrać w węźle modelowania. Ponadto wszystkie modele drzew decyzyjnych umożliwiają generowanie zestawu reguł na podstawie modelu użytkowego. Więcej informacji można znaleźć w temacie “Generowanie zestawu reguł z drzewa decyzyjnego” na stronie 95.

Węzeł C&RT

Węzeł drzewa klasyfikacji i regresji (C&RT) jest metodą klasyfikacji i predykcji w oparciu o drzewo. W metodzie tej, podobnie jak w algorytmie C5.0, stosuje się rekursywny podział rekordów uczących na segmenty o podobnych wartościach zmiennych przewidywanych. Działanie węzła C&RT rozpoczyna się od analizy zmiennych wejściowych w poszukiwaniu najlepszych podziałów, przy czym jakość podziału mierzona jest ograniczeniem wskaźnika zanieczyszczenia uzyskanego wskutek podziału. W wyniku podziału powstają dwie podgrupy, z których każda jest następnie dzielona na następne dwie podgrupy i tak dalej, aż do spełnienia kryterium zatrzymania. Wszystkie podziały są binarne (tylko na dwie podgrupy).

Przycinanie

W przypadku algorytmu C&RT możliwe jest najpierw zbudowanie dużego drzewa, a następnie przycięcie go z zastosowaniem algorytmu analizy kosztu i złożoności, który koryguje oszacowanie ryzyka na podstawie liczby węzłów końcowych. Ta metoda, która umożliwia rozrost drzewa przed przycięciem go na podstawie bardziej złożonych kryteriów, pozwala na uzyskanie mniejszych drzew, które lepiej poddają się walidacji krzyżowej. Zwiększenie ryzyka węzłów końcowych co do zasady zmniejsza ryzyko błędu w odniesieniu do bieżących danych (tj. danych uczących), ale faktyczne ryzyko może być wyższe, gdy model zostanie uogólniony dla danych nieznanych wcześniej. Wyobraźmy sobie skrajny przypadek, w którym dla każdego rekordu w zbiorze uczącym istnieje osobny węzeł końcowy. Oszacowanie ryzyka wyniosłoby 0%, ponieważ każdy rekord ma swój węzeł, ale ryzyko błędnej klasyfikacji na danych nieznanych (testowych) niemal na pewno byłoby większe od 0. Miara kosztu i kompletności jest próbą skompensowania tego zjawiska.

Przykład. Załóżmy, że operator telewizji kablowej zamówił badanie marketingowe mające ustalić, którzy klienci gotowi są kupić subskrypcję interaktywnego serwisu informacyjnego. Mając do dyspozycji dane z takiego badania, można utworzyć strumień, w którym zmienną przewidywaną jest zamiar zakupu subskrypcji, a predyktorami są wiek, płeć, wykształcenie, kategoria dochodów, liczba godzin spędzanych dziennie przed telewizorem i liczba dzieci. Stosując w strumieniu węzeł C&RT, można przewidywać i klasyfikować odpowiedzi, aby uzyskać jak najwyższy wskaźnik pozytywnych reakcji na kampanię.

Wymagania. Do uczenia modelu C&RT potrzeba co najmniej jednej zmiennej *wejściowej* i dokładnie jednej zmiennej *przewidywanej*. Zmienne przewidywana i wejściowa mogą być ilościowe (przedział liczbowy) lub jakościowe. Zmienne o roli *Łącznie* lub *Żadna* są ignorowane. Typy wszystkich zmiennych używane w modelu muszą być zrealizowane jako instancje zmiennych, a wszelkie zmienne porządkowe (uporządkowany zestaw) stosowane w modelu muszą być przechowywane jako liczby (nie łańcuchy). W razie potrzeby do ich przekształcenia można użyć węzła rekodowania.

Mocne strony. Modele C&RT wykazują się dużą odpornością na takie problemy, jak brak danych i duża liczba zmiennych. Zwykle nie wymagają długiego uczenia, by generować prawidłowe oszacowania. Ponadto modele C&RT

bywają bardziej zrozumiałe niż modele innego typu, ponieważ reguły wywiedzione z modelu dają się bardzo łatwo i bezpośrednio zinterpretować. W odróżnieniu od węzła C5.0, węzeł C&RT obsługuje zarówno zmienne przewidywane typu ilościowego, jak i jakościowego.

Węzeł CHAID

CHAID (ang. Chi-squared Automatic Interaction Detection) to metoda klasyfikacji umożliwiająca budowanie drzew decyzyjnych z użyciem statystyki chi-kwadrat w celu identyfikacji optymalnych podziałów.

CHAID bada najpierw tabele krzyżowe między każdą ze zmiennych wejściowych a wynikiem oraz testuje istotność za pomocą testu niezależności chi-kwadrat. Jeśli więcej niż jedna z tych relacji jest statystycznie znacząca, CHAID wybierze najbardziej znaczącą zmienną wejściową (o najmniejszej wartości *p*). Jeśli dane wejściowe należą do dwu lub większej liczby kategorii, są one porównywane, a kategorie niewykazujące różnic w wynikach są zwijane razem. Realizuje się to przez sukcesywne łączenie par kategorii wykazujących najmniej znaczące różnice. Ten proces scalania kategorii jest zatrzymywany w chwili, gdy wszystkie pozostałe kategorie różnią się na danym poziomie testowania. W przypadku wejściowych zmiennych nominalnych można scalać dowolne kategorie; w przypadku zestawu porządkowego możliwe jest scalenie tylko kategorii zmiennych ilościowych.

Wyczerpujący CHAID stanowi modyfikację CHAID umożliwiającą dokładniejsze badanie wszystkich możliwych podziałów dla każdego predyktora, lecz obliczenia w jego przypadku zajmują więcej czasu.

Wymagania. Zmienne przewidywane i wejściowe mogą być ilościowe lub jakościowe; węzły mogą być dzielone na dwie lub więcej podgrup na każdym poziomie. Wszelkie zmienne porządkowe stosowane w modelu muszą charakteryzować się składowaniem typu numerycznego (nie łańcuchowego). W razie potrzeby do ich przekształcenia można użyć węzła rekodowania.

Mocne strony. W odróżnieniu od węzłów C&RT i węzłów QUEST, CHAID może generować drzewa niebinarne, co oznacza, że niektóre podziały mają więcej niż dwie gałęzie. Oznacza to tendencję do tworzenia szerszych drzew, niż w przypadku binarnych metod wzrostu. CHAID działa w przypadku wszystkich typów danych wejściowych, i akceptuje zarówno wagi obserwacji, jak i zmienne częstości.

Węzeł QUEST

QUEST — z ang. Quick, Unbiased, Efficient Statistical Tree (szybkie, nieobciążone, wydajne drzewo statystyczne) — to metoda klasyfikacji binarnej służąca do budowania drzew decyzyjnych. Główną motywacją jego opracowania było skrócenie czasu przetwarzania niezbędnego do analiz dużych drzew decyzyjnych C&R z wieloma zmiennymi lub z wieloma obserwacjami. Drugim celem stworzenia drzewa QUEST było zmniejszenie tendencji obecnej w metodach drzew klasyfikacji do preferowania danych wejściowych umożliwiających więcej podziałów, to jest, ilościowych zmiennych wejściowych (zakresów liczbowych) lub zmiennych z wieloma kategoriami.

- W drzewie QUEST stosowana jest, bazująca na istotności testów, sekwencja reguł umożliwiająca ocenę zmiennych wejściowych w węźle. Do celów wyboru dla każdych danych wejściowych w danym węźle może być konieczne przeprowadzenie zaledwie jednego testu. Inaczej niż w przypadku drzewa C&RT, podziały nie są oceniane, i inaczej niż w przypadku drzew K&R i CHAID, podczas kwalifikacji zmiennej wejściowej do wyboru nie są testowane kombinacje kategorii. Pozwala to skrócić czas analizy.
- Podziały są wyznaczane przez kwadratową analizę dyskryminacyjną z użyciem wybranych danych wejściowych w grupach tworzonych przez kategorie zmiennych przewidywanych. Ta metoda również skutkuje skróceniem czasu znajdowania optymalnego podziału względem wyszukiwania dokładnego (C&RT).

Wymagania. Zmienne wejściowe mogą być zakresami liczbowymi (ilościowymi), lecz zmienna przewidywana musi być jakościowa. Wszystkie podziały są binarne. Nie można stosować zmiennych ważących. Wszelkie zmienne porządkowe (uporządkowany zestaw) stosowane w modelu muszą charakteryzować się składowaniem typu numerycznego (nie łańcuchowego). W razie potrzeby do ich przekształcenia można użyć węzła rekodowania.

Mocne strony. Podobnie jak CHAID, lecz inaczej niż C&RT, drzewo QUEST używa testów statystycznych do decydowania o tym, czy zmienna wejściowa jest, czy nie jest, używana. Oddziela ono także problemy związane z wyborem danych wejściowych i podziałów, stosując do każdego z nich inne kryteria. Stoi to w kontraście z CHAID, w

przypadku którego wynik testu statystycznego określający wybór zmiennej jednocześnie generuje podział. Podobnie, C&RT korzysta z miary zanieczyszczenie-zmiana, umożliwiając zarówno wybór zmiennej wejściowej, jak i określenie podziału.

Opcje zmiennych węzła Drzewo decyzyjne

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Tę opcję należy wybrać, aby ręcznie przypisać zmienne przewidywane, predyktory i inne role.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji.

Predyktory (dane wejściowe). Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Waga analizy. (Tylko CHAID, C&RT i Drzewo-AS) Należy tutaj wskazać zmienną, która ma być używana jako waga obserwacji. Wagi obserwacji są stosowane w celu uwzględniania różnic w wariancji między poziomami zmiennej wyjściowej. Więcej informacji można znaleźć w temacie “Użycie zmiennych częstości i ważących” na stronie 33.

Opcje budowania węzła Drzewo decyzyjne

Na karcie Opcje budowania można ustawić wszystkie opcje budowania modelu. Można oczywiście po prostu kliknąć przycisk **Uruchom** w celu zbudowania modelu z wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik potrzebuje zwykle dostosować proces budowy do swoich celów.

Karta zawiera kilka różnych okien, w których można dostosować ustawienia odpowiednio do specyfiki własnego modelu.

Węzły drzew decyzyjnych — cele

W przypadku węzłów C&RT, QUEST i CHAID w panelu Cele na karcie Opcje budowania można wybrać budowę nowego modelu lub aktualizację istniejącego. Można również ustawić główny cel węzła: budowa modelu standardowego, budowa modelu o zwiększonej dokładności lub stabilności, albo budowa modelu dla dużych zbiorów danych.

Co zamierzasz zrobić?

Zbuduj nowy model. (Ustawienie domyślne) Tworzy całkowicie nowy model przy każdym uruchomieniu strumienia zawierającego ten węzeł modelowania.

Kontynuuj uczenie istniejącego modelu. Domyślnie po każdym wykonaniu węzła modelowania tworzony jest całkowicie nowy model. Jeśli ta opcja jest zaznaczona, uczenie jest kontynuowane z użyciem ostatniego modelu pomyślnie utworzonego przez węzeł. Dzięki temu możliwa jest aktualizacja lub odświeżenie istniejącego modelu bez konieczności uzyskania dostępu do oryginalnych danych, co może skutkować znacznie bardziej wydajnym działaniem, ponieważ *tylko* nowe lub zaktualizowane rekordy są podawane do strumienia. Szczegóły dotyczące poprzedniego

modelu są zapisywane z węzłem modelowania, umożliwiając używanie tej opcji nawet, jeśli poprzednie wartościowe informacje z modelu są już niedostępne w strumieniu lub w palecie Modeli.

Uwaga: Ta opcja jest aktywna tylko, jeśli jako cel wybrano **Zbudować pojedyncze drzewo** (w przypadku węzłów C&RT, CHAID i QUEST), **Zbudować model standardowy** (w przypadku sieci neuronowej i liniowej) lub **Utworzyć model dla dużych zbiorów danych**.

Jaki chcesz osiągnąć cel?

- **Zbudować pojedyncze drzewo.** Tworzy standardowy, pojedynczy model drzewa decyzyjnego. Ogólnie rzecz biorąc, standardowe modele są łatwiejsze w interpretacji, a ponadto ich ocena może okazać się szybsza, niż modeli budowanych z użyciem innych celów.

  Uwaga: Aby użyć tej opcji w modelach rozdzielonych razem z opcją **Kontynuować uczenie istniejącego modelu**, należy mieć połączenie z produktem Analytic Server.

  Dominanta. Określa metodę używaną do budowy modelu. Opcja **Model** powoduje utworzenie modelu automatycznie, po uruchomieniu strumienia. Opcja **Drzewo interakcyjne** powoduje otwarcie konstruktora drzewa, umożliwiającego tworzenie jednego poziomu drzewa naraz, edytowanie podziałów oraz przycinanie odpowiednio do potrzeb przed utworzeniem modelu użytkowego.

  Stosuj dyrektywy drzewa. Wybór tej opcji pozwala określić dyrektywy, które mają być stosowane podczas generowania drzewa interaktywnego z węzła. Na przykład można wskazać podziały pierwszego i drugiego poziomu, a zostaną one automatycznie zastosowane po uruchomieniu konstruktora drzewa. Można także zapisać dyrektywy z interaktywnej sesji budowania drzewa w celu ponownego utworzenia drzewa w przyszłości. Więcej informacji można znaleźć w temacie “Aktualizowanie dyrektyw drzewa” na stronie 94.
- **Zwiększyć dokładność modelu (boosting).** Tę opcję należy wybrać, aby użyć metody specjalnej, zwanej **boostingiem**, do poprawy wskaźnika dokładności modelu. Boosting polega na budowie kolejno wielu modeli. Pierwszy model jest budowany w zwykły sposób. Drugi model jest budowany w taki sposób, że koncentruje się on na rekordach, które zostały błędnie sklasyfikowane przez pierwszy model. Trzeci model koncentruje się na błędach drugiego modelu, i tak dalej. Ostatecznie obserwacje są klasyfikowane przez zastosowanie do nich całego zbioru modeli, z użyciem procedury głosowania ważonego w celu połączenia poszczególnych predykcji w jedną predykcję ogólną. Boosting może znacząco poprawić dokładność modelu drzewa decyzyjnego, wymaga też jednak dłuższego okresu uczenia.
- **Wzmocnić stabilność modelu (agregacja bootstrapowa).** Tę opcję należy wybrać, aby użyć metody specjalnej, znanej także jako **agregacja bootstrap** do poprawy stabilności modelu, a zarazem w celu uniknięcia przeuczenia. Opcja ta tworzy wiele modeli i łączy je w celu uzyskania bardziej wiarygodnych predykcji. Budowa i ocena modeli uzyskanych za pomocą tej opcji może zająć więcej czasu.
- **Utworzyć model dla dużych zbiorów danych.** Tę opcję należy wybrać w przypadku pracy ze zbiorami danych, które są zbyt duże, aby możliwe było zbudowanie dla nich modelu z użyciem którejkolwiek z pozostałych opcji. Opcja ta dzieli dane na mniejsze bloki danych i buduje model dla każdego bloku. Następnie następuje automatyczna selekcja najdokładniejszych modeli i łączenie ich w jeden model użytkowy. Zaznaczenie na tym ekranie opcji **Kontynuuj uczenie istniejącego modelu** pozwala na przyrostową aktualizację modelu.

  Uwaga: Ta opcja w przypadku bardzo dużych zbiorów danych wymaga połączenia z programem IBM SPSS Modeler Server.

Węzły drzew decyzyjnych — Podstawowe

Można tutaj określić podstawowe opcje określające sposób budowania drzewa decyzyjnego.

Algorytm wzrostu drzewa (tylko CHAID i Drzewo-AS) Umożliwia wybór typu algorytmu **CHAID**, który ma być używany. **Wyczerpujący CHAID** stanowi modyfikację CHAID umożliwiającą dokładniejsze badanie wszystkich możliwych podziałów dla każdego predyktora, lecz obliczenia w jego przypadku zajmują więcej czasu.

Maksymalna głębokość drzewa Umożliwia określenie maksymalnej liczby poziomów poniżej węzła głównego (liczby rekurencyjnych podziałów próby). Wartość domyślna to 5; należy wybrać opcję **Użytkownika** i wprowadzić wartość określającą inną liczbę poziomów.

Przycinanie (tylko C&RT i QUEST)

Przytnij drzewo, aby uniknąć przeuczenia Przycinanie składa się z usuwania podziałów dolnego poziomu, niewpływających istotnie na dokładność drzewa. Przycinanie może pomóc w uproszczeniu drzewa, ułatwiając jego interpretację oraz, w niektórych przypadkach, poprawiając uogólnienie. W przypadku potrzeby zachowania całego drzewa, bez przycinania, należy pozostawić tę opcję niezaznaczoną.

- **Ustaw maksymalną różnicę w ryzyku (w błędach standardowych)** Ta opcja umożliwia określenie bardziej liberalnej reguły przycinania. Reguła błędu standardowego umożliwia wybór przez algorytm najprostszego drzewa, którego oszacowanie ryzyka jest bliskie (lecz najprawdopodobniej większe) niż dla drzewa podrzędnego o najmniejszym ryzyku. Wartość ta wskazuje wielkość dozwolonej różnicy w oszacowaniu ryzyka między przyciętym drzewem a drzewem o najmniejszym ryzyku w warunkach oszacowania ryzyka. Na przykład w przypadku wskazania wartości 2 może zostać wybrane drzewo, którego oszacowanie ryzyka jest (2 × błąd standardowy) większe niż to dla pełnego drzewa.

Maksimum substytutów. Substytuty to metoda postępowania w przypadku braków danych. Dla każdego podziału w drzewie algorytm identyfikuje zmienne wejściowe najbardziej podobne do wybranej zmiennej podziału. Zmienne te stanowią *substytuty* dla tego podziału. Jeśli rekord wymaga sklasyfikowania, lecz występuje w nim brak danych dla zmiennej podziału, podział może zostać dokonany z użyciem wartości w zmiennej substytutu rekordu. Zwiększenie tej wartości zwiększy elastyczność postępowania z brakami danych, lecz może także spowodować zwiększenie wykorzystania pamięci i wydłużenie czasów uczenia.

Węzły drzew decyzyjnych — reguły zatrzymujące

Te opcje pozwalają sterować procesem tworzenia drzewa. Reguły zatrzymujące określają, kiedy ma nastąpić zatrzymanie podziału określonych gałęzi drzewa. Należy ustawić minimalną wielkość gałęzi, aby uniknąć podziału na bardzo małe podgrupy. Opcja **Minimum rekordów w gałęzi nadrzędnej** uniemożliwi podział, jeśli liczba rekordów w węźle do podziału (*nadrzędny*) będzie mniejsza niż określona wartość. Opcja **Minimum rekordów w gałęzi podrzędnej** uniemożliwi podział, jeśli liczba rekordów w dowolnej gałęzi utworzonej w wyniku podziału (*podrzędna*) będzie mniejsza niż określona wartość.

- **Wartość procentowa** Umożliwia określenie wielkości jako wartości procentowej wszystkich danych uczących.
- **Wartość bezwzględna** Umożliwia określenie wielkości jako bezwzględnej liczby rekordów.

Węzły drzew decyzyjnych — zespoły

Ustawienia te determinują zachowanie tworzenia zespołów, które występuje, gdy w Celach pożądane jest wspomaganie, agregacja metodą bootstrap lub bardzo duże zbiory danych. Opcje, które nie mają zastosowania do wybranego celu są ignorowane.

Agregacja metodą bootstrap i bardzo duże zbiory danych. Podczas wybierania zestawu jest to reguła służąca do łączenia przewidywanych wartości z modeli podstawowych w celu wyliczenia wartości oceny zestawu.

- **Domyślna reguła zespolenia dla przewidywanych zmiennych jakościowych.** Przewidywane wartości zestawu dla przewidywanych zmiennych jakościowych mogą być połączone przy pomocy głosowania, największego prawdopodobieństwa lub największego, średniego prawdopodobieństwa. **Głosowanie** wybiera kategorię, która ma największe prawdopodobieństwo, najczęściej wśród modeli podstawowych. **Największe prawdopodobieństwo** wybiera kategorię, która uzyskuje największe, pojedyncze prawdopodobieństwo wśród modeli podstawowych. **Największe średnie prawdopodobieństwo** wybiera kategorię z najwyższą wartością, gdy prawdopodobieństwa kategorii wśród modeli podstawowych są uśrednione.
- **Domyślna reguła zespolenia dla docelowych wartości ilościowych.** Przewidywane wartości zestawu dla jakościowych zmiennych docelowych można połączyć przy pomocy średniej lub mediany przewidywanych wartości z modeli podstawowych.

Należy zwrócić uwagę, że gdy celem jest zwiększenie dokładności modelu, wybory reguły łączenia są ignorowane. Wzmocnienie zawsze wykorzystuje głos ważonej większości do oceny jakościowych zmiennych docelowych i ważonej mediany do oceny jakościowych zmiennych docelowych.

Boosting i agregacja bootstrapowa. Podaj liczbę modeli podstawowych do utworzenia, gdy celem jest zwiększenie dokładności lub stabilności modelu; dla agregacji metodą bootstrap jest to liczba prób agregacji metodą bootstrap. Powinna to być dodatnia liczba całkowita.

Węzły C&RT i QUEST — koszty błędnej klasyfikacji i prawdopodobieństwa a priori

Koszty błędnej klasyfikacji

W niektórych kontekstach pewne błędy są bardziej kosztowne od innych. Przykładowo, bardziej kosztowne może być sklasyfikowanie osób składających wniosek kredytowy z wysokim poziomem ryzyka jako osób z niskim poziomem ryzyka (jeden rodzaj błędu) niż sklasyfikowanie osób składających wniosek z niskim poziomem ryzyka jako osób z wysokim poziomem ryzyka (inny rodzaj błędu). Koszty błędnej klasyfikacji umożliwiają określenie względnej ważności różnych rodzajów błędów predykcji.

Kosztami błędnej klasyfikacji zwykle są wagi zastosowane do określonych danych wynikowych. Wagi te są uwzględniane w modelu i rzeczywiście mogą zmienić predykcję (jako sposób ochrony przed kosztownymi błędami).

Z wyjątkiem modeli C5.0 koszty błędnej klasyfikacji nie mają zastosowania podczas oceniania modelu i nie są brane pod uwagę podczas rangowania lub porównywania modeli za pomocą węzła Auto Klasyfikacja, wykresu ewaluacyjnego lub węzła analizy. Model, który uwzględnia koszty, może nie wygenerować mniejszej liczby błędów niż ten, który ich nie uwzględnia, i może nie mieć wyższej rangi pod względem całkowitej dokładności, ale prawdopodobnie lepiej sprawdzi się w warunkach praktycznych, ponieważ generuje błędy *mniej kosztowne*.

Macierz kosztów przedstawia koszty dla każdej możliwej kombinacji przewidywanej kategorii rzeczywistej. Domyślnie wszystkie koszty błędnej klasyfikacji są ustawione na wartość 1,0. Aby wprowadzić niestandardowe wartości kosztów, należy wybrać opcję **Stosuj koszty błędnej klasyfikacji** i wprowadzić do macierzy kosztów niestandardowe wartości.

Aby zmienić koszt błędnej klasyfikacji, należy zaznaczyć komórkę odpowiadającą odpowiedniej kombinacji wartości przewidywanych i rzeczywistych, usunąć istniejącą zawartość komórki i wprowadzić do niej żądany koszt. Koszty nie są automatycznie symetryczne. Przykładowo, jeśli koszt błędnej klasyfikacji *A* jako *B* zostanie ustawiony na 2,0, koszt błędnej klasyfikacji *B* jako *A* nadal będzie miał domyślną wartość 1,0, chyba że zostanie ona również jawnie zmieniona.

Prawdopodobieństwa a priori

Opcje te umożliwiają określenie prawdopodobieństwa a priori dla kategorii podczas predykcji przewidywanej zmiennej jakościowej. **Prawdopodobieństwa a priori** to oszacowania ogólnej względnej częstości dla każdej kategorii zmiennej przewidywanej w populacji, z której pochodzą dane uczące. Inaczej ujmując, są to oszacowania prawdopodobieństwa wykonywane dla każdej możliwej przewidywanej wartości *przed* uzyskaniem jakichkolwiek informacji na temat wartości predykcyjnych. Istnieją trzy metody ustawienia prawdopodobieństwa a priori:

- **W oparciu o dane uczące.** Jest to ustawienie domyślne. Prawdopodobieństwa a priori określane są w oparciu o względne częstości kategorii w danych uczących.
- **Równe dla wszystkich klas.** Prawdopodobieństwa a priori dla wszystkich kategorii są definiowane jako $1/k$, gdzie k to liczba kategorii zmiennej przewidywanej.
- **Użytkownika.** Użytkownik może określić własne prawdopodobieństwa a priori. Wartości początkowe dla prawdopodobieństw a priori są ustawiane jako równe dla wszystkich klas. Można dostosować prawdopodobieństwo dla poszczególnych kategorii, ustawiając wartości zdefiniowane przez użytkownika. Aby ustawić określone prawdopodobieństwo dla kategorii, należy wybrać komórkę prawdopodobieństwa w tabeli odpowiadającej żądanej kategorii, usunąć zawartość komórki i wprowadzić żądaną wartość.

Suma prawdopodobieństw a priori dla wszystkich kategorii powinna wynosić 1,0 (**ograniczenie prawdopodobieństwa**). Jeśli suma nie wynosi 1,0, wyświetlane jest ostrzeżenie, z opcją automatycznego znormalizowania wartości. To automatyczne dostosowanie zachowuje proporcje we wszystkich kategoriach,

wymuszając ograniczenie prawdopodobieństwa. Takie dostosowanie można przeprowadzić w dowolnym czasie, klikając przycisk **Normalizuj**. Aby w tabeli ponownie ustawiać jednakowe wartości dla wszystkich kategorii, należy kliknąć przycisk **Wyrównaj**.

Korekta prawdopodobieństw a priori z użyciem kosztów błędnej klasyfikacji Ta opcja umożliwia dostosowanie prawdopodobieństw wstępnych w oparciu o koszty błędnej klasyfikacji (określone na karcie Koszty). Umożliwia to umieszczenie informacji o kosztach bezpośrednio w procesie wzrostu drzew korzystających z miary zanieczyszczenia Twoing. (Jeśli ta opcja nie zostanie zaznaczona, informacje o koszcie zostaną użyte wyłącznie do klasyfikowania rekordów i obliczania oszacowań ryzyka dla drzew w oparciu o miarę Twoing).

Węzeł CHAID — Koszty

W niektórych kontekstach pewne błędy są bardziej kosztowne od innych. Przykładowo, bardziej kosztowne może być sklasyfikowanie osób składających wniosek kredytowy z wysokim poziomem ryzyka jako osób z niskim poziomem ryzyka (jeden rodzaj błędu) niż sklasyfikowanie osób składających wniosek z niskim poziomem ryzyka jako osób z wysokim poziomem ryzyka (inny rodzaj błędu). Koszty błędnej klasyfikacji umożliwiają określenie względnej ważności różnych rodzajów błędów predykcji.

Kosztami błędnej klasyfikacji zwykle są wagi zastosowane do określonych danych wynikowych. Wagi te są uwzględniane w modelu i rzeczywiście mogą zmienić predykcję (jako sposób ochrony przed kosztownymi błędami).

Z wyjątkiem modeli C5.0 koszty błędnej klasyfikacji nie mają zastosowania podczas oceniania modelu i nie są brane pod uwagę podczas rangowania lub porównywania modeli za pomocą węzła Auto Klasyfikacja, wykresu ewaluacyjnego lub węzła analizy. Model, który uwzględnia koszty, może nie wygenerować mniejszej liczby błędów niż ten, który ich nie uwzględnia, i może nie mieć wyższej rangi pod względem całkowitej dokładności, ale prawdopodobnie lepiej sprawdzi się w warunkach praktycznych, ponieważ generuje błędy *mniej kosztowne*.

Macierz kosztów przedstawia koszty dla każdej możliwej kombinacji przewidywanej kategorii rzeczywistej. Domyślnie wszystkie koszty błędnej klasyfikacji są ustawione na wartość 1,0. Aby wprowadzić niestandardowe wartości kosztów, należy wybrać opcję **Stosuj koszty błędnej klasyfikacji** i wprowadzić do macierzy kosztów niestandardowe wartości.

Aby zmienić koszt błędnej klasyfikacji, należy zaznaczyć komórkę odpowiadającą odpowiedniej kombinacji wartości przewidywanych i rzeczywistych, usunąć istniejącą zawartość komórki i wprowadzić do niej żądany koszt. Koszty nie są automatycznie symetryczne. Przykładowo, jeśli koszt błędnej klasyfikacji *A* jako *B* zostanie ustawiony na 2,0, koszt błędnej klasyfikacji *B* jako *A* nadal będzie miał domyślną wartość 1,0, chyba że zostanie ona również jawnie zmieniona.

Węzeł C&RT — Zaawansowane

Opcje zaawansowane umożliwiają precyzyjne dostosowywanie procesu budowy drzewa do potrzeb użytkownika.

Minimalna zmiana zanieczyszczenia. Umożliwia określenie minimalnej zmiany zanieczyszczenia powodującej utworzenie w drzewie nowego podziału. **Zanieczyszczenie** odnosi się do zakresu, w jakim podgrupy definiowane przez drzewo mają szeroki rozstęp wartości zmiennych wyjściowych w ramach każdej grupy. W przypadku zmiennych jakościowych węzeł uważa się za „czysty”, jeśli 100% obserwacji w tym węźle przypada na konkretną kategorię zmiennej przewidywanej. Celem budowy drzewa jest utworzenie podgrup o podobnych wartościach wyjściowych — innymi słowy, zminimalizowanie zanieczyszczeń w ramach każdego węzła. Jeśli najlepszy podział dla gałęzi zmniejsza zanieczyszczenie o mniej niż określoną wartość, podział nie zostanie dokonany.

Miara zanieczyszczenia dla przewidywanych zmiennych jakościowych. W przypadku przewidywanych zmiennych jakościowych należy wskazać metodę używaną do pomiaru zanieczyszczenia drzewa. (W przypadku przewidywanych zmiennych ilościowych ta opcja jest ignorowana i jako miara zanieczyszczenia używane jest zawsze **najmniejsze odchylenie kwadratowe**).

- **Gini** to ogólna miara zanieczyszczenia bazująca na prawdopodobieństwach członkostwa w kategorii dla gałęzi.
- **Twoing** to miara zanieczyszczenia uwypuklająca podział binarny i z większym prawdopodobieństwem prowadząca do uzyskania w wyniku podziału w przybliżeniu równomiernych, jeśli chodzi o wielkość, gałęzi.

- **Twoing porządkowy** wnosi dodatkowe ograniczenie polegające na grupowaniu razem tylko klas ilościowych zmiennych przewidywanych, i ma zastosowanie tylko w przypadku porządkowych zmiennych przewidywanych. W przypadku wyboru tej opcji dla nominalnej zmiennej przewidywanej domyślnie używana jest standardowa miara Twoing.

Zbiór zabezpieczający przed przeuczeniem. Algorytm wewnętrznie rozdziela rekordy między podzbiór budowania modelu oraz zbiór zabezpieczający przed przeuczeniem, który jest niezależnym zbiorem rekordów danych używanym do śledzenia błędów podczas uczenia i zapobiegania modelowaniu przez metodę zmienności prawdopodobieństwa w danych. Należy określić procent rekordów. Domyślną wartością jest 30.

Replikacja wyników. Ustawienie wartości początkowej generatora liczb losowych umożliwia powielenie analizy. Podaj liczbę całkowitą lub kliknij przycisk **Generuj**, co spowoduje utworzenie pseudolosowej liczby całkowitej między 1 a 2147483647, włącznie.

Węzeł QUEST — Zaawansowane

Opcje zaawansowane umożliwiają precyzyjne dostosowywanie procesu budowy drzewa do potrzeb użytkownika.

Poziom istotności dla podziału. Określa poziom istotności (alpha) dla dzielenia węzłów. Wartość musi należeć do zakresu od 0 do 1. Niższe wartości mają tendencję do tworzenia drzew z mniejszą liczbą węzłów.

Zbiór zabezpieczający przed przeuczeniem. Algorytm wewnętrznie rozdziela rekordy między podzbiór budowania modelu oraz zbiór zabezpieczający przed przeuczeniem, który jest niezależnym zbiorem rekordów danych używanym do śledzenia błędów podczas uczenia i zapobiegania modelowaniu przez metodę zmienności prawdopodobieństwa w danych. Należy określić procent rekordów. Domyślną wartością jest 30.

Replikacja wyników. Ustawienie wartości początkowej generatora liczb losowych umożliwia powielenie analizy. Podaj liczbę całkowitą lub kliknij przycisk **Generuj**, co spowoduje utworzenie pseudolosowej liczby całkowitej między 1 a 2147483647, włącznie.

Węzeł CHAID — Zaawansowane

Opcje zaawansowane umożliwiają precyzyjne dostosowywanie procesu budowy drzewa do potrzeb użytkownika.

Poziom istotności dla podziału. Określa poziom istotności (alpha) dla dzielenia węzłów. Wartość musi należeć do zakresu od 0 do 1. Niższe wartości mają tendencję do tworzenia drzew z mniejszą liczbą węzłów.

Poziom istotności dla łączenia. Określa poziom istotności (alpha) dla łączenia kategorii. Wartość musi być większa od 0 i mniejsza lub równa 1. Aby uniemożliwić łączenie kategorii, należy podać wartość 1. W przypadku ilościowych zmiennych przewidywanych oznacza to, że liczba kategorii dla zmiennej w drzewie końcowym odpowiada podanej liczbie przedziałów. Opcja ta jest niedostępna w przypadku modeli Wyczerpujący CHAID.

Skoryguj wartości istotności metodą Bonferroniego. Ta opcja powoduje korygowanie wartości istotności podczas testowania różnych kombinacji kategorii predyktora. Wartości są korygowane w oparciu o liczbę testów, które bezpośrednio odnoszą się do liczby kategorii i poziomów pomiaru predyktora. Jest to zwykle pożądane, ponieważ zapewnia lepszą kontrolę wskaźnika wyników fałszywie dodatnich. Wyłączenie tej opcji powoduje zwiększenie możliwości analitycznych w zakresie znajdowania prawdziwych różnic, lecz odbywa się to kosztem zwiększenia wskaźnika wyników fałszywie dodatnich. Wyłączenie tej opcji może być zalecane w szczególności w przypadku niewielkich prób.

Zezwalaj na podział kategorii połączonych w węźle. Algorytm CHAID próbuje połączyć kategorie w celu utworzenia najprostszego drzewa opisującego model. Zaznaczenie tej opcji umożliwia ponowny podział połączonych kategorii, o ile efektem jest lepsze rozwiązanie.

Chi-kwadrat dla przewidywanych zmiennych jakościowych. W przypadku zmiennych jakościowych można określić metodę używaną do obliczania statystyki chi-kwadrat.

- **Pearsona.** Ta metoda skraca czas obliczeń, lecz należy zachować ostrożność w przypadku stosowania jej do niewielkich prób.

- **Iloraz wiarygodności.** Ta metoda jest bardziej odporna niż metoda Pearsona, lecz wydłuża czas obliczeń. Jest to metoda preferowana w przypadku niewielkich prób. Metoda ta jest używana zawsze w przypadku ilościowych zmiennych przewidywanych.

Minimalna zmiana w oczekiwanych częstościach komórek. Podczas szacowania częstości komórek (zarówno w przypadku modelu nominalnego, jak i modelu porządkowego efektów dla wierszy) używana jest procedura iteracyjna (epsilon) pozwalająca uzyskać zbieżność dla optymalnego oszacowania używanego w teście chi-kwadrat dla określonego podziału. Epsilon determinuje, w jakim stopniu zmiana musi wystąpić, aby iteracje mogły być kontynuowane; jeśli zmiana w porównaniu z ostatnią iteracją jest mniejsza niż zadana wartość, iteracje są zatrzymywane. W przypadku napotkania problemów związanych z brakiem zbieżności algorytmu można zwiększyć tę wartość lub zwiększyć maksymalną liczbę iteracji do wystąpienia zbieżności.

Maksimum iteracji dla uzyskania zbieżności. Określa maksymalną liczbę iteracji przed zatrzymaniem, niezależnie od tego, czy uzyskano zbieżność, czy nie.

Zbiór zabezpieczający przed przeuczeniem. (Ta opcja jest dostępna tylko w przypadku użycia interaktywnego konstruktora drzewa). Algorytm wewnętrznie rozdziela rekordy między podzbiór budowania modelu oraz zbiór zabezpieczający przed przeuczeniem, który jest niezależnym zbiorem rekordów danych używanym do śledzenia błędów podczas uczenia i zapobiegania modelowaniu przez metodę zmienności prawdopodobieństwa w danych. Należy określić procent rekordów. Domyślną wartością jest 30.

Replikacja wyników. Ustawienie wartości początkowej generatora liczb losowych umożliwia powielenie analizy. Podaj liczbę całkowitą lub kliknij przycisk **Generuj**, co spowoduje utworzenie pseudolosowej liczby całkowitej między 1 a 2147483647, włącznie.

Opcje modelu węzła Drzewo decyzyjne

Na karcie Opcje modelu można zdecydować o wyborze nazwy dla modelu lub o jej wygenerowaniu automatycznie. Można także zdecydować się na uzyskanie informacji o ważności predyktora, a także surowych i skorygowanych ocen skłonności dla zmiennych przewidywanych typu flaga.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Ocena modelu

Oblicz ważność predyktora. W przypadku modeli generujących odpowiednią miarę ważności możliwe jest wyświetlenie tabeli wskazującej ważność względną każdego predyktora w procesie estymacji modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Należy zauważyć, że obliczenie ważności predyktora może potrwać dłużej dla niektórych modeli, szczególnie w przypadku pracy z dużymi zbiorami danych, i domyślnie ta opcja dla niektórych modeli jest wyłączona. Ważność predyktorów jest niedostępna dla modeli listy decyzyjnej. Więcej informacji można znaleźć w “Ważność predyktorów” na stronie 43.

Oceny skłonności

Oceny skłonności można aktywować w węźle modelowania oraz na karcie Ustawienia w modelu użytkowym. Ta funkcja jest dostępna tylko wówczas, gdy wybrana zmienna przewidywana jest zmienną typu flaga. Więcej informacji można znaleźć w temacie “Oceny skłonności” na stronie 35.

Wylicz surowe oceny skłonności. Surowe oceny skłonności są wyznaczane z modelu wyłącznie w oparciu o dane uczące. Jeśli model przewiduje wartość *true* (udzieli odpowiedzi), wówczas skłonność jest taka sama jak P, gdzie P to prawdopodobieństwo predykcji. Jeśli model przewidzi wartość typu false, wówczas skłonność jest obliczana jako (1 – P).

- W przypadku wybrania tej opcji podczas budowania modelu oceny skłonności będą domyślnie aktywowane w modelu użytkowym. Surowe oceny skłonności można jednak aktywować w modelu użytkowym w dowolnym czasie, niezależnie od tego, czy zostały wybrane w węźle modelowania.
- Podczas oceniania modelu surowe oceny skłonności zostaną dodane do zmiennej z literami *RP* dodanymi do standardowego przedrostka. Przykładowo, jeśli predykcje znajdują się w zmiennej o nazwie *$R-churn*, wprowadzona nazwa zmiennej oceny skłonności będzie następująca: *$RRP-churn*.

Wylicz skorygowane oceny skłonności. Surowe skłonności są wyznaczane wyłącznie w oparciu o oszacowania udostępnione przez model, które mogą być nadmiernie dopasowane, co może doprowadzić do zbyt optymistycznych oszacowań skłonności. Skorygowane skłonności spróbują przeprowadzić wyrównanie, sprawdzając, jak model działa w podzbiorze testowym lub walidacyjnym i korygując skłonności, tak aby uzyskać lepsze oszacowanie.

- To ustawienie wymaga, aby w strumieniu obecna była poprawna zmienna dzieląca na podzbiory.
- W przeciwieństwie do surowych ocen ufności skorygowane oceny skłonności muszą być obliczone podczas budowania modelu; w przeciwnym razie nie będą dostępne podczas oceniania modelu użytkowego.
- Podczas oceniania modelu skorygowane oceny skłonności zostaną dodane do zmiennej z literami *AP* dodanymi do standardowego przedrostka. Przykładowo, jeśli predykcje znajdują się w zmiennej o nazwie *$R-churn*, wprowadzona nazwa zmiennej oceny skłonności będzie następująca: *$RAP-churn*. Skorygowane oceny skłonności są niedostępne dla modeli regresji logistycznej.
- Podczas obliczania skorygowanych ocen skłonności podzbiór testowy lub walidacyjny używany do obliczeń nie może być zrównoważony. Aby tego uniknąć, należy sprawdzić, czy opcja **Równoważ tylko dane uczące** jest zaznaczona w którymkolwiek poprzedzającym węźle ważenia. Ponadto, jeśli w poprzedzającej części strumienia przeprowadzona została złożona próba, spowoduje to unieważnienie skorygowanych ocen skłonności.
- Skorygowane oceny skłonności są niedostępne w przypadku modeli drzewa wzmacnianego i zestawu reguł. Więcej informacji można znaleźć w temacie “Wzmacniane modele C5.0” na stronie 124.

Na podstawie. Aby możliwe było obliczenie skorygowanych ocen skłonności, w strumieniu musi znajdować się zmienna dzieląca na podzbiory. Można określić, czy do obliczenia ma być używany podzbiór testowy czy walidacyjny. Aby uzyskać jak najlepsze wyniki, podzbiór testowy lub walidacyjny powinien zawierać co najmniej tyle rekordów, ile podzbiór użyty do uczenia oryginalnego modelu.

Węzeł C5.0

Ta funkcja jest dostępna w programach SPSS Modeler Professional i SPSS Modeler Premium.

Ten węzeł używa algorytmu C5.0 do utworzenia **drzewa decyzyjnego** albo **zestawu reguł**. Działanie modelu C5.0 polega na podziale próby na podstawie zmiennej oferującej największy **zysk informacyjny**. Każda podpróba zdefiniowana w wyniku pierwszego podziału jest ponownie dzielona, zwykle na podstawie innej zmiennej, a proces powtarzany jest do momentu, aż podprób nie da się już dalej podzielić. Po podziale podpróby na najniższym poziomie są ponownie analizowane, a te z nich, które nie przyczyniają się istotnie do budowania wartości modelu, są usuwane lub **przycinane**.

Uwaga: Węzeł C5.0 może przewidywać tylko zmienną jakościową. Podczas analizowania danych ze zmiennymi jakościowymi (nominalnymi lub porządkowymi) węzeł z większym prawdopodobieństwem będzie grupował kategorie niż węzeł C5.0 w wersjach wcześniejszych niż 11.0.

Węzeł C5.0 może generować dwa rodzaje modeli. **Drzewo decyzyjne** jest prostym opisem podziałów znalezionych przez algorytm. Każdy węzeł końcowy („liść”) opisuje konkretny podzbiór danych uczących, a każda obserwacja w danych uczących należy do dokładnie jednego węzła końcowego w drzewie. Innymi słowy dla każdego konkretnego rekordu danych odzwierciedlonego w drzewie decyzyjnym możliwa jest dokładnie jedna predykcja.

Z kolei **zestaw reguł** jest zbiorem reguł próbujących dokonać predykcji dla poszczególnych rekordów. Zestawy reguł są wywodzone z drzew decyzyjnych i mogą być traktowane jako uproszczone lub wydestylowane wersje informacji obecnych w drzewie decyzyjnym. Zestawy reguł często zachowują większość istotnych informacji z całego drzewa decyzyjnego, ale w postaci mniej złożonego modelu. Ze względu na sposób działania zestawów reguł nie mają one

tych samych właściwości, co drzewa decyzyjne. Najważniejszą różnicą dotyczącą zestawu reguł jest to, że do dowolnego rekordu może mieć zastosowanie więcej niż jedna reguła lub może nie mieć zastosowania żadna reguła. Jeśli zastosowanie ma wiele reguł, każda z nich otrzymuje ważony „głos" w oparciu o ufność powiązaną z tą regułą; ostateczna decyzja dotycząca predykcji jest podejmowana poprzez połączenie ważonych głosów ze wszystkich reguł mających zastosowanie do danego rekordu. Jeśli żadna reguła nie ma zastosowania, do rekordu przypisywana jest domyślna predykcja.

Przykład. Załóżmy, że lekarz prowadzący badania naukowe zebrał dane o zbiorze pacjentów cierpiących na tę samą chorobę. W trakcie leczenia każdy pacjent zareagował na jeden z pięciu leków. Można zastosować model C5.0 w połączeniu z innymi węzłami, aby dowiedzieć się, który lek byłby odpowiedni dla przyszłego pacjenta cierpiącego na tę samą chorobę.

Wymagania. Do uczenia modelu C5.0 potrzebna jest jedna jakościowa (tj. nominalna lub porządkowa) zmienna *Przewidywana* i co najmniej jedna zmienna *Wejściowa* dowolnego typu. Zmienne o roli *Łącznie* lub *Żadna* są ignorowane. Typy zmiennych używanych w modelu muszą być w pełni zrealizowane jako instancje. Można również określić zmienną ważącą.

Mocne strony. Modele C5.0 wykazują się dużą odpornością na takie problemy, jak braki danych i duża liczba predyktorów. Zwykle nie wymagają długiego uczenia, by generować prawidłowe oszacowania. Ponadto modele C5.0 bywają bardziej zrozumiałe niż modele innego typu, ponieważ reguły wywiedzione z modelu dają się bardzo łatwo i bezpośrednio zinterpretować. Modele C5.0 oferują także metodę **wzmacniania**, która zwiększa dokładność klasyfikacji.

Uwaga: Szybkość budowania modelu C5.0 może zostać zwiększona poprzez aktywowanie przetwarzania równoległego.

Opcje modelu węzła C5.0

Nazwa modelu. Określ nazwę modelu, który ma być generowany.

- **Auto.** Gdy ta opcja jest wybrana, nazwa modelu zostanie wygenerowana automatycznie na podstawie nazw zmiennych przewidywanych. Jest to ustawienie domyślne.
- **Użytkownika.** Wybierz tę opcję, aby określić własną nazwę modelu użytkowego tworzonego przez ten węzeł.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Utwórz modele rozdzielone. Tworzy osobny model dla każdej możliwej wartości zmiennych wejściowych, jakie są określone jako zmienne podziału. Więcej informacji można znaleźć w "Budowanie modeli rozdzielonych" na stronie 28.

Typ wyjściowy. Określ typ wynikowego modelu użytkowego: **Drzewo decyzyjne** albo **Zestaw reguł**.

Grupuj jakościowe. Gdy ta opcja jest wybrana, model C5.0 podejmuje próbę połączenia wartości symbolicznych o podobnej charakterystyce względem zmiennej przewidywanej. Gdy ta opcja nie jest wybrana, model C5.0 tworzy węzeł podrzędny dla każdej wartości zmiennej jakościowej użytej do podziału węzła nadrzędnego. Na przykład, jeśli model C5.0 dokonuje podziału wg zmiennej *COLOR* (o wartościach *RED*, *GREEN* i *BLUE*), to domyślnie tworzy podział na trzy. Jeśli jednak ta opcja jest wybrana, a rekordy, w których *COLOR* = *RED*, są bardzo podobne do rekordów, w których *COLOR* = *BLUE*, algorytm utworzy podział na dwie grupy: rekordy *GREEN* znajdują się w jednej grupie, a rekordy *BLUE* i *RED* znajdą się razem w drugiej grupie.

Użyj boostingu. Algorytm C5.0 oferuje specjalną metodę zwiększania dokładności, nazywaną wzmacnianiem lub **boostingiem**. Jej działanie polega na tworzeniu po kolei wielu modeli. Pierwszy model jest budowany w zwykły sposób. Drugi model jest budowany w taki sposób, że koncentruje się on na rekordach, które zostały błędnie sklasyfikowane przez pierwszy model. Trzeci model koncentruje się na błędach drugiego modelu, i tak dalej. Ostatecznie obserwacje są klasyfikowane przez zastosowanie do nich całego zbioru modeli, z użyciem procedury głosowania ważonego w celu połączenia poszczególnych predykcji w jedną predykcję ogólną. Boosting może istotnie

zwiększyć dokładność modelu C5.0, ale wymaga także dłuższego uczenia. Opcja **Liczba prób** pozwala określić, z ilu modeli będzie składał się model wzmocniony. Ta funkcja jest oparta na badaniach, które prowadzili Freunda i Schapire oraz własnych udoskonaleniach autorów programu związanych z obsługą danych zaszumionych.

Sprawdź krzyżowo. Gdy ta opcja jest wybrana, algorytm C5.0 używa zestawu modeli zbudowanych na podstawie podzbiorów danych uczących, aby oszacować dokładność modelu zbudowanego na podstawie całego zbioru danych. Jest to przydatne, gdy zbiór danych jest za mały, by podzielić go na tradycyjny zbiór uczący i testowy. Modele używane do walidacji krzyżowej są odrzucane po oszacowaniu dokładności. Można określić **liczbę podzbiorów** lub liczbę modeli używanych do walidacji krzyżowej. Należy zauważyć, że w poprzednich wersjach programu IBM SPSS Modeler budowanie modelu i jego walidacja krzyżowa były dwiema odrębnymi operacjami. W bieżącej wersji nie jest wymagane osobne budowanie modelu. Budowanie modelu i walidacja krzyżowa odbywa się jednocześnie.

Dominanta. W przypadku uczenia **prostego** większość parametrów algorytmu C5.0 ustawiana jest automatycznie. Szkolenie **zaawansowane** umożliwia bardziej bezpośrednie sterowanie parametrami uczenia.

Opcje trybu prostego

Preferuj. Domyślnie algorytm C5.0 próbuje utworzyć drzewo o jak największej dokładności. W niektórych przypadkach może to prowadzić do przeuczenia, co z kolei przełoży się na niską wydajność modelu po zastosowaniu go do nowych danych. Wybierz opcję **Ogólność**, aby użyć ustawień, przy których algorytm jest mniej podatny na ten problem.

Uwaga: Nie ma gwarancji, że modele utworzone z wybraną opcją **Ogólność** będą uogólniały lepiej niż inne modele. Gdy ogólność ma kluczowe znaczenie, należy zawsze zwalidować model względem zachowanej do tego celu próby testującej.

Oczekiwany szum (%). Określ oczekiwany odsetek danych zaszumionych lub błędnych w zbiorze uczącym.

Opcje trybu zaawansowanego

Istotność przycinania. Określa, w jakim stopniu zostanie przycięte drzewo decyzyjne lub zestaw reguł. Zwiększenie tej wartości spowoduje wygenerowanie mniejszego, bardziej zwartego drzewa. Zmniejszenie spowoduje wygenerowanie bardziej dokładnego drzewa. To ustawienie wpływa tylko na przycinanie lokalne (zob. opcja „Stosuj globalne przycinanie” poniżej).

Minimum rekordów w gałęzi podrzędnej. Regulując wielkość podgrup, można ograniczyć liczbę podziałów w jednej gałęzi drzewa. Gałąź drzewa zostanie podzielona tylko wtedy, gdy co najmniej dwie wynikowe podgałęzie zawierałyby co najmniej określoną tutaj liczbę rekordów ze zbioru uczącego. Wartość domyślna to 2. Zwiększenie tej wartości może zapobiec **przeuczeniu** w sytuacji, gdy dane są zaszumione.

Stosuj globalne przycinanie. Drzewa są przycinane w dwóch etapach: najpierw odbywa się etap przycinania lokalnego, w którym poddrzewa są analizowane, a gałęzie zwijane w celu zwiększenia dokładności modelu. Drugim etapem jest przycinanie globalne, w którym drzewo rozpatrywane jest jako całość i możliwe jest zwijanie słabych poddrzew. Domyślnie globalne przycinanie jest wykonywane. Aby pominąć etap przycinania globalnego, usuń zaznaczenie tej opcji.

Przetestuj przydatność atrybutów. Gdy ta opcja jest wybrana, algorytm C5.0 analizuje przydatność predyktorów przed rozpoczęciem tworzenia modelu. Predyktory uznane za nieistotne są wykluczane z procesu budowania modelu. Ta opcja bywa przydatna w przypadku modeli z wieloma predyktorami i może pomóc w unikaniu przeuczenia.

Uwaga: Szybkość budowania modelu C5.0 może zostać zwiększona poprzez aktywowanie przetwarzania równoległego.

Węzeł Drzewo-AS

Węzeł Drzewo-AS może być używany z danymi w środowisku rozproszonym. W węźle tym można wybrać budowę drzew decyzyjnych za pomocą modelu CHAID lub modelu Wyczerpujący CHAID.

CHAID (ang. Chi-squared Automatic Interaction Detection) to metoda klasyfikacji umożliwiająca budowanie drzew decyzyjnych z użyciem statystyki chi-kwadrat w celu identyfikacji optymalnych podziałów.

CHAID bada najpierw tabele krzyżowe między każdą ze zmiennych wejściowych a wynikiem oraz testuje istotność za pomocą testu niezależności chi-kwadrat. Jeśli więcej niż jedna z tych relacji jest statystycznie znacząca, CHAID wybierze najbardziej znaczącą zmienną wejściową (o najmniejszej wartości *p*). Jeśli dane wejściowe należą do dwu lub większej liczby kategorii, są one porównywane, a kategorie niewykazujące różnic w wynikach są zwijane razem. Realizuje się to przez sukcesywne łączenie par kategorii wykazujących najmniej znaczące różnice. Ten proces scalania kategorii jest zatrzymywany w chwili, gdy wszystkie pozostałe kategorie różnią się na danym poziomie testowania. W przypadku wejściowych zmiennych nominalnych można scalać dowolne kategorie; w przypadku zestawu porządkowego możliwe jest scalenie tylko kategorii zmiennych ilościowych.

Wyczerpujący CHAID stanowi modyfikację CHAID umożliwiającą dokładniejsze badanie wszystkich możliwych podziałów dla każdego predyktora, lecz obliczenia w jego przypadku zajmują więcej czasu.

Wymagania. Zmienne przewidywane i wejściowe mogą być ilościowe lub jakościowe; węzły mogą być dzielone na dwie lub więcej podgrup na każdym poziomie. Wszelkie zmienne porządkowe stosowane w modelu muszą charakteryzować się składowaniem typu numerycznego (nie łańcuchowego). W razie potrzeby do ich przekształcania należy użyć węzła Rekodowanie.

Mocne strony. CHAID może generować drzewa niebinarne, co oznacza, że niektóre podziały mają więcej niż dwie gałęzie. Oznacza to tendencję do tworzenia szerszych drzew, niż w przypadku binarnych metod wzrostu. CHAID działa w przypadku wszystkich typów danych wejściowych, i akceptuje zarówno wagi obserwacji, jak i zmienne częstości.

Opcje zmiennych węzła Drzewo-AS

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Tę opcję należy wybrać, aby ręcznie przypisać zmienne przewidywane, predyktory i inne role.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji.

Predyktory Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Waga analizy Należy tutaj wskazać zmienną, która ma być używana jako waga obserwacji. Wagi obserwacji są stosowane w celu uwzględniania różnic w wariancji między poziomami zmiennej wyjściowej. Aby uzyskać więcej informacji, zobacz “Użycie zmiennych częstości i ważących” na stronie 33.

Opcje budowania węzła Drzewo-AS

Na karcie Opcje budowania można ustawić wszystkie opcje budowania modelu. Można oczywiście po prostu kliknąć przycisk **Uruchom** w celu zbudowania modelu z wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik potrzebuje zwykle dostosować proces budowy do swoich celów.

Karta zawiera kilka różnych okien, w których można dostosować ustawienia odpowiednio do specyfiki własnego modelu.

Węzeł Drzewo-AS — informacje podstawowe

Można tutaj określić podstawowe opcje określające sposób budowania drzewa decyzyjnego.

Algorytm wzrostu drzewa Wybierz typ algorytmu **CHAID**, którego chcesz użyć. **Wyczerpujący CHAID** stanowi modyfikację CHAID umożliwiającą dokładniejsze badanie wszystkich możliwych podziałów dla każdego predyktora, lecz obliczenia w jego przypadku zajmują więcej czasu.

Maksymalna głębokość drzewa Podaj maksymalną liczbę poziomów poniżej węzła głównego (liczbę rekurencyjnych podziałów próby); wartość domyślna to 5. Maksymalna liczba poziomów (określana także jako *węzły*) to 50 000.

Kategoryzacja W przypadku użycia danych ilościowych konieczna jest kategoryzacja danych wejściowych. Można ją przeprowadzić w poprzedzającym węźle; węzeł Drzewo-AS automatycznie kategoryzuje wszelkie ilościowe dane wejściowe. W przypadku automatycznej kategoryzacji danych za pomocą węzła Drzewo-AS należy wybrać wartość opcji **Liczba kategorii** określającą, na ile kategorii zostaną rozdzielone dane wejściowe. Dane są dzielone na kategorie z równą częstością; dostępne opcje to 2, 4, 5, 10, 20, 25, 50 i 100.

Węzeł Drzewo-AS — Wzrost

Opcje wzrostu umożliwiają precyzyjne dostosowanie procesu budowy drzewa do potrzeb użytkownika.

Próg przejścia z wartości p na wielkości efektów Umożliwia określenie liczby rekordów, przy której na potrzeby budowania modelu nastąpi przełączenie z opcji **Ustawienia wartości p** na opcję **Ustawienia wielkości efektów**. Wartość domyślna to 1 000 000.

Poziom istotności dla podziału Umożliwia określenie poziomu istotności (alfa) dla podziału węzłów. Wartość musi mieścić się w przedziale od 0,01 do 0,99. Niższe wartości wykazują tendencję do tworzenia drzew o mniejszej liczbie węzłów.

Poziom istotności dla łączenia Umożliwia określenie poziomu istotności (alfa) dla łączenia węzłów. Wartość musi mieścić się w przedziale od 0,01 do 0,99. Opcja ta jest niedostępna w przypadku modeli Wyczerpujący CHAID.

Skoryguj wartości istotności metodą Bonferroniego Umożliwia korygowanie wartości istotności podczas testowania różnych kombinacji kategorii predyktorów. Wartości są korygowane w oparciu o liczbę testów, które bezpośrednio odnoszą się do liczby kategorii i poziomów pomiaru predyktora. Jest to zwykle pożądane, ponieważ zapewnia lepszą kontrolę wskaźnika wyników fałszywie dodatnich. Wyłączenie tej opcji powoduje zwiększenie możliwości analitycznych w zakresie znajdowania prawdziwych różnic, lecz odbywa się to kosztem zwiększenia wskaźnika wyników fałszywie dodatnich. Wyłączenie tej opcji może być zalecane w szczególności w przypadku niewielkich prób.

Próg wielkości efektów (tylko przewidywane zmienne ilościowe) Umożliwia ustawienie progu wielkości efektów obowiązującego podczas podziału węzłów i łączenia kategorii w przypadku używania ilościowej zmiennej przewidywanej. Wartość musi mieścić się w przedziale od 0,01 do 0,99.

Próg wielkości efektów (tylko przewidywane zmienne jakościowe) Umożliwia ustawienie progu wielkości efektów obowiązującego podczas podziału węzłów i łączenia kategorii w przypadku używania jakościowej zmiennej przewidywanej. Wartość musi mieścić się w przedziale od 0,01 do 0,99.

Zezwalaj na podział kategorii połączonych w węźle Algorytm CHAID próbuje połączyć kategorie w celu utworzenia najprostszego drzewa opisującego model. Zaznaczenie tej opcji umożliwia ponowny podział połączonych kategorii, o ile efektem jest lepsze rozwiązanie.

Poziom istotności dla grupowania liści Umożliwia określenie poziomu istotności określającego sposób tworzenia liści lub sposób identyfikacji liści nietypowych.

Chi-kwadrat dla przewidywanych zmiennych jakościowych W przypadku zmiennych jakościowych można określić metodę używaną do obliczania statystyki chi-kwadrat.

- **Pearsona** Ta metoda skraca czas obliczeń, lecz należy zachować ostrożność w przypadku stosowania jej do niewielkich prób.
- **Iloraz wiarygodności** Ta metoda jest bardziej odporna niż metoda Pearsona, lecz wydłuża czas obliczeń. Jest to metoda preferowana w przypadku niewielkich prób. Metoda ta jest używana zawsze w przypadku ilościowych zmiennych przewidywanych.

Węzeł Drzewo-AS — reguły zatrzymujące

Te opcje pozwalają sterować procesem tworzenia drzewa. Reguły zatrzymujące określają, kiedy ma nastąpić zatrzymanie podziału określonych gałęzi drzewa. Należy ustawić minimalną wielkość gałęzi, aby uniknąć podziału na bardzo małe podgrupy. Opcja **Minimum rekordów w gałęzi nadrzędnej** uniemożliwi podział, jeśli liczba rekordów w węźle do podziału (*nadrzędny*) będzie mniejsza niż określona wartość. Opcja **Minimum rekordów w gałęzi podrzędnej** uniemożliwi podział, jeśli liczba rekordów w dowolnej gałęzi utworzonej w wyniku podziału (*podrzędna*) będzie mniejsza niż określona wartość.

- **Wartość procentowa** Umożliwia określenie wielkości jako wartości procentowej wszystkich danych uczących.
- **Wartość bezwzględna** Umożliwia określenie wielkości jako bezwzględnej liczby rekordów.

Minimalna zmiana w oczekiwanych częstościach komórek Podczas szacowania częstości komórek (zarówno w przypadku modelu nominalnego, jak i modelu porządkowego efektów dla wierszy) używana jest procedura iteracyjna (epsilon) pozwalająca uzyskać zbieżność dla optymalnego oszacowania używanego w teście chi-kwadrat dla określonego podziału. Epsilon determinuje, w jakim stopniu zmiana musi wystąpić, aby iteracje mogły być kontynuowane; jeśli zmiana w porównaniu z ostatnią iteracją jest mniejsza niż zadana wartość, iteracje są zatrzymywane. W przypadku napotkania problemów związanych z brakiem zbieżności algorytmu można zwiększyć tę wartość lub zwiększyć maksymalną liczbę iteracji do wystąpienia zbieżności.

Maksimum iteracji dla uzyskania zbieżności Określa maksymalną liczbę iteracji przed zatrzymaniem, niezależnie od tego, czy uzyskano zbieżność, czy nie.

Węzeł Drzewo-AS — Koszty

W niektórych kontekstach pewne błędy są bardziej kosztowne od innych. Przykładowo, bardziej kosztowne może być sklasyfikowanie osób składających wniosek kredytowy z wysokim poziomem ryzyka jako osób z niskim poziomem ryzyka (jeden rodzaj błędu) niż sklasyfikowanie osób składających wniosek z niskim poziomem ryzyka jako osób z wysokim poziomem ryzyka (inny rodzaj błędu). Koszty błędnej klasyfikacji umożliwiają określenie względnej ważności różnych rodzajów błędów predykcji.

Kosztami błędnej klasyfikacji zwykle są wagi zastosowane do określonych danych wynikowych. Wagi te są uwzględniane w modelu i rzeczywiście mogą zmienić predykcję (jako sposób ochrony przed kosztownymi błędami).

Model, który uwzględnia koszty, może nie wygenerować mniejszej liczby błędów niż ten, który ich nie uwzględnia, i może nie mieć wyższej rangi pod względem całkowitej dokładności, ale prawdopodobnie lepiej sprawdzi się w warunkach praktycznych, ponieważ generuje kosztowne błędy.

Macierz kosztów przedstawia koszty dla każdej możliwej kombinacji przewidywanej kategorii rzeczywistej. Domyślnie wszystkie koszty błędnej klasyfikacji są ustawione na wartość 1,0. Aby wprowadzić niestandardowe wartości kosztów, należy wybrać opcję **Stosuj koszty błędnej klasyfikacji** i wprowadzić do macierzy kosztów niestandardowe wartości.

Aby zmienić koszt błędnej klasyfikacji, należy zaznaczyć komórkę odpowiadającą odpowiedniej kombinacji wartości przewidywanych i rzeczywistych, usunąć istniejącą zawartość komórki i wprowadzić do niej żądany koszt. Koszty nie są automatycznie symetryczne. Przykładowo, jeśli koszt błędnej klasyfikacji *A* jako *B* zostanie ustawiony na 2,0, koszt błędnej klasyfikacji *B* jako *A* nadal będzie miał domyślną wartość 1,0, chyba że zostanie ona również jawnie zmieniona.

W przypadku tylko porządkowych zmiennych przewidywanych można wybrać opcję **Domyślny wzrost kosztu dla porządkowej zmiennej przewidywanej** i ustawić wartości domyślne w macierzy kosztów. Opis dostępnych opcji zawiera poniższa lista.

- **Brak wzrostu** — wartość domyślna wynosząca 1,0 w przypadku każdej poprawnej predykcji.
- **Liniowa** — każda kolejna niepoprawna predykcja zwiększa koszty o 1.
- **Kwadrat** — każda kolejna niepoprawna predykcja stanowi kwadrat wartości liniowej. W takim przypadku wartości mogą wynosić: 1, 4, 9 itd.
- **Użytkownika** — w przypadku ręcznej edycji jakichkolwiek wartości w tabeli opcja na liście rozwijanej automatycznie przyjmuje wartość **Użytkownika**. W przypadku zmiany opcji z listy rozwijanej na dowolną z pozostałych edytowanych opcji wartości są zastępowane wartościami dla wybranej opcji.

Opcje modelu węzła Drzewo-AS

Na karcie Opcje modelu można zdecydować o wyborze nazwy dla modelu lub o jej wygenerowaniu automatycznie. Można także zdecydować o wyliczaniu współczynników ufności i dodawaniu identyfikatora podczas oceniania modelu.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Wylicz ufności To pole wyboru należy zaznaczyć, aby dodać zmienną ufności podczas oceny modelu.

Identyfikator reguły To pole wyboru należy zaznaczyć, aby podczas oceniania modelu dodać zmienną, która zawiera identyfikator liścia, do którego przypisano rekord.

Model użytkowy Drzewo-AS

Wynik modelu użytkowego Drzewo-AS

Po utworzeniu modelu Drzewo-AS w oknie wyników dostępne są następujące informacje.

Tabela Informacje o modelu

Tabela Informacje o modelu zawiera kluczowe informacje o modelu. Tabela określa niektóre ustawienia modelu wysokiego poziomu, takie jak:

- Typ używanego algorytmu; CHAID lub Wyczerpujący CHAID.
- Nazwa zmiennej przewidywanej wybranej w węźle Typ lub na karcie Zmienne węzła Drzewo-AS.
- Nazwy zmiennych wybranych jako predyktory w węźle Typ lub na karcie Zmienne węzła Drzewo-AS.
- Liczba rekordów w danych. W przypadku budowania modelu z wagą liczebności, ta wartość jest ważoną liczbą odzwierciedlającą rekordy, na podstawie których budowane jest drzewo.
- Liczba *liści* w wygenerowanym drzewie.
- Liczba poziomów w drzewie, innymi słowy, głębokość drzewa.

Ważność predyktorów

Wykres Ważność predyktorów przedstawia ważność pierwszych 10 danych wejściowych (predyktorów) w modelu jako wykres słupkowy.

W przypadku, gdy na wykresie jest więcej niż 10 zmiennych, można zmienić wybór predyktorów uwzględnianych na wykresie, korzystając z suwaka pod wykresem. Wskaźniki na suwaku mają stałą szerokość, a każdy znak na suwaku reprezentuje 10 zmiennych. Wskaźniki można przemieszczać wzdłuż suwaka, wyświetlając w ten sposób 10 kolejnych lub poprzednich zmiennych, uporządkowanych według ważności predyktora.

Dwukrotne kliknięcie wykresu powoduje otwarcie osobnego okna dialogowego, w którym można edytować ustawienia wykresu. Można na przykład zmodyfikować cechy, takie jak wielkość wykresu, a także rozmiar i kolor używanych czcionek. Po zamknięciu tego osobnego okna dialogowego do edycji zmiany są odzwierciedlane na wykresie wyświetlanym na karcie Wynik.

Tabela Pierwsze reguły decyzyjne

Domyślnie w tej tabeli interaktywnej wyświetlane są statystyki reguł dla pierwszych pięciu liści w wynikach, jeśli chodzi o wartość procentową sumy rekordów zawartych w liściu.

Dwukrotne kliknięcie tabeli powoduje otwarcie osobnego okna dialogowego, w którym można edytować informacje o regule prezentowane w tabeli. Wyświetlane informacje oraz opcje dostępne w oknie dialogowym zależą od tego, jakiego typu są dane zmiennej przewidywanej: jakościowe czy ilościowe.

W tabeli wyświetlane są następujące informacje o regule:

- Identyfikator reguły
- Szczegóły dotyczące sposobu stosowania reguły i jej składowych
- Liczebność rekordów dla każdej reguły. W przypadku budowania modelu z wagą liczebności, ta wartość jest ważoną liczbą odzwierciedlającą rekordy, na podstawie których budowane jest drzewo.
- Wartość procentowa rekordów dla każdej reguły

Ponadto, w przypadku ilościowej zmiennej przewidywanej, dodatkowa kolumna w tabeli przedstawia wartość **Średnia** dla każdej reguły.

Układ tabeli reguł można zmienić, korzystając z następujących opcji **Zawartość tabeli**:

- **Pierwsze reguły decyzyjne** Pierwszych pięć reguł decyzyjnych jest sortowanych według wartości procentowej sumy rekordów zawartych w liściach.
- **Wszystkie reguły** Tabela zawiera wszystkie liście utworzone przez model, lecz wyświetlanych jest tylko 20 reguł na stronę. Po wybraniu tego układu można wyszukiwać reguły, korzystając z dodatkowych opcji **Znajdź identyfikator reguły** i **Str.**.

Ponadto, w przypadku jakościowej zmiennej przewidywanej można zmienić układ tabeli reguł, korzystając z opcji **Pierwsze reguły według kategorii**. Pierwszych pięć reguł decyzyjnych jest sortowanych wg wartości procentowej sumy rekordów dla kategorii zmiennych przewidywanych wybranej w polu **Kategoria zmiennej przewidywanej**.

W przypadku zmiany układu tabeli reguł można skopiować zmodyfikowaną tabelę reguł z powrotem do przeglądarki wyników, klikając przycisk Kopiuj do okna raportów w lewym górnym rogu okna dialogowego.

Ustawienia modelu użytkowego Drzewo-AS

Karta Ustawienia modelu użytkowego Drzewo-AS umożliwia określenie opcji ufności oraz generowania kodu SQL podczas oceny modelu. Ta karta jest dostępna tylko po dodaniu modelu użytkowego do strumienia.

Wylicz ufności To pole wyboru należy zaznaczyć, aby uwzględnić ufności w operacjach oceniania. Podczas oceniania modeli w bazie danych wykluczenie ufności oznacza możliwość generowania bardziej wydajnego kodu SQL. W przypadku drzew regresji ufności nie są przypisywane.

Identyfikator reguły To pole wyboru należy zaznaczyć, aby dodać zmienną w wynikach oceniania, określającą identyfikator dla węzła końcowego, do którego przypisany jest każdy rekord.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji:

- **Domyślne: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Po podłączeniu do bazy danych z zainstalowanym składnikiem Scoring Adapter generuje kod SQL, korzystając ze składnika Scoring Adapter oraz powiązanych funkcji zdefiniowanych przez użytkownika (UDF) i ocenia model w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Węzeł Drzewa losowe

Węzeł Drzewa losowe może być używany z danymi w środowisku rozproszonym. Ten węzeł służy do budowania modelu zespolonego składającego się z wielu drzew decyzyjnych.

Węzeł Drzewa losowe jest metodą klasyfikacji i predykcji w oparciu o drzewo. W metodzie tej, podobnie jak w algorytmie C&R, stosuje się rekursywny podział rekordów uczących na segmenty o podobnych wartościach zmiennych przewidywanych. Działanie węzła rozpoczyna się od analizy dostępnych zmiennych wejściowych w poszukiwaniu najlepszych podziałów, przy czym jakość podziału mierzona jest ograniczeniem wskaźnika zanieczyszczenia uzyskanego wskutek podziału. W wyniku podziału powstają dwie podgrupy, z których każda jest następnie dzielona na następne dwie podgrupy i tak dalej, aż do spełnienia kryterium zatrzymania. Wszystkie podziały są binarne (tylko na dwie podgrupy).

Algorytm Drzewa losowe oferuje dwie funkcje niedostępne w algorytmie C&RT:

- Pierwszą jest *agregacja bootstrapowa*, w której tworzone są repliki zbiorów danych uczących poprzez próbkowanie z zastąpieniem oryginalnego zbioru danych. Taki sposób działania powoduje powstawanie prób bootstrapowych, które mają rozmiar równy oryginalnemu zbiorowi danych, po czym na każdej replice budowany jest *model zespolony*. Te modele składników tworzą razem model zespolony.
- Druga funkcja polega na tym, że przy każdym podziale drzewa tylko próbka zmiennych wejściowych jest uwzględniana na potrzeby miary zanieczyszczenia.

Wymagania. Do uczenia modelu Drzewa losowe potrzeba co najmniej jednej zmiennej *wejściowej* i jednej zmiennej *przewidywanej*. Zmienne przewidywana i wejściowa mogą być ilościowe (przedział liczbowy) lub jakościowe. Zmienne o roli *Łącznie* lub *Żadna* są ignorowane. Typy wszystkich zmiennych używane w modelu muszą być zrealizowane jako instancje zmiennych, a wszelkie zmienne porządkowe (uporządkowany zestaw) stosowane w modelu muszą być przechowywane jako liczby (nie łańcuchy). W razie potrzeby do ich przekształcenia można użyć węzła rekodowania.

Mocne strony. Modele Drzewa losowe są odporne w przypadku dużych zbiorów danych i dużej liczby zmiennych. Z powodu stosowania agregacji bootstrapowej i próbkowania zmiennych są one dużo mniej podatne na przeuczenie i dzięki temu wyniki uzyskiwane w testach mogą zostać z większym prawdopodobieństwem powtórzone w przypadku użycia nowych danych.

Opcje zmiennych węzła Drzewa losowe

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Tę opcję należy wybrać, aby ręcznie przypisać zmienne przewidywane, predyktory i inne role.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji.

Predyktory Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Waga analizy Należy tutaj wskazać zmienną, która ma być używana jako waga obserwacji. Wagi obserwacji są stosowane w celu uwzględniania różnic w wariancji między poziomami zmiennej wyjściowej. Aby uzyskać więcej informacji, zobacz “Użycie zmiennych częstości i ważących” na stronie 33.

Opcje budowania węzła Drzewa losowe

Na karcie Opcje budowania można ustawić wszystkie opcje budowania modelu. Można oczywiście po prostu kliknąć przycisk **Uruchom** w celu zbudowania modelu z wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik potrzebuje zwykle dostosować proces budowy do swoich celów.

Karta zawiera kilka różnych okien, w których można dostosować ustawienia odpowiednio do specyfiki własnego modelu.

Węzeł Drzewa losowe — ustawienia podstawowe

Można tutaj określić podstawowe opcje określające sposób budowania drzewa decyzyjnego.

Liczba modeli do zbudowania Określ maksymalną liczbę modeli, jaka może być zbudowana.

Wielkość próby Domyślnie wielkość próby bootstrapowej jest równa wielkości pierwotnych danych uczących. W przypadku pracy z dużymi zbiorami danych zmniejszenie próby może przyczynić się do wzrostu wydajności.

Obsługuj niezrównoważone dane Jeśli zmienna przewidywana modelu jest flagą (np. zakup albo brak zakupu), a liczba wyników pożądanych jest bardzo mała w stosunku do liczby wyników niepożądanych, dane są niezrównoważone, a próbkowanie bootstrapowe przeprowadzane przez model może wpłynąć na jego dokładność. Aby poprawić dokładność, zaznacz to pole wyboru; model uwzględni wówczas większy odsetek wyników pożądanych i będzie miał wyższą jakość.

Zastosuj próbkowanie ważone przy wyborze zmiennych Domyślnie zmienne dla każdego węzła-liścia są wybierane losowo z tym samym prawdopodobieństwem. Zaznacz to pole, aby zastosować ważenie zmiennych i poprawić wyniki wyboru.

Maksymalna liczba węzłów Określ maksymalną liczbę węzłów-liści dozwoloną w jednym drzewie. Jeśli następny podział spowodowałby przekroczenie tej liczby, rozrost drzewa jest zatrzymywany przed tym podziałem.

Maksymalna głębokość drzewa Określ maksymalną liczbę poziomów *węzłów-liści* poniżej węzła głównego, tj. liczbę rekurencyjnych podziałów próby.

Minimalna wielkość węzła podrzędnego Określ minimalną liczbę rekordów, które muszą być zawarte w węźle podrzędnym po podziale węzła nadrzędnego. Gdyby węzeł podrzędny miał zawierać mniej rekordów, węzeł nadrzędny nie jest dzielony.

Określ liczbę predyktorów, jaka ma być używana do podziału W przypadku budowania modeli rozdzielonych określ minimalną liczbę predyktorów, jaka ma być używana do budowy dla każdego podziału. Zapobiegnie to tworzeniu w wyniku podziału zbyt małych podgrup.

Uwaga: Liczba predyktorów do podziału nie może być większa od łącznej liczby predyktorów w danych.

Przerwij budowanie, gdy nie można będzie już poprawić dokładności W celu skrócenia czasów budowania modeli zaznacz tę opcję, co spowoduje zatrzymanie procesu budowania modelu, gdy nie ma już możliwości poprawy dokładności wyników.

Węzeł Drzewa losowe — koszty

W niektórych kontekstach pewne błędy są bardziej kosztowne od innych. Przykładowo, bardziej kosztowne może być sklasyfikowanie osób składających wniosek kredytowy z wysokim poziomem ryzyka jako osób z niskim poziomem ryzyka (jeden rodzaj błędu) niż sklasyfikowanie osób składających wniosek z niskim poziomem ryzyka jako osób z wysokim poziomem ryzyka (inny rodzaj błędu). Koszty błędnej klasyfikacji umożliwiają określenie względnej ważności różnych rodzajów błędów predykcji.

Kosztami błędnej klasyfikacji zwykle są wagi zastosowane do określonych danych wynikowych. Wagi te są uwzględniane w modelu i rzeczywiście mogą zmienić predykcję (jako sposób ochrony przed kosztownymi błędami).

Model, który uwzględnia koszty, może nie wygenerować mniejszej liczby błędów niż ten, który ich nie uwzględnia, i może nie mieć wyższej rangi pod względem całkowitej dokładności, ale prawdopodobnie lepiej sprawdzi się w warunkach praktycznych, ponieważ generuje kosztowne błędy.

Macierz kosztów przedstawia koszty dla każdej możliwej kombinacji przewidywanej kategorii rzeczywistej. Domyślnie wszystkie koszty błędnej klasyfikacji są ustawione na wartość 1,0. Aby wprowadzić niestandardowe wartości kosztów, należy wybrać opcję **Stosuj koszty błędnej klasyfikacji** i wprowadzić do macierzy kosztów niestandardowe wartości.

Aby zmienić koszt błędnej klasyfikacji, należy zaznaczyć komórkę odpowiadającą odpowiedniej kombinacji wartości przewidywanych i rzeczywistych, usunąć istniejącą zawartość komórki i wprowadzić do niej żądany koszt. Koszty nie są automatycznie symetryczne. Przykładowo, jeśli koszt błędnej klasyfikacji *A* jako *B* zostanie ustawiony na 2,0, koszt błędnej klasyfikacji *B* jako *A* nadal będzie miał domyślną wartość 1,0, chyba że zostanie ona również jawnie zmieniona.

W przypadku tylko porządkowych zmiennych przewidywanych można wybrać opcję **Domyślny wzrost kosztu dla porządkowej zmiennej przewidywanej** i ustawić wartości domyślne w macierzy kosztów. Opis dostępnych opcji zawiera poniższa lista.

- **Brak wzrostu** — wartość domyślna wynosząca 1,0 w przypadku każdej niepoprawnej predykcji.
- **Liniowa** — każda kolejna niepoprawna predykcja zwiększa koszty o 1.
- **Kwadrat** — każda kolejna niepoprawna predykcja stanowi kwadrat wartości liniowej. W takim przypadku wartości mogą wynosić: 1, 4, 9 itd.
- **Użytkownika** — w przypadku ręcznej edycji jakichkolwiek wartości w tabeli opcja na liście rozwijanej automatycznie przyjmuje wartość **Użytkownika**. W przypadku zmiany opcji z listy rozwijanej na dowolną z pozostałych edytowanych opcji wartości są zastępowane wartościami dla wybranej opcji.

Węzeł Drzewa losowe — ustawienia zaawansowane

Można tutaj wybrać zaawansowane opcje określające sposób budowania drzewa decyzyjnego.

Maksymalny procent braków danych. Określ maksymalny odsetek braków danych dozwolony w każdej zmiennej wejściowej. Jeśli odsetek przekracza tę wartość, zmienna wejściowa jest wykluczana z budowania modelu.

Wyklucz zmienne o większości pojedynczych kategorii powyżej. Określ maksymalny odsetek rekordów, jaki może zawierać jedna kategoria w zmiennej. Jeśli którakolwiek kategoria reprezentuje więcej rekordów, cała zmienna jest wykluczana z budowania modelu.

Maksymalna liczba kategorii zmiennej. Określ maksymalną liczbę kategorii, jaką może zawierać jedna zmienna. Jeśli liczba kategorii przekracza tę wartość, zmienna jest wykluczana z budowania modelu.

Minimalna zmienność zmiennej. Jeśli współczynnik zmienności zmiennej ilościowej jest mniejszy od podanej tutaj wartości, zmienna jest wykluczana z budowania modelu.

Liczba przedziałów. Podaj liczbę równych przedziałów liczebności, która ma być używana dla wejściowych zmiennych ilościowych. Dostępne są liczby: 2, 4, 5, 10, 20, 25 ,50 i 100.

Liczba interesujących reguł do ujęcia w raporcie. Określ liczbę reguł do ujęcia w raporcie (minimum 1, maksymalnie 1000, domyślnie 50).

Opcje modelu węzła Drzewa losowe

Na karcie Opcje modelu można zdecydować o wyborze nazwy dla modelu lub o jej wygenerowaniu automatycznie. Można także zdecydować o wyliczaniu ważności predyktorów podczas oceniania modelu.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Model użytkowy Drzewa losowe

Wyniki modelu użytkowego Drzewa losowe

Po utworzeniu modelu Drzewa losowe w oknie wyników dostępne są następujące informacje:

Tabela Informacje o modelu

Tabela Informacje o modelu zawiera kluczowe informacje o modelu. Tabela zawiera niektóre ustawienia modelu wysokiego poziomu, takie jak:

- Nazwa zmiennej przewidywanej wybranej w węźle Typ lub na karcie Zmienne węzła Drzewa losowe.
- Zastosowana metoda budowania modelu - Drzewa losowe.
- Liczba predyktorów wprowadzonych do modelu.

Dodatkowe szczegóły widoczne w tabeli są zależne od tego, czy budowany jest model klasyfikacji, czy model regresji, oraz od tego, czy model jest zbudowany w celu obsługi niezrównoważonych danych:

- Model klasyfikacji (ustawienia domyślne)
  - Dokładność modelu
  - Reguła błędnych klasyfikacji
- Model klasyfikacji (wybrana jest opcja **Obsługuj niezrównoważone dane**)
  - Średnia G
  - Współczynnik rzeczywiście dodatnich podzielony na klasy
- Model regresji
  - Pierwiastek błędu średniokwadratowego
  - Błąd względny
  - Wariancja wyjaśniona

Podsumowanie rekordów

Podsumowanie zawiera informację o liczbie rekordów użytych do dopasowywania modelu oraz o liczbie rekordów wykluczonych. Przedstawiona jest liczba rekordów oraz procent liczby całkowitej. Jeśli model został zbudowany w taki sposób, aby zawierał wagę liczebności, wówczas przedstawiona jest także nieważona liczba rekordów, które są uwzględnione i wykluczone.

Ważność predyktorów

Wykres Ważność predyktorów przedstawia ważność pierwszych 10 danych wejściowych (predyktorów) w modelu jako wykres słupkowy.

W przypadku, gdy na wykresie jest więcej niż 10 zmiennych, można zmienić wybór predyktorów uwzględnianych na wykresie, korzystając z suwaka pod wykresem. Wskaźniki na suwaku mają stałą szerokość, a każdy znak na suwaku reprezentuje 10 zmiennych. Wskaźniki można przemieszczać wzdłuż suwaka, wyświetlając w ten sposób 10 kolejnych lub poprzednich zmiennych, uporządkowanych według ważności predyktora.

Dwukrotne kliknięcie wykresu powoduje otwarcie osobnego okna dialogowego, w którym można edytować rozmiar wykresu. Po zamknięciu tego osobnego okna dialogowego do edycji zmiany są odzwierciedlane na wykresie wyświetlanym na karcie Wynik.

Tabela Pierwsze reguły decyzyjne

Domyślnie w tej tabeli interaktywnej wyświetlane są statystyki reguł pierwszych i są posortowane wg atrakcyjności.

Dwukrotne kliknięcie tabeli powoduje otwarcie osobnego okna dialogowego, w którym można edytować informacje o regule prezentowane w tabeli. Wyświetlane informacje oraz opcje dostępne w oknie dialogowym zależą od tego, jakiego typu są dane zmiennej przewidywanej: jakościowe czy ilościowe.

W tabeli wyświetlane są następujące informacje o regule:

- Szczegóły dotyczące sposobu stosowania reguły i jej składowych
- Jeśli wyniki znajdują się w kategorii o największej częstości
- Dokładność reguły
- Dokładność drzew
- Wskaźnik stopnia atrakcyjności

Wskaźnik stopnia atrakcyjności jest obliczany przy użyciu następującej formuły:

$$I_{index}(t) = P(A(t)) * P(B(t)) * (P(B(t)|A(t)) + P(\bar{B}(t)|\bar{A}(t)))$$

W tej formule:

- P(A(t)) jest dokładnością drzew
- P(B(t)) jest dokładnością reguły
- P(B(t)|A(t)) reprezentuje poprawne predykcje wg drzew oraz węzła
- Pozostały fragment formuły reprezentuje niepoprawne predykcje wg drzew oraz węzła.

Układ tabeli reguł można zmienić, korzystając z następujących opcji **Zawartość tabeli**:

- **Pierwsze reguły decyzyjne** Pierwszych pięć reguł decyzyjnych jest sortowanych według wskaźnika stopnia atrakcyjności.
- **Wszystkie reguły** Tabela zawiera wszystkie reguły utworzone przez model, lecz wyświetlanych jest tylko 20 reguł na stronę. Po wybraniu tego układu można wyszukiwać reguły, korzystając z dodatkowych opcji **Znajdź identyfikator reguły** i **Str.**.

Ponadto, w przypadku jakościowej zmiennej przewidywanej można zmienić układ tabeli reguł, korzystając z opcji **Pierwsze reguły według kategorii**. Pierwszych pięć reguł decyzyjnych jest sortowanych wg wartości procentowej sumy rekordów dla kategorii zmiennych przewidywanych wybranej w polu **Kategoria zmiennej przewidywanej**.

Uwaga: W przypadku zmiennych jakościowych tabela jest dostępna tylko wówczas, gdy opcja **Obsługuj niezrównoważone dane** nie jest wybrana na karcie Podstawowe w obszarze Opcje budowania.

W przypadku zmiany układu tabeli reguł można skopiować zmodyfikowaną tabelę reguł z powrotem do przeglądarki wyników, klikając przycisk Kopiuj do okna raportów w lewym górnym rogu okna dialogowego.

Macierz pomyłek

W przypadku modeli klasyfikacji macierz pomyłek przedstawia liczbę przewidywanych wyników w porównaniu z liczbą rzeczywiście obserwowanych, z uwzględnieniem odsetka prawidłowych predykcji.

Uwaga: Macierz pomyłek jest niedostępna dla modeli regresji ani w sytuacji, gdy opcja **Obsługuj niezrównoważone dane** jest wybrana na karcie Podstawowe w obszarze Opcje budowania.

Ustawienia modelu użytkowego Drzewa losowe

Karta Ustawienia modelu użytkowego Drzewa losowe umożliwia określenie opcji ufności oraz generowania kodu SQL podczas oceny modelu. Ta karta jest dostępna tylko po dodaniu modelu użytkowego do strumienia.

Wylicz ufności To pole wyboru należy zaznaczyć, aby uwzględnić ufności w operacjach oceniania. Podczas oceniania modeli w bazie danych wykluczenie ufności oznacza możliwość generowania bardziej wydajnego kodu SQL. W przypadku drzew regresji ufności nie są przypisywane.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji:

- **Domyślne: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Po podłączeniu do bazy danych z zainstalowanym składnikiem Scoring Adapter generuje kod SQL, korzystając ze składnika Scoring Adapter oraz powiązanych funkcji zdefiniowanych przez użytkownika (UDF) i ocenia model w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Modele użytkowe drzew decyzyjnych C&RT, CHAID, QUEST i C5.0

Modele użytkowe drzew decyzyjnych reprezentują struktury drzew umożliwiające predykcję określonej zmiennej wyjściowej przez jeden z węzłów modelowania drzew decyzyjnych (C&RT, CHAID, QUEST lub C5.0). Modele drzewa można generować bezpośrednio z węzła budowania drzewa lub pośrednio za pomocą interaktywnego konstruktora drzewa. Więcej informacji można znaleźć w temacie “Interaktywny konstruktor drzewa” na stronie 85.

Modele drzew oceniania

Po uruchomieniu strumienia zawierającego model użytkowy drzewa konkretny wynik zależy od typu drzewa.

- W przypadku drzew klasyfikacji (zmienna jakościowa) do danych dodawane są dwie nowe zmienne, zawierające wartość przewidywaną oraz ufność dla każdego rekordu. Predykcja bazuje na najczęstszej kategorii dla węzła końcowego, do którego przypisany jest rekord; jeśli większość respondentów w danym węźle udziela odpowiedzi *tak*, wówczas predykcja dla wszystkich rekordów w tym węźle to „tak”.
- W przypadku drzew regresji generowane są tylko wartości przewidywane; ufności nie są przypisywane.
- Opcjonalnie w przypadku modeli CHAID, QUEST i C&RT możliwe jest dodanie dodatkowej zmiennej wskazującej identyfikator węzła, do którego przypisany jest każdy rekord.

Nazwy nowych zmiennych stanowią pochodne nazwy modeli poprzedzone przedrostkiem. W przypadku modeli C&RT, CHAID i QUEST prefiksy to *$R-* dla zmiennej predykcyjnej, *$RC-* dla zmiennej ufności oraz *$RI-* dla zmiennej identyfikatora węzła. W przypadku drzew C5.0 przedrostki to *$C-* dla zmiennej predykcyjnej oraz *$CC-* dla zmiennej ufności. W przypadku wielu węzłów w modelu drzewa nazwy nowych zmiennych będą zawierały w *przedrostku* odróżniające je numery — na przykład, *$R1-*, *$RC1-*, *$R2-*.

Praca z modelami użytkowymi drzewa

Informacje dotyczące modelu można zapisywać i eksportować na szereg sposobów.

Uwaga: Wiele z tych opcji jest również dostępnych w oknie konstruktora drzewa.

Korzystając z konstruktora drzew lub modelu użytkowego drzewa, można:

- Wygenerować węzeł Filtrowanie lub Selekcja w oparciu o bieżące drzewo. Więcej informacji można znaleźć w “Generowanie węzłów filtrowania i selekcji” na stronie 94.
- Wygenerować model użytkowy Zestaw reguł reprezentujący strukturę drzewa jako zestaw reguł definiujących gałęzie końcowe drzewa. Więcej informacji można znaleźć w “Generowanie zestawu reguł z drzewa decyzyjnego” na stronie 95.
- Ponadto model można wyeksportować w formacie PMML (dot. tylko modeli użytkowych drzewa). Więcej informacji można znaleźć w “Paleta modeli” na stronie 40. Jeśli model zawiera niestandardowe podziały, taka informacja nie jest zachowywana w wyeksportowanym pliku PMML. (Podział jest zachowywany, jednak fakt, że jest on niestandardowy, a nie wybrany przez algorytm, już nie).
- Wygenerować wykres na podstawie wybranej części bieżącego drzewa. Ma to zastosowanie tylko w przypadku węzła użytkowego, jeśli jest on dołączony do innego węzła w strumieniu. Więcej informacji można znaleźć w “Tworzenie wykresów” na stronie 124.
- W przypadku wzmacnianych modeli C5.0 możliwy jest wybór opcji **Pojedyncze drzewo decyzyjne (obszar roboczy)** lub **Pojedyncze drzewo decyzyjne (paleta modeli)** pozwalający tworzyć nowy zestaw reguł na podstawie aktualnie wybranej reguły. Więcej informacji można znaleźć w temacie “Wzmacniane modele C5.0” na stronie 124.

Uwaga: Mimo, że węzeł tworzenia reguły zastąpiono węzłem C&RT, węzły drzew decyzyjnych w istniejących strumieniach, oryginalnie utworzone za pomocą węzła tworzenia reguły, będą nadal działały prawidłowo.

Modele użytkowe pojedynczego drzewa

W przypadku wybrania jako głównego celu dla węzła modelowania opcji **Zbudować pojedyncze drzewo** wynikowy model użytkowy zawiera następujące karty.

Tabela 7. Karty w modelu użytkowym pojedynczego drzewa

Tabulator	Opis	Dalsze informacje
Model	Wyświetla reguły definiujące model.	Więcej informacji można znaleźć w temacie “Reguły modelu drzewa decyzyjnego”.
Okno raportu	Wyświetla widok drzewa modelu.	Więcej informacji można znaleźć w temacie “Karta Przegląd modelu drzewa decyzyjnego” na stronie 122.
Podsumowanie	Wyświetla informacje o zmiennych, ustawieniach budowania i procesie estymacji modelu.	Więcej informacji można znaleźć w temacie “Podsumowanie modelu użytkowego/informacje” na stronie 43.
Ustawienia	Umożliwia określenie opcji ufności i generowania kodu SQL podczas oceniania modelu.	Więcej informacji można znaleźć w temacie “Drzewo decyzyjne/Ustawienia modelu użytkowego zestawu reguł” na stronie 123.
Adnotacja	Umożliwia dodawanie adnotacji opisowych, określanie nazw niestandardowych, dodawanie tekstów podpowiedzi oraz określanie słów kluczowych dla modelu.	

Reguły modelu drzewa decyzyjnego

Karta Model dla modelu użytkowego drzewa decyzyjnego przedstawia reguły definiujące model. Opcjonalnie może również zawierać wykres ważności predyktora oraz trzeci panel z informacjami na temat historii, częstości i substytutów.

Uwaga: Po wybraniu opcji **Utworzyć model dla dużych zbiorów danych** na karcie Opcje budowania w węźle CHAID (panel Cele) na karcie Model wyświetlane są tylko szczegóły dot. reguły drzewa.

Reguły drzewa

W panelu po lewej stronie wyświetlana jest lista warunków definiujących podział danych wykrytych przez algorytm na segmenty — zwykle jest to szereg reguł, jakie można zastosować w celu przypisania poszczególnych rekordów do węzłów podrzędnych w oparciu o wartości różnych predyktorów.

Drzewa decyzyjne działają w ten sposób, że rekurencyjnie dzielą wartości zmiennych wejściowych. Podzbiory danych są nazywane *gałęziami*. Gałąź początkowa (niekiedy nazywana *główną*) obejmuje wszystkie rekordy danych. Gałąź główna jest podzielona na podzbiory lub *gałęzie podrzędne* w oparciu o wartości poszczególnych zmiennych wejściowych. Każda gałąź podrzędna może być dalej podzielona na podgałęzie, które z kolei mogą dalej się dzielić. Na najniższym poziomie drzewa występują gałęzie, które nie są podzielone. Gałęzie te są nazywane *gałęziami końcowymi* (lub *liśćmi*).

Szczegóły dot. reguły drzewa

Przeglądarka reguł przedstawia wprowadzane wartości, które definiują poszczególne podzbiory lub gałęzie oraz podsumowanie wartości zmiennych wyjściowych rekordów dla tego podziału. Ogólne informacje na temat korzystania z przeglądarki modelu zawiera sekcja “Przeglądanie modeli użytkowych” na stronie 42.

W przypadku podziałów na podstawie zmiennych liczbowych gałąź jest przedstawiana w następujący sposób:

```
fieldname relation value [summary]
```

gdzie *relation* oznacza relację liczbową. Przykładowo, gałąź zdefiniowana przez wartości większe niż 100 dla zmiennej *revenue* będzie miała następującą postać:

```
revenue > 100 [summary]
```

W przypadku podziałów w oparciu o zmienne symboliczne gałąź ma następującą postać:

```
fieldname = value [summary] lub fieldname in [values] [summary]
```

gdzie *values* oznacza wartości zmiennych definiujących gałąź. Przykładowo, gałąź obejmująca rekordy, w której wartość *region* to *North*, *West* lub *South*, może wyglądać w następujący sposób:

```
region in ["North" "West" "South"] [summary]
```

W przypadku gałęzi końcowych predykcja również jest przeprowadzana poprzez dodanie strzałki i przewidywanej wartości na końcu warunku reguły. Przykładowo, liść zdefiniowany przez regułę *revenue* > 100, który przewiduje wartość *high* dla zmiennej wyjściowej, będzie wyświetlany w następujący sposób:

```
revenue > 100 [Mode: high] → high
```

Wartość *summary* dla tej gałęzi jest definiowana w różny sposób dla symbolicznych i liczbowych zmiennych wyjściowych. W przypadku drzew z liczbowymi zmiennymi wyjściowymi podsumowanie stanowi wartość *average* dla gałęzi, a *effect* dla gałęzi stanowi różnica pomiędzy średnią dla tej gałęzi a średnią jej gałęzi nadrzędnej. W przypadku drzew z symbolicznymi zmiennymi wyjściowymi podsumowanie stanowi *mode* lub najczęściej występująca wartość dla rekordów w gałęzi.

Aby w pełni opisać gałąź, konieczne jest uwzględnienie warunku, który zdefiniuje gałąź oraz warunków definiujących dalsze podziały drzewa. Na przykład w drzewie:

```
revenue > 100
  region = "North"
  region in ["South" "East" "West"]
    revenue <= 200
```

gałąź, którą reprezentuje drugi wiersz, jest zdefiniowana przez warunki: *revenue* > *100* i *region* = *"North"*.

Po kliknięciu opcji **Pokaż instancje/ufność** na pasku narzędzi każda reguła będzie również zawierała informacje na temat liczby rekordów, do których reguła ma zastosowanie (*Instancje*) oraz proporcji rekordów, dla których reguła jest prawdziwa (*Ufność*).

Ważność predyktorów

Opcjonalnie na karcie Model może być również wyświetlany wykres przedstawiający względną ważność poszczególnych predyktorów w oszacowaniu modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne.

Uwaga: Ten wykres jest dostępny tylko po wybraniu opcji **Oblicz ważność predyktora** na karcie Analiza przed wygenerowaniem modelu. Więcej informacji można znaleźć w temacie “Ważność predyktorów” na stronie 43.

Dodatkowe informacje o modelu

Po kliknięciu opcji **Pokaż panel informacji dodatkowych** na pasku narzędzi u dołu okna zostanie wyświetlony panel przedstawiający szczegółowe informacje dla wybranej reguły. Panel informacji składa się z trzech kart.

Historia. Ta karta pozwala śledzić warunki podziału od węzła głównego w dół do wybranego węzła. Przedstawiana jest lista warunków określających, kiedy rekord jest przypisywany do wybranego węzła. Do węzła przypisane zostaną te rekordy, dla których spełnione zostały wszystkie warunki.

Częstości. W przypadku modeli z symbolicznymi zmiennymi przewidywanymi na tej karcie dla każdej możliwej wartości przewidywanej przedstawiana jest liczba rekordów przypisanych do tego węzła (w danych uczących), które zawierają wartość przewidywaną. Wyświetlana jest również częstość wyrażona jako wartość procentowa (w postaci liczby z maksymalnie trzema miejscami dziesiętnymi). W modelach z liczbowymi zmiennymi przewidywanymi ta karta jest pusta.

Substytuty. O ile ma to zastosowanie, wyświetlane są wszelkie substytuty głównej zmiennej podziału dla wybranego węzła. Substytuty to zmienne alternatywne, używane w przypadku, gdy brakuje wartości głównego predyktora dla danego rekordu. Maksymalna dozwolona liczba substytutów dla danego podziału jest określona w węźle budowania drzewa, ale liczba rzeczywista zależy od danych uczących. Ogólnie, im więcej danych brakuje, tym więcej substytutów można użyć. Dla innych modeli drzewa decyzyjnego ta karta jest pusta.

Uwaga: Aby substytuty były uwzględnione w modelu, muszą zostać zidentyfikowane podczas fazy uczenia. Jeśli w próbie uczącej nie brakuje wartości, wówczas nie zostanie określony żaden substytut, a wszystkie rekordy z brakami danych napotkane podczas testowania lub oceniania zostaną automatycznie przeniesione do węzła podrzędnego z największą liczbą rekordów. Jeśli podczas testowania lub oceniania spodziewane są braki danych, należy upewnić się, że wartości tych nie ma również w próbie uczącej. Substytuty są niedostępne dla drzew CHAID.

Karta Przegląd modelu drzewa decyzyjnego

Karta Przegląd modelu użytkowego drzewa decyzyjnego przypomina ekran konstruktora drzewa. Główna różnica polega na tym, że podczas przeglądania modelu użytkowego nie można rozwinąć ani zmodyfikować drzewa. Pozostałe opcje wyświetlania i dostosowywania wyświetlania tych dwu składników są podobne. Więcej informacji można znaleźć w temacie “Dostosowywanie widoku drzewa” na stronie 87.

Uwaga: Karta Przegląd nie jest wyświetlana w przypadku modeli użytkowych CHAID w przypadku zaznaczenia opcji **Utworzyć model dla dużych zbiorów danych** na karcie Opcje budowania — Cele.

Podczas wyświetlania reguł podziału na karcie Przegląd nawiasy kwadratowe oznaczają, że sąsiednia wartość jest uwzględniona w zakresie, zaś cudzysłów oznacza, że sąsiednia wartość jest wykluczona z zakresu. Wyrażenie (23,37] oznacza więc: od 23 wyłącznie do 37 włącznie, innymi słowy: od ponad 23 do 37. Na karcie Model ten sam warunek będzie wyświetlany następująco:

```
Age > 23 i Age <= 37
```

Drzewo decyzyjne/Ustawienia modelu użytkowego zestawu reguł

Karta Ustawienia dla drzewa decyzyjnego lub dla modelu użytkowego Zestaw reguł umożliwia określenie opcji ufności oraz generowania kodu SQL podczas oceny modelu. Ta karta jest dostępna tylko po dodaniu modelu użytkowego do strumienia.

Wylicz ufności Zaznaczenie tej opcji pozwala uwzględnić ufności w operacjach oceniania. Podczas oceniania modeli w bazie danych wykluczenie ufności pozwala na generowanie bardziej wydajnego kodu SQL. W przypadku drzew regresji ufności nie są przypisywane.

Uwaga: W przypadku zaznaczenia opcji **Utworzyć model dla dużych zbiorów danych** na karcie opcji kompilacji, w panelu Metoda dla modeli CHAID to pole wyboru jest dostępne tylko w modelach użytkowych dla przewidywanych zmiennych jakościowych — nominalnych lub flag.

Wylicz surowe oceny skłonności W przypadku modeli z przewidywaną zmienną typu flaga (zwracających tak lub brak predykcji) można wyliczyć oceny skłonności definiujące prawdopodobieństwo prawdziwego wyniku określonego dla zmiennej przewidywanej. Stanowią one uzupełnienie pozostałych współczynników ufności i wartości predykcyjnych, jakie mogą zostać wygenerowane podczas oceniania.

Uwaga: W przypadku zaznaczenia opcji **Utworzyć model dla dużych zbiorów danych** na karcie opcji kompilacji, w panelu Metoda dla modeli CHAID to pole wyboru jest dostępne tylko w modelach użytkowych dla przewidywanych zmiennych jakościowych typu flaga.

Wylicz skorygowane oceny skłonności Surowe oceny skłonności bazują na danych uczących i mogą być zbyt optymistyczne z uwagi na tendencję wielu modeli do przeuczania tych danych. Opcja wyliczania skłonności skorygowanych próbuje skompensować tę tendencję na drodze oceny wydajności modelu w teście lub w podzbiorze walidacyjnym. Opcja ta wymaga, aby zmienna dzieląca na podzbiory była zdefiniowana w strumieniu oraz aby skorygowane oceny skłonności były włączone w węźle modelowania przed przystąpieniem do generowania modelu.

Uwaga: Skorygowane oceny skłonności nie są dostępne w przypadku drzewa wzmacnianego i modeli zestawu reguł. Więcej informacji można znaleźć w temacie “Wzmacniane modele C5.0” na stronie 124.

Identyfikator reguły W przypadku modeli CHAID, QUEST i C&RT ta opcja pozwala dodać zmienną w danych wynikowych oceniania, określającą identyfikator węzła końcowego, do którego przypisany jest każdy rekord.

Uwaga: Kiedy ta opcja jest wybrana, opcja generowania kodu SQL nie jest dostępna.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę, wykorzystując natywny kod SQL bez obsługi brakujących wartości** Po wybraniu tej opcji generuje natywny kod SQL umożliwiający ocenę modelu w bazie danych, bez narzutu wynikającego z obsługi braków danych. Ta opcja ustawia predykcję na wartość null ($null$) w przypadku napotkania braku danych podczas oceniania obserwacji.

  Uwaga: Opcja ta jest niedostępna w przypadku modeli CHAID. W przypadku pozostałych typów modeli jest ona dostępna tylko dla drzew decyzyjnych (nie zaś dla zestawów reguł).
- **Przeprowadź ocenę, wykorzystując natywny kod SQL z obsługą brakujących wartości** W przypadku modeli CHAID, QUEST i C&RT możliwe jest wygenerowanie natywnego kodu SQL w celu oceny modelu w bazie danych

z pełną obsługą braków danych. Oznacza to generowanie kodu SQL w taki sposób, aby braki danych były obsługiwane zgodnie ze specyfikacją w modelu. Na przykład Drzewa K&R korzystają z reguł substytucyjnych i największego wycofania podrzędnego.

Uwaga: W przypadku modeli C5.0 ta opcja jest dostępna tylko dla zestawów reguł (nie zaś dla drzew decyzyjnych).

- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Wzmacniane modele C5.0

Ta funkcja jest dostępna w programach SPSS Modeler Professional i SPSS Modeler Premium.

Tworzenie wzmocnionego modelu C5.0 (albo zestawu reguł, albo drzewa decyzyjnego) polega faktycznie na utworzeniu zestawu powiązanych modeli. Przeglądarka reguł wzmocnionego modelu C5.0 wyświetla listę modeli na najwyższym poziomie hierarchii wraz z oszacowaną dokładnością każdego modelu i ogólną dokładnością modelu zespolonego. Aby zapoznać się z regułami lub podziałami dla konkretnego modelu, wybierz ten model i rozwiń go, tak jak gdyby był regułą lub gałęzią jednego modelu.

Można także wyodrębnić konkretny model ze zbioru modeli wzmocnionych i utworzyć nowy model użytkowy Zestaw reguł zawierający tylko ten wyodrębniony model. Aby utworzyć nowy zestaw reguł ze wzmocnionego modelu C5.0, wybierz interesujący zestaw reguł lub drzewo i z menu Utwórz wybierz albo opcję **Pojedyncze drzewo decyzyjne (paleta modeli)**, albo **Pojedyncze drzewo decyzyjne (obszar roboczy)**.

Tworzenie wykresów

Węzły drzew dostarczają bardzo wielu informacji, które jednak nie zawsze podane są w formie łatwo dostępnej dla użytkowników biznesowych. Aby przedstawić dane w postaci odpowiedniej do uwzględnienia w raportach biznesowych, prezentacjach itp., można tworzyć wykresy na podstawie wybranych danych. Na przykład z karty Model lub Przegląd modelu użytkowego lub z karty Przegląd drzewa interaktywnego można utworzyć wykres na podstawie wybranej części drzewa, który obejmował będzie tylko obserwacje z wybranej części, gałęzi lub węzła.

Uwaga: Wykres z modelu użytkowego można wygenerować tylko wtedy, gdy model użytkowy jest połączony z innymi węzłami w strumieniu.

Tworzenie wykresu

Pierwszym krokiem jest wybranie informacji, które mają być widoczne na wykresie:

- Na karcie Model modelu użytkowego rozwiń listę warunków i reguł w lewym panelu i wybierz element, który chcesz przedstawić na wykresie.
- Na karcie Przegląd modelu użytkowego rozwiń listę gałęzi i wybierz węzeł, który chcesz przedstawić na wykresie.
- Na karcie Przegląd drzewa interaktywnego rozwiń listę gałęzi i wybierz węzeł, który chcesz przedstawić na wykresie.

  Uwaga: Na karcie Przegląd nie można wybrać najwyższego węzła.

Sposób tworzenia wykresu jest zawsze taki sam, niezależnie od sposobu wyboru danych:

1. Z menu Utwórz wybierz opcję **Wykres (z wyboru)**; zamiast tego na karcie Przegląd kliknij przycisk **Wykres (z wyboru)** w lewym dolnym rogu. Zostanie wyświetlona karta podstawowej wizualizacji.

   Uwaga: Po wyświetleniu wizualizacji opisanym sposobem dostępna będzie tylko karta Podstawowe i Zaawansowane.
2. Korzystając z ustawień na karcie podstawowej lub zaawansowanej wizualizacji, określ informacje, które mają być widoczne na wykresie.
3. Kliknij przycisk OK, aby utworzyć wykres.

W nagłówku wykresu podane są informacje o węzłach lub regułach, które zostały wybrane do uwzględnienia.

Modele użytkowe dla boostingu, agregacji bootstrapowej i bardzo dużych zbiorów danych

W przypadku wyboru opcji **Zwiększyć dokładność modelu (boosting)**, **Wzmocnić stabilność modelu (agregacja bootstrapowa)** lub **Utworzyć model dla dużych zbiorów danych** jako głównego celu modelowania program IBM SPSS Modeler tworzy zespół składający się z wielu modeli. Więcej informacji można znaleźć w temacie “Modele dla zestawów” na stronie 45.

Wynikowy model użytkowy zawiera następujące karty. Karta Model zawiera pewną liczbę różnych widoków modelu.

Tabela 8. Karty dostępne w modelu użytkowym

Tabulator	Widok	Opis	Dalsze informacje
Model	Podsumowanie modelu	Wyświetla podsumowanie jakości zespołu i (z wyjątkiem modeli wzmacnianych i ilościowych zmiennych przewidywanych) różnorodności, będącej miarą stopnia odróżniania się predykcji między poszczególnymi modelami.	Więcej informacji można znaleźć w temacie “Podsumowanie modelu” na stronie 45.
	Ważność predyktorów	Wyświetla wykres określający względną ważność każdego predyktora (zmiennej wejściowej) w oszacowaniu modelu.	Więcej informacji można znaleźć w temacie “Ważność predyktorów” na stronie 45.
	Częstotliwość predyktorów	Wyświetla wykres przedstawiający względną częstość, z jaką każdy predyktor jest używany w zestawie modeli.	Więcej informacji można znaleźć w temacie “Częstotliwość predyktorów” na stronie 46.
	Dokładność modeli składowych	Kreśli wykres dokładności predykcyjnej każdego z różnych modeli w zespole.	
	Szczegóły dotyczące modeli składowych	Wyświetla informacje na temat każdego z różnych modeli w zespole.	Więcej informacji można znaleźć w temacie “Szczegóły dotyczące modeli składowych” na stronie 46.
	Informacje	Wyświetla informacje o zmiennych, ustawieniach budowania i procesie estymacji modelu.	Więcej informacji można znaleźć w temacie “Podsumowanie modelu użytkowego/informacje” na stronie 43.
Ustawienia		Umożliwia uwzględnienie ufności w operacjach oceniania.	Więcej informacji można znaleźć w temacie “Drzewo decyzyjne/Ustawienia modelu użytkowego zestawu reguł” na stronie 123.
Adnotacja		Umożliwia dodawanie adnotacji opisowych, określanie nazw niestandardowych, dodawanie tekstów podpowiedzi oraz określanie słów kluczowych dla modelu.	

Modele użytkowe zestawu reguł C&RT, CHAID, QUEST, C5.0 i Apriori

Model użytkowy Zestaw reguł odzwierciedla reguły predykcji konkretnej zmiennej wynikowej wykrytej przez węzeł modelowania reguł asocjacyjnych (Apriori) lub jeden z węzłów budujących drzewo (C&RT, CHAID, QUEST lub C5.0). W przypadku reguł asocjacyjnych zestaw reguł musi być utworzony z surowego modelu użytkowego reguły. W przypadku drzew zestaw reguł można utworzyć z interaktywnego konstruktora drzewa, z węzła budującego model C5.0

lub z dowolnego modelu użytkowego drzewa. W odróżnieniu od surowych modeli użytkowych Reguła, modele użytkowe Zestaw reguł można umieszczać w strumieniach w celu generowania predykcji.

Po uruchomieniu strumienia zawierającego model użytkowy Zestaw reguł do strumienia tego dodawane są dwie nowe zmienne zawierające wartość przewidywaną i ufność dla każdego rekordu w danych. Nazwy nowych zmiennych stanowią pochodne nazwy modeli poprzedzone przedrostkiem. W przypadku zestawów reguł asocjacyjnych zmienna predykcji ma przedrostek *$A-*, a zmienna ufności ma przedrostek *$AC-*. W przypadku zestawów reguł C5.0 przedrostki to *$C-* dla zmiennej predykcji oraz *$CC-* dla zmiennej ufności. W przypadku zestawów reguł C&RT zmienna predykcji ma przedrostek *$R-*, a zmienna ufności ma przedrostek *$RC-*. W strumieniu z wieloma modelami użytkowymi Zestaw reguł asocjacyjnych umieszczonymi szeregowo i przewidującymi te same zmienne wynikowe nowe nazwy zmiennych będą opatrzone *przedrostkami liczbowymi* odróżniającym je od siebie. W pierwszym modelu użytkowym Zestaw reguł asocjacyjnych w strumieniu będą używane zwykłe nazwy, w drugim nazwy rozpoczynające się od *$A1-* i *$AC1-*, zaś w trzecim nazwy rozpoczynające się od *$A2-* i *$AC2-* itd.

Sposób stosowania reguł. Zestawy reguł generowane na podstawie reguł asocjacyjnych różnią się od innych modeli użytkowych, ponieważ dla danego rekordu może być wygenerowana więcej niż jedna predykcja, a predykcje te mogą nie być ze sobą zgodne. Istnieją dwie metody generowania predykcji z zestawów reguł.

Uwaga: Zestawy reguł generowane z drzew decyzyjnych zwracają te same wyniki niezależnie od zastosowanej metody, ponieważ reguły wywiedzione z drzewa decyzyjnego wykluczają się wzajemnie.

- **Głosowanie.** W tej metodzie podejmowana jest próba łączenia predykcji wszystkich reguł mających zastosowanie do rekordu. Dla każdego rekordu analizowane są wszystkie reguły i każda reguła mająca do niego zastosowanie jest używana do wygenerowania predykcji i powiązanej z nią ufności. Obliczana jest suma ufności dla każdej wartości wynikowej, a wartość o największej sumie ufności jest wybierana jako ostateczna predykcja. Ufność ostatecznej predykcji jest sumą ufności tej wartości podzieloną przez liczbę reguł zastosowanych do tego rekordu.
- **Pierwsze trafienie.** W tej metodzie reguły są po prostu testowane po kolei, a do wygenerowania predykcji używana jest pierwsza reguła mająca zastosowanie do rekordu.

Wyboru metody można dokonać w opcjach strumienia.

Generowanie węzłów. Menu Utwórz umożliwia tworzenie nowych węzłów na podstawie zestawu reguł.

- **Węzeł filtrowania** Tworzy nowy węzeł filtrowania w celu odfiltrowania zmiennych nieużywanych przez reguły w zestawie reguł.
- **Węzeł selekcji** Tworzy nowy węzeł selekcji do wyboru rekordów, do których ma zastosowanie wybrana reguła. Utworzony węzeł będzie wybierał rekordy, dla których wszystkie poprzedniki reguły są prawdziwe. Ta opcja wymaga wybrania reguły.
- **Węzeł śledzenia reguł** Tworzy superwęzeł, który wyznaczy wartość zmiennej wskazującą na regułę, której użyto do dokonania predykcji dla rekordu. Gdy ewaluacja zestawu reguł przeprowadzana jest metodą pierwszego trafienia, jest to po prostu symbol wskazujący na pierwszą regułę, która miała zastosowanie. Gdy ewaluacja zestawu reguł przeprowadzana jest metodą głosowania, jest to bardziej złożony łańcuch przedstawiający dane wejściowe dla mechanizmu głosowania.
- **Pojedyncze drzewo decyzyjne** (obszar roboczy) / **Pojedyncze drzewo decyzyjne** (paleta modeli). Tworzy pojedynczy nowy model użytkowy Zestaw reguł wywiedziony z obecnie wybranej reguły. Opcja dostępna tylko w przypadku **wzmacnianych** modeli C5.0. Więcej informacji można znaleźć w temacie “Wzmacniane modele C5.0” na stronie 124.
- **Model do palety** Zwraca model do palety modeli. Opcja ta jest przydatna w sytuacjach, gdy kolega/koleżanka wysłał(a) użytkownikowi strumień zawierający model, a nie sam model.

Uwaga: Karty Ustawienia i Podsumowanie w modelu użytkowym Zestaw reguł są identyczne, jak w modelach drzewa decyzyjnego.

Karta Model zestawu reguł

Karta Model modelu użytkowego Zestaw reguł zawiera listę reguł wyodrębnionych z danych przez algorytm.

Reguły są rozbite według następnika (przewidywana kategoria) i zaprezentowane w następującym formacie:

```
if antecedent_1
and antecedent_2
...
and antecedent_n
then predicted value
```

gdzie consequent i antecedent_1 do antecedent_n są warunkami. Reguła jest interpretowana następująco „dla rekordów, w których antecedent_1 do antecedent_n wszystkie są prawdziwe, all true, consequent także jest prawdopodobnie prawdziwy". Po kliknięciu opcji **Pokaż instancje/ufność** na pasku narzędzi każda reguła będzie również zawierała informacje na temat liczby rekordów, do których reguła ma zastosowanie, tj. dla których poprzedniki są prawdziwe (**Instancje**) oraz odsetek rekordów, dla których cała reguła jest prawdziwa (**Ufność**).

Należy zwrócić uwagę, że w przypadku zestawów reguł C5.0 ufność obliczana jest w specyficzny sposób. W regułach C5.0 ufność reguły jest obliczana według wzoru:

```
(1 + liczba rekordów, w których reguła jest poprawna)
/
(2 + liczba rekordów, dla których poprzedniki reguły są prawdziwe)
```

W takim sposobie szacowania ufności uwzględniany jest proces generalizacji reguł z drzewa decyzyjnego (model C5.0 realizuje ten proces, tworząc zestaw reguł).

Importowanie projektów z programu AnswerTree 3.0

IBM SPSS Modeler może importować projekty zapisane w programie AnswerTree 3.0 lub 3.1 za pośrednictwem standardowego okna dialogowego Plik > Otwórz. Procedura postępowania jest następująca:

1. Z menu programu IBM SPSS Modeler wybierz:

   Plik > Otwórz strumień
2. Z menu rozwijanego Pliki typu wybierz pozycję **Pliki projektu AT (*.atp, *.ats)**.

   Każdy zaimportowany projekt jest przekształcany w strumień programu IBM SPSS Modeler zawierający następujące węzły:
   - Jeden węzeł źródłowy, który definiuje używane źródło danych (na przykład pliki danych IBM SPSS Statistics lub źródło bazodanowe).
   - Dla każdego drzewa w projekcie (może ich być kilka) tworzony jest jeden węzeł wprowadzania danych, który definiuje właściwości poszczególnych zmiennych, w tym typ, rolę (dane wejściowe lub predyktor albo wyniki lub zmienna przewidywana), braki danych i inne opcje.
   - Dla każdego drzewa w projekcie tworzony jest węzeł podziału na podzbiory, który dzieli dane dla próby uczącej lub testowej, oraz węzeł budujący drzewo definiujący parametry generowania drzewa (węzeł C&RT, QUEST albo CHAID).
3. Aby wyświetlić wygenerowane drzewa, uruchom strumień.

Komentarze

- Drzew decyzyjnych utworzonych w programie IBM SPSS Modeler nie można eksportować do programu AnswerTree; import z programu AnswerTree do programu IBM SPSS Modeler jest operację jednokierunkową.
- Zyski zdefiniowane w programie AnswerTree nie są zachowywane podczas importowania projektu do programu IBM SPSS Modeler.

Rozdział 7. Modele sieci bayesowskiej

Węzeł sieci bayesowskiej

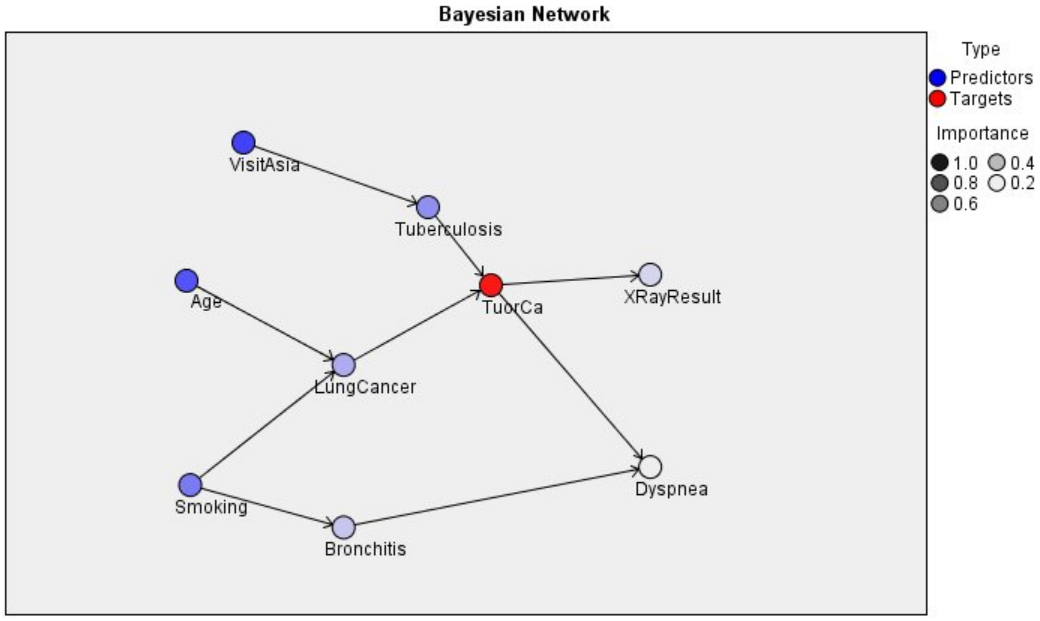
Węzeł **sieci bayesowskiej** umożliwia utworzenie modelu prawdopodobieństwa przez połączenie zaobserwowanych i zarejestrowanych dowodów ze „zdroworozsądkową" wiedzą rzeczywistą w celu ustanowienia prawdopodobieństwa występowania zdarzeń na podstawie pozornie niepowiązanych ze sobą atrybutów. Węzeł koncentruje się na sieciach Tree Augmented Naïve Bayes (TAN) i Markov Blanket, używanych głównie do klasyfikacji.

Sieci bayesowskie stosowane są do predykcji w wielu różnych sytuacjach; oto niektóre przykłady:

- Wybór wniosków kredytowych o niskim ryzyku zaległości w spłatach.
- Szacowanie, kiedy sprzęt będzie wymagał serwisowania, nowych części lub wymiany na nowy na podstawie danych z czujników i istniejącej dokumentacji.
- Rozwiązywanie problemów klientów za pośrednictwem narzędzi internetowych.
- Diagnozowanie działania sieci telefonii komórkowej i rozwiązywanie wykrytych problemów w czasie rzeczywistym.
- Ocena potencjalnego ryzyka i potencjalnych zysków z projektów badawczo-rozwojowych w celu skierowania zasobów do najlepiej rokujących inicjatyw.

Sieć bayesowska jest modelem graficznym prezentującym zmienne (często nazywanych **węzłami**) w zbiorze danych oraz prawdopodobnych lub warunkowych współzależności między tymi zmiennymi. Sieć bayesowska może odzwierciedlać relacje przyczynowe między węzłami; jednak łącza w sieci (nazywane także **łukami**) nie zawsze odzwierciedlają bezpośrednią przyczynę i skutek. Sieci bayesowskiej można na przykład użyć do obliczenia prawdopodobieństwa, że pacjent cierpi na określoną chorobę, na podstawie obecności lub braku określonych objawów i innych istotnych danych, jeśli prawdopodobne współzależności między objawami a chorobą uwidocznione na wykresie są prawdziwe. Sieci są bardzo odporne na braki danych i generują najlepsze predykcje możliwe do uzyskania na podstawie dostępnych informacji.

Typowy podstawowy przykład sieci bayesowskiej opracowali Lauritzen i Spiegelhalter (1988). Często nazywa się go modelem „Asia". Stanowi on uproszczoną wersję sieci, którą można stosować do diagnozowania nowych pacjentów lekarza; kierunek łączy w przybliżeniu odpowiada relacjom przyczynowym. Każdy węzeł odzwierciedla jeden aspekt, który może mieć związek ze stanem pacjenta. „Smoking" oznacza, że pacjent pali tytoń, a „VisitAsia" oznacza, że niedawno był w Azji. Relacje prawdopodobieństw są uwidocznione przez łącza między węzłami; na przykład palenie zwiększa prawdopodobieństwo zarówno zachorowania na zapalenie oskrzeli, jak i na raka płuc, natomiast wiek wydaje się być skorelowany tylko z ryzykiem wystąpienia raka płuc. W ten sam sposób nieprawidłowości na zdjęciu RTG płuc mogą być wywołane albo gruźlicą, albo rakiem płuc, natomiast ryzyko występowania duszności jest większe, jeśli pacjent jednocześnie choruje na zapalenie oskrzeli albo raka płuc.



Rysunek 29. Przykład sieci Asia (autorzy Lauritzen i Spiegelhalter)

Istnieje kilka uzasadnień dla zastosowania sieci bayesowskiej:

- Sieć taka pomaga w ujawnieniu relacji przyczynowych. Dzięki temu pomaga określić obszar, którego dotyczy problem, i przewidywać konsekwencje podejmowanych interwencji.
- Zastosowanie sieci jest skuteczną strategią unikania przeuczenia.
- Relacje są uwidocznione w przejrzystej postaci.

Wymagania. Zmienne przewidywane muszą być jakościowe i mogą mieć poziom pomiaru *Nominalne*, *Porządkowa* lub *Flaga*. Zmienne wejściowe mogą być dowolnego typu. Zmienne ciągłe (przedziały liczbowe) będą automatycznie kategoryzowane; jeśli jednak rozkład jest skośny, lepsze wyniki można uzyskać poprzez ręczną kategoryzację zmiennych za pomocą węzła kategoryzacji umieszczonego przed węzłem sieci bayesowskiej. Na przykład można zastosować kategoryzację optymalną, w której **zmienna nadzorcy** będzie taka sama, jak zmienna **przewidywana** sieci bayesowskiej.

Przykład. Analityk w banku chce przewidzieć, którzy klienci lub potencjalni klienci prawdopodobnie będą zalegać ze spłatą długów. Można użyć sieci bayesowskiej do określenia cech klientów o największym prawdopodobieństwie zalegania ze spłatami i zbudować kilka różnych typów modeli, aby wybrać ten, który będzie najlepiej przewidywał potencjalnie problematycznych kredytobiorców.

Przykład. Operator telekomunikacyjny chce ograniczyć liczbę klientów, którzy odchodzą z jego sieci, i co miesiąc aktualizuje model danymi z poprzedniego miesiąca. W takim scenariuszu można użyć modelu sieci bayesowskiej do określenia cech klientów o największym prawdopodobieństwie odejścia, i co miesiąc uczyć model na podstawie nowych danych.

Opcje modelu węzła sieci bayesowskiej

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Zbuduj model dla każdego rozdziału. Tworzy osobny model dla każdej możliwej wartości zmiennych wejściowych, jakie są określone jako zmienne podziału. Więcej informacji można znaleźć w temacie “Budowanie modeli rozdzielonych” na stronie 28.

Podział. To pole umożliwia określenie zmiennej używanej do podziału danych na osobne próby do uczenia, testowania i walidacji podczas budowania modelu. Korzystając z jednej próby do generowania modelu oraz innej do testowania go, można uzyskać wiarygodne informacje o tym, jak dobrze model pozwala uogólnić większe zbiory danych, podobne do danych bieżących. Jeśli korzystając z węzłów Typ lub Partycja, zdefiniowano wiele zmiennych dzielących na podzbiory, na karcie Zmienne każdego węzła modelowania korzystającego z tego podziału na podzbiory należy wybrać jedną zmienną dzielącą na podzbiory. (Jeśli obecna jest tylko jedna zmienna dzieląca na podzbiory, jest ona automatycznie używana po aktywowaniu dzielenia). Należy również pamiętać, że aby zastosować wybrany podział w analizie, dzielenie musi być również włączone na karcie Opcje modelu danego węzła. (Usunięcie zaznaczenia tej opcji umożliwia wyłączenie podziału bez zmiany ustawień zmiennych).

Rozdzielone. W przypadku modeli rozdzielonych należy wybrać zmienne lub zmienną podziału. Działanie jest podobne, jak w przypadku ustawienia roli zmiennej na wartość *Rozdzielone* w węźle Typ. Na zmienne podziału można wyznaczyć tylko zmienne z poziomem pomiaru **Flaga**, **Nominalna**, **Porządkowa** lub **Ilościowa**. Zmienne wybrane jako zmienne podziału nie mogą być używane jako zmienne przewidywane, wejściowe, zmienne dzielące na podzbiory, zmienne częstości ani zmienne ważące. Więcej informacji można znaleźć w temacie “Budowanie modeli rozdzielonych” na stronie 28.

Kontynuuj uczenie istniejącego modelu. W wypadku wybrania tej opcji wyniki widoczne na karcie Model modelu użytkowego będą generowane od nowa i aktualizowane po każdym uruchomieniu modelu. Jest to celowe na przykład po dodaniu nowego lub zaktualizowanego źródła danych do istniejącego modelu.

Uwaga: Ta operacja może tylko zaktualizować istniejącą sieć; nie można dodać ani usunąć węzłów lub połączeń. Po każdym ponownym uczeniu modelu sieć zachowa ten sam kształt. Zmienią się jedynie prawdopodobieństwa warunkowe i ważności predyktorów. Jeśli nowe dane są ogólnie podobne do starych, to powyższe ograniczenie nie ma znaczenia, ponieważ spodziewamy się, że ważne będą te same czynniki. Jeśli jednak chcemy sprawdzić lub na nowo ustalić, *co* jest istotne (a nie jak istotne), to musimy zbudować nowy model, tj. stworzyć nową sieć.

Typ struktury. Wybierz typ struktury, która ma być używana przy budowaniu sieci bayesowskiej:

- **TAN.** Model Tree Augmented Naïve Bayes (TAN) tworzy prosty model sieci bayesowskiej będący udoskonaloną wersją standardowego modelu Naïve Bayes. Udoskonalenie polega na tym, że każdy predyktor może zależeć od innego predyktora, a nie tylko od zmiennej przewidywanej, zatem dokładność klasyfikacji jest potencjalnie większa.
- **Markov Blanket.** Ten model wybiera w zbiorze danych zestaw węzłów, które zawierają zmienne nadrzędne zmiennej przewidywanej, jej zmienne podrzędne oraz zmienne nadrzędne jej zmiennych podrzędnych. Zasadniczo model Markov Blanket wykrywa w sieci wszystkie zmienne potrzebne do przewidzenia zmiennej przewidywanej. Tę metodę budowania sieci uważa się za bardziej dokładną; jednak w przypadku dużych zbiorów danych czas przetwarzania może wydłużyć się ze względu na konieczność analizowania dużej liczby zmiennych. Aby skrócić czas przetwarzania można skorzystać z opcji **Dobór predyktorów** na karcie Zaawansowane i wybrać zmienne istotnie powiązane ze zmienną przewidywaną.

Uwzględnij krok wstępnego przetwarzania przy wyborze predyktorów. Zaznaczenie tej opcji umożliwia skorzystanie z opcji **Dobór predyktorów** na karcie Zaawansowane.

Metoda uczenia parametrów Parametry sieci bayesowskiej są to warunkowe prawdopodobieństwa dla każdego węzła wyznaczone na podstawie wartości węzłów nadrzędnych. Dostępne są dwie opcje sterujące szacowaniem tabel prawdopodobieństwa warunkowego między węzłami, gdy znane są wartości węzłów nadrzędnych:

- **Największej wiarygodności.** Zaznacz to pole, gdy zbiór danych jest obszerny. Jest to ustawienie domyślne.
- **Korekta Bayesa dla niewielkiej liczby komórek.** W przypadku mniejszych zbiorów danych istnieje ryzyko przeuczenia modelu oraz możliwość wystąpienia dużej liczby liczebności zerowych. Wybranie tej opcji umożliwia złagodzenie tych problemów poprzez zastosowanie wygładzania w celu ograniczenia wpływu liczebności zerowych i niewiarygodnych oszacowań.

Opcje zaawansowane węzła sieci bayesowskiej

Opcje ustawień zaawansowanych węzła umożliwiają precyzyjne dostosowanie procesu tworzenia modelu. W celu uzyskania dostępu do opcji zaawansowanych należy ustawić opcję Tryb na wartość **Zaawansowany** na karcie Zaawansowany.

Braki danych. Domyślnie program IBM SPSS Modeler korzysta tylko z rekordów zawierających poprawne wartości dla wszystkich zmiennych użytych w modelu. (Niekiedy jest to zwane **usuwaniem obserwacjami** brakujących wartości). Jeśli istnieje dużo braków danych, może okazać się, że to rozwiązanie eliminuje zbyt wiele rekordów, a ilość pozostałych danych jest zbyt mała, aby wygenerować dobry model. W takich przypadkach można usunąć zaznaczenie opcji **Używaj tylko kompletnych rekordów**. Wówczas program IBM SPSS Modeler podejmie próbę użycia jak największej ilości informacji do oszacowania modelu, łącznie z rekordami, w których dla niektórych zmiennych istnieją braki danych. (Niekiedy jest to zwane *usuwaniem parami* brakujących wartości). Jednak w niektórych sytuacjach użycie niekompletnych rekordów w taki sposób może prowadzić do problemów obliczeniowych podczas szacowania modelu.

Dołącz wszystkie prawdopodobieństwa. Określa, czy prawdopodobieństwa dla poszczególnych kategorii zmiennych wyjściowych są dodawane do poszczególnych rekordów przetwarzanych przez węzeł. Jeśli ta opcja nie zostanie wybrana, wówczas zostanie dodane tylko prawdopodobieństwo przewidywanej kategorii.

Test niezależności. Test niezależności ocenia, czy połączone w parę obserwacje dwóch zmiennych są od siebie niezależne. Wybierz typ testu, który ma być zastosowany. Dostępne są następujące opcje:

- **Iloraz wiarygodności.** Testuje niezależność zmienna przewidywana-predyktor, obliczając stosunek maksymalnego prawdopodobieństwa wyniku przy dwóch różnych hipotezach.
- **Chi-kwadrat Pearsona.** Testuje niezależność zmienna przewidywana-predyktor, używając hipotezy zerowej mówiącej, że względne częstości występowania zaobserwowanych zdarzeń mają określony rozkład.

Modele sieci bayesowskiej warunkowo testują niezależność z wykorzystaniem nie tylko pary testowanej, lecz również innych zmiennych. Ponadto modele eksplorują nie tylko relacje między zmiennymi przewidywanymi a predyktorami, lecz również relacje między samymi predyktorami.

Uwaga: Opcje testowania niezależności są dostępne tylko, jeśli na karcie Model modelu Markov Blanket wybrano albo opcję **Uwzględnij krok wstępnego przetwarzania przy wyborze predyktorów**, albo **Typ struktury**.

Poziom istotności. W połączeniu z ustawieniami testowania niezależności umożliwia określenie wartości odcięcia obowiązującej przy wykonywaniu testów. Im mniejsza wartość, tym mniej łączy pozostanie w sieci; domyślny poziom to 0,01.

Uwaga: Ta opcja jest dostępna tylko, jeśli na karcie Model modelu Markov Blanket wybrano albo opcję **Uwzględnij krok wstępnego przetwarzania przy wyborze predyktorów**, albo **Typ struktury**.

Maksymalny rozmiar zbioru warunkującego. W algorytmie tworzenia struktury Markov Blanket do przeprowadzania testów niezależności i usuwania zbędnych łączy z sieci stosowane są coraz większe zbiory warunkujące. Ponieważ testy z udziałem dużej liczby zmiennych warunkujących wymagają więcej czasu i pamięci, można ograniczyć liczbę uwzględnianych zmiennych. Bywa to szczególnie przydatne w przypadku przetwarzania danych z silnymi zależnościami między wieloma zmiennymi. Należy jednak zwrócić uwagę, że wynikowa sieć może wówczas zawierać zbędne łącza.

Określ maksymalną liczbę zmiennych warunkujących, która może być używana do testowania niezależności. Wartość domyślna to 5.

Uwaga: Ta opcja jest dostępna tylko, jeśli na karcie Model modelu Markov Blanket wybrano albo opcję **Uwzględnij krok wstępnego przetwarzania przy wyborze predyktorów**, albo **Typ struktury**.

Dobór predyktorów. Ta opcja umożliwia ograniczenie liczby zmiennych wejściowych używanych przy przetwarzaniu w celu przyspieszenia budowania modelu. Jest to szczególnie użyteczne przy tworzeniu struktury Markov Blanket ze

względu na możliwą dużą liczbę potencjalnych zmiennych wejściowych; opcja ta umożliwia wybranie zmiennych wejściowych istotnie powiązanych ze zmienną przewidywaną.

Uwaga: Opcje doboru predyktorów są dostępne tylko, jeśli na karcie Model wybrano opcję **Uwzględnij krok wstępnego przetwarzania przy wyborze predyktorów**.

- **Dane wejściowe zawsze wybierane.** Korzystając z selektora zmiennych (przycisk po prawej stronie pola tekstowego) wybierz ze zbioru danych zmienne, które mają być zawsze używane przy budowaniu modelu sieci bayesowskiej. Zmienna przewidywana jest zawsze wybrana. Należy także zwrócić uwagę na to, że w procesie tworzenia modelu sieć Bayesa może usuwać elementy z tej listy, jeśli inne testy nie będą traktowały ich jako istotnych. Ta opcja po prostu zapewnia, że elementy na liście będą używane w procesie tworzenia modelu, a nie, że pojawią się w wynikowym modelu Bayesa.
- **Maksymalna liczba predyktorów.** Określ łączną liczbę predyktorów ze zbioru danych, które mają być używane przy budowaniu sieci bayesowskiej. Najwyższą liczbą, jaką można wprowadzić, jest łączna liczba zmiennych wejściowych w zbiorze danych.

  Uwaga: Jeśli liczba zmiennych wybranych w polu **Dane wejściowe zawsze wybierane** przekroczy wartość **Maksymalna liczba predyktorów**, zostanie wyświetlony komunikat o błędzie.

Modele użytkowe sieci bayesowskiej

Uwaga: Jeśli na karcie Model węzła modelowania wybrano opcję **Kontynuuj uczenie istniejącego modelu**, to informacje widoczne na karcie Model modelu użytkowego będą aktualizowane każdorazowo podczas ponownego generowania modelu.

Karta Model modelu użytkowego jest podzielona na dwa panele:

Lewy panel

Podstawowy Ten widok zawiera wykres sieci węzłów obrazujący relacje między zmienną przewidywaną a jej najważniejszymi predyktorami oraz relacje między predyktorami. Ważność poszczególnych predyktorów odzwierciedlona jest gęstością koloru; intensywny kolor oznacza, że predyktor jest ważny; słaby kolor — odwrotnie.

Wartości kategoryzacji węzłów reprezentujących przedziały są wyświetlane w podpowiedzi po zatrzymaniu wskaźnika myszy nad węzłem.

Za pomocą narzędzi produktu IBM SPSS Modeler przeznaczonych do tworzenia wykresów można interaktywnie korzystać z wykresów, edytować je i zapisywać, na przykład w celu wykorzystania w innej aplikacji, takiej jak MS Word.

Wskazówka: Jeśli sieć zawiera wiele węzłów, można kliknąć węzeł, aby go zaznaczyć, i przeciągnąć w celu poprawienia czytelności wykresu.

Rozkład Ten widok przedstawia w formie minikwykresu warunkowe prawdopodobieństwa poszczególnych węzłów sieci. Aby wyświetlić wartości, należy zatrzymać wskaźnik myszy nad wykresem.

Prawy panel

Ważność predyktorów Widok przedstawiający względną ważność poszczególnych predyktorów przy szacowaniu modelu. Aby uzyskać więcej informacji, zobacz “Ważność predyktorów” na stronie 43.

Prawdopodobieństwa warunkowe Gdy użytkownik zaznaczy węzeł lub miniwykres rozkładu w lewym panelu, w prawym panelu pojawia się tabela odpowiednich prawdopodobieństw warunkowych. Tabela ta zawiera wartości prawdopodobieństw warunkowych dla poszczególnych węzłów i kombinacji wartości w ich węzłach nadrzędnych. Ponadto zawiera liczbę rekordów zaobserwowanych dla każdej wartości i każdej kombinacji wartości w węzłach nadrzędnych.

Ustawienia modelu sieci bayesowskiej

Karta Ustawienia modelu użytkowego sieci bayesowskiej określa opcje modyfikacji budowanego modelu. Na przykład węzeł sieci bayesowskiej może służyć do budowy kilku różnych modeli z użyciem tych samych danych i ustawień. Następnie, za pomocą tej karty dla każdego z modeli można nieznacznie zmodyfikować ustawienia, obserwując jednocześnie, jak wpłynie to na wyniki.

Uwaga: Ta karta jest dostępna tylko po dodaniu modelu użytkowego do strumienia.

Wylicz surowe oceny skłonności. W przypadku modeli z przewidywaną zmienną typu flaga (zwracających predykcje tak lub nie) można wyliczyć oceny skłonności definiujące prawdopodobieństwo prawdziwego wyniku określonego dla zmiennej przewidywanej. Stanowią one uzupełnienie pozostałych współczynników ufności i wartości predykcyjnych, jakie mogą zostać wygenerowane podczas oceniania.

Wylicz skorygowane oceny skłonności. Surowe oceny skłonności bazują na danych uczących i mogą być zbyt optymistyczne z uwagi na tendencję wielu modeli do przeuczania tych danych. Opcja wyliczania skłonności skorygowanych próbuje skompensować tę tendencję na drodze oceny wydajności modelu w teście lub w podzbiorze walidacyjnym. Opcja ta wymaga, aby zmienna dzieląca na podzbiory była zdefiniowana w strumieniu oraz aby skorygowane oceny skłonności były włączone w węźle modelowania przed przystąpieniem do generowania modelu.

Dołącz wszystkie prawdopodobieństwa Określa, czy prawdopodobieństwa dla poszczególnych kategorii zmiennych wyjściowych są dodawane do poszczególnych rekordów przetwarzanych przez węzeł. Jeśli ta opcja nie zostanie wybrana, wówczas zostanie dodane tylko prawdopodobieństwo przewidywanej kategorii.

Ustawienie domyślne dla tego pola wyboru jest określane przez odpowiednie pole na karcie Zaawansowane w węźle modelowania. Więcej informacji można znaleźć w temacie “Opcje zaawansowane węzła sieci bayesowskiej” na stronie 132.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Podsumowanie modelu sieci bayesowskiej

Karta Podsumowanie modelu użytkowego zawiera informacje na temat samego modelu (*Analiza*), zmiennych użytych w modelu (*Zmienne*), ustawień użytych podczas budowania modelu (*Ustawienia budowania*) i uczenia modelu (*Podsumowanie uczenia*).

Podczas przeglądania węzła po raz pierwszy karta Podsumowanie jest zwinięta. Aby zobaczyć wyniki będące przedmiotem zainteresowania, należy użyć rozszerzanego elementu sterującego po lewej stronie pozycji, aby ją rozwinąć lub kliknąć przycisk **Rozwiń wszystko**, aby wyświetlić wszystkie wyniki. W celu ukrycia wyników po zakończeniu ich przeglądania należy użyć rozszerzanego elementu sterującego, aby zwinąć konkretne wyniki, jakie mają zostać ukryte lub kliknąć przycisk **Zwiń wszystko**, aby zwinąć wszystkie wyniki.

Analiza. Wyświetla informacje na temat konkretnego modelu.

Zmienne. Na liście znajdują się zmienne użyte jako zmienne przewidywane i wejściowe podczas budowania modelu.

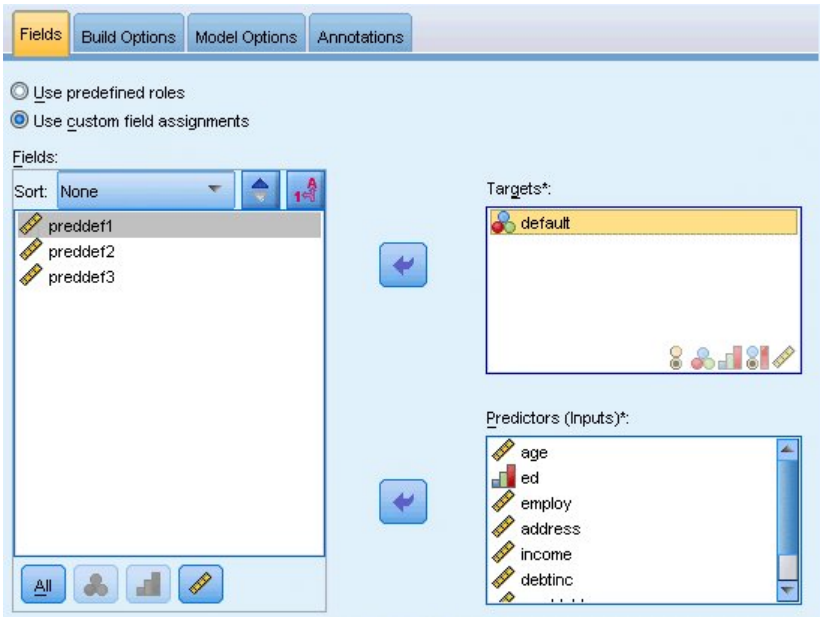
Ustawienia budowania. Zawiera informacje na temat ustawień użytych podczas budowania modelu.

Podsumowanie uczenia. Przedstawia typ modelu, strumień użyty do jego utworzenia, użytkownika, który go utworzył, informację, kiedy został utworzony oraz czas, jaki był potrzebny do zbudowania modelu.

Rozdział 8. Sieci neuronowe

Sieci neuronowe może przybliżyć szeroką gamę modeli predykcyjnych, przy czym charakteryzuje się minimalnymi wymaganiami w odniesieniu do struktury i założeń modelu. Forma powiązania jest określona podczas procesu uczenia. Jeśli odpowiednie jest powiązanie liniowe między zmiennymi przewidywanymi a predyktorami, wówczas wyniki sieci neuronowej powinny dokładnie przybliżać wyniki tradycyjnego modelu liniowego. Jeśli bardziej odpowiednie jest powiązanie nieliniowe, wówczas sieć neuronowa będzie automatycznie przybliżać „prawidłową" strukturę modelu.

Kosztem takiego elastycznego działania jest brak możliwości łatwej interpretacji sieci neuronowej. Jeśli próbujesz wyjaśnić proces bazowy, który wywołuje powiązania między zmienną przewidywaną a predyktorami, wówczas lepiej będzie użyć bardziej tradycyjnego modelu statystycznego. Jeśli jednak możliwość interpretacji modelu nie jest istotna, dobre wyniki można uzyskać, stosując sieć neuronową.



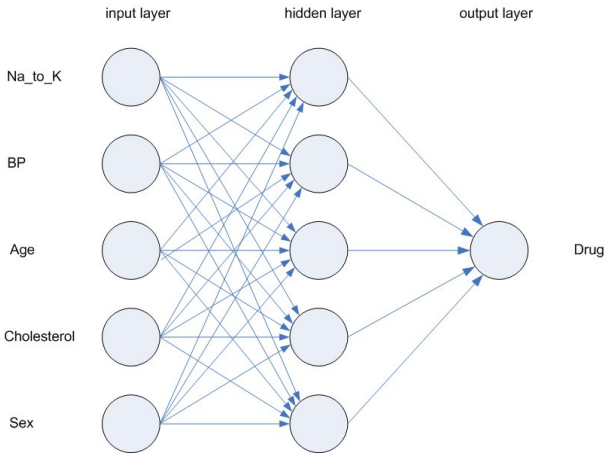
Rysunek 30. Zakładka zmiennych

Wymagania dotyczące zmiennych. Musi istnieć co najmniej jedna zmienna przewidywana i jedna zmienna wejściowa. Zmienne ustawione jako Łącznie i Brak są ignorowane. Nie istnieją żadne ograniczenia dotyczące poziomu pomiarów dla zmiennych przewidywanych ani predyktorów (zmiennych wejściowych). Więcej informacji można znaleźć w "Opcje zmiennych węzła modelowania" na stronie 31.

Wagi początkowe są przypisywane do sieci neuronowych podczas budowania modeli, dlatego wygenerowany model ostateczny zależy od porządku zmiennych w danych. SPSS Modeler sortuje dane automatycznie według nazwy zmiennej przed zaprezentowaniem ich sieci neuronowej w celu uczenia. Oznacza to, że jawna zmiana porządku zmiennych w poprzedzających danych nie ma wpływu na wygenerowane modele sieci neuronowych, gdy w konstruktorze modelu ustawiona jest wartość początkowa generatora liczb losowych. Jednak zmiana nazw zmiennych wejściowych na takie, które spowodują zmianę porządku sortowania spowoduje wygenerowanie innych modeli sieci neuronowych, nawet jeśli w konstruktorze modelu będzie ustawiona wartość początkowa generatora liczb losowych. Inny porządek sortowania zmiennych nie wpłynie istotnie na jakość modelu.

Model sieci neuronowych

Sieci neuronowe są prostymi modelami działającymi w sposób przypominający układ nerwowy. Podstawowymi jednostkami są **neurony**, które zwykle są rozmieszczone w **warstwach**, co przedstawia następujący rysunek.



Rysunek 31. Struktura sieci neuronowej

Sieci neuronowe to uproszczony model procesu przetwarzania informacji przez ludzki umysł. Polega on na symulowaniu dużej liczby połączonych wzajemnie jednostek przetwarzania, które przypominają abstrakcyjne wersje neuronów.

Jednostki przetwarzania są rozmieszczone w postaci warstw. Zwykle sieć neuronowa składa się z trzech części: **warstwa wejściowa**, z jednostkami reprezentującymi zmienne wejściowe, co najmniej jedna **warstwa ukryta** oraz **warstwa wyjściowa**, z jednostkami reprezentującymi zmienne przewidywane. Jednostki są połączone połączeniami o różnej sile (lub **wadze**). Dane wejściowe są prezentowane na pierwszej warstwie, a wartości są przekazywane z poszczególnych neuronów do wszystkich neuronów w następnej warstwie. Ostatecznie z warstwy wyjściowej uzyskiwany jest wynik.

Sieć uczy się poprzez sprawdzanie pojedynczych rekordów, generowanie predykcji dla poszczególnych rekordów i wprowadzanie korekty wag, jeśli powodują niepoprawną predykcję. Ten proces jest powtarzany wiele razy, a sieć coraz bardziej udoskonala predykcje, aż do spełnienia co najmniej jednego kryterium zatrzymywania.

Początkowo wszystkie wagi mają charakter losowy, a odpowiedzi wychodzące z sieci są prawdopodobnie bezsensowne. Sieć rozwija się w procesie **uczenia**. Przykłady, dla których wynik jest znany, są wielokrotnie wprowadzane do sieci, a przedstawiane odpowiedzi są porównywane ze znanymi wynikami. Informacje uzyskane na podstawie tego porównania są z powrotem przekazywane do sieci, przy czym następuje stopniowa zmiana wag. W trakcie uczenia sieć staje się coraz bardziej dokładna w replikowaniu znanych wyników. Po zakończeniu uczenia sieć można zastosować do przyszłych obserwacji, w których wynik jest nieznany.

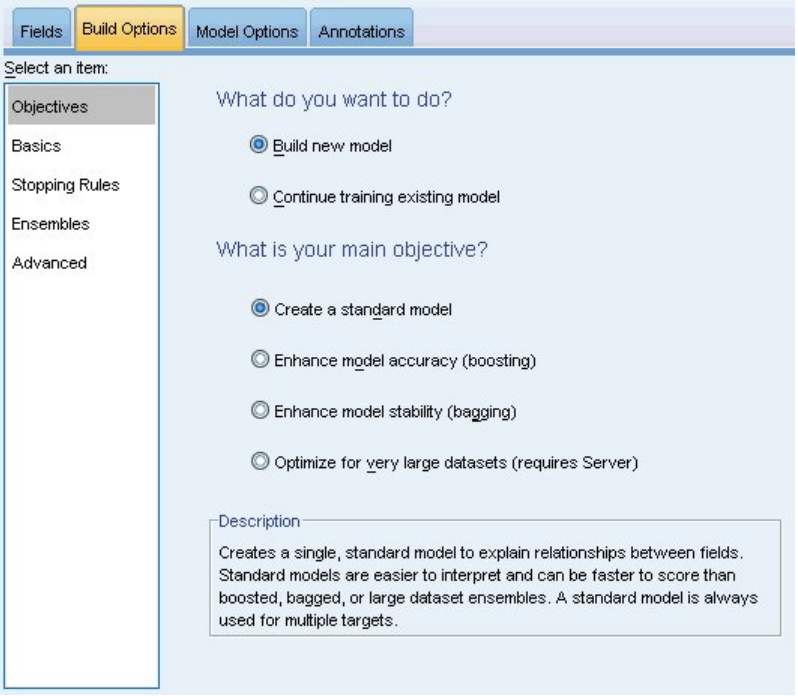
Korzystanie z sieci neuronowych ze starszymi strumieniami

W wersji 14 produktu IBM SPSS Modeler wprowadzono nowy węzeł sieci neuronowej obsługujący techniki boosting i agregacji bootstrapowej, a także optymalizację dla bardzo dużych zbiorów danych. Istniejące strumienie zawierające stary węzeł nadal mogą budować i oceniać model w tej wersji produktu. Jednak w przyszłej wersji te funkcje będą niedostępne, dlatego zalecamy używanie nowej wersji.

Począwszy od wersji 13 zmienne o nieznanych wartościach (czyli wartościach nieobecnych w danych uczących) nie są automatycznie traktowane jako brakujące wartości, a ich ocena zwraca wartość $null$. Oznacza to, że jeśli zmienne z brakującymi wartościami mają być oceniane w wersji 13 lub późniejszej jako wartości inne niż null przy użyciu starszego modelu sieci neuronowej (z wersji wcześniejszych niż 13), wówczas należy oznaczyć nieznane wartości jako brakujące wartości (na przykład, stosując węzeł Typ).

Należy zwrócić uwagę na to, że w celu zapewnienia zgodności starsze strumienie, które nadal zawierają stary węzeł, mogą nadal używać opcji *ograniczenia rozmiaru zbioru*, która jest dostępna po wybraniu opcji **Narzędzia > Właściwości strumienia > Opcje**; ta opcja obowiązuje tylko w przypadku sieci Kohonena i węzłów *K*-średnich z wersji 14 i kolejnych.

Cele



Rysunek 32. Ustawienia celów

Co zamierzasz zrobić?

- **Utworzyć nowy model.** Stworzyć całkowicie nowy model. Jest to zwykle działanie węzła.
- **Kontynuować uczenie istniejącego modelu.** Szkolenie jest kontynuowane z wykorzystaniem ostatniego modelu, utworzonego z powodzeniem przez węzeł. Dzięki temu możliwa jest aktualizacja lub odświeżenie istniejącego modelu bez konieczności wejścia do oryginalnych danych i może skutkować znacznie wydajniejszym działaniem, ponieważ tylko nowe lub zaktualizowane rekordy są podawane do strumienia. Szczegóły dotyczące poprzedniego modelu są zapisywane z węzłem modelowania, umożliwiając używanie tej opcji nawet, jeśli poprzednie wartościowe informacje z modelu są już niedostępne w strumieniu lub w palecie Modeli.

  Uwaga: Gdy opcja ta jest włączona, wszystkie inne elementy sterowania w zakładkach Zmienne i Opcje budowania są wyłączone.

Jaki chcesz osiągnąć cel? Zaznacz odpowiedni cel.

- **Zbudować model standardowy.** Ta metoda tworzy pojedynczy model do przewidywania przy pomocy predyktorów. Ogólnie rzecz biorąc standardowe modele są łatwiejsze w interpretacji i można je szybciej ocenić w porównaniu ze wzmocnionymi, spakowanymi lub dużymi zestawami zbiorów danych.

  Uwaga: Aby użyć tej opcji w modelach rozdzielonych razem z opcją **Kontynuuj uczenie istniejącego modelu** należy mieć połączenie z produktem Analytic Server.
- **Zwiększyć dokładność modelu (boosting).** Metoda ta tworzy model zespolony przy pomocy wzmocnienia, który generuje sekwencję modeli w celu uzyskania bardziej precyzyjnych predykcji. Tworzenie i ocena zestawów mogą trwać dłużej niż w przypadku standardowego modelu.

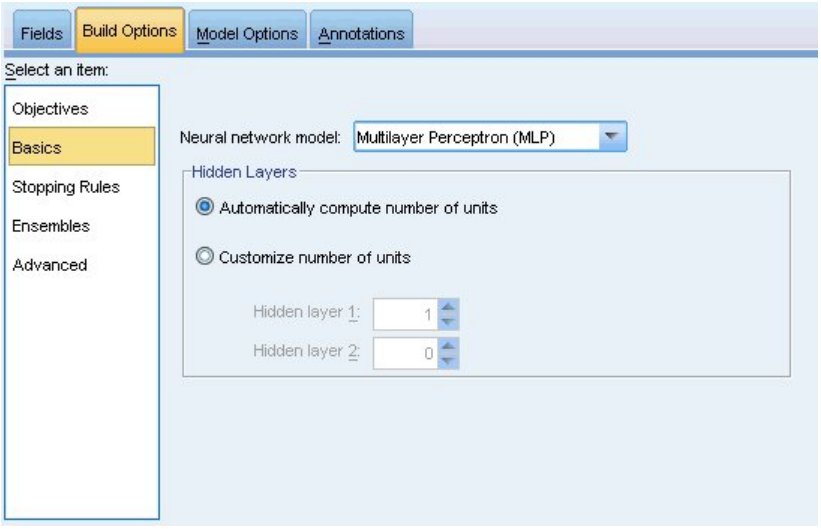
Wzmocnienie tworzy kolejność „modeli składników”, z których każdy został skompilowany na podstawie całego zbioru danych. Przed skompilowaniem każdego kolejnego modelu składników, rekordy są ważone na podstawie reszt po poprzednich modelach składników. Obserwacje o dużej wartości reszt dają stosunkowo wyższe wagi analizy tak, że kolejny model składników będzie się skupiał na dobrym przewidywaniu tych rekordów. Te modele składników tworzą razem model zespolony. Model zespolony ocenia nowe rekordy przy pomocy reguły łączenia; dostępne reguły zależą od poziomu pomiaru celu.

- **Wzmocnić stabilność modelu (agregacja bootstrapowa).** Metoda ta tworzy model zespolony przy pomocy spakowania (agregacja metodą bootstrap), które generuje wiele modeli w celu uzyskania bardziej wiarygodnych predykcji. Tworzenie i ocena zestawów mogą trwać dłużej niż w przypadku standardowego modelu.

  Agregacja metodą bootstrap (bagging) powiela zespół danych z przyuczenia, tworząc próbkowanie poprzez zastąpienie oryginalnego zbioru danych. W wyniku tego powstają próby bootstrap, które mają taki sam rozmiar, jak oryginalny zbiór danych. Następnie na podstawie każdego powielania kompilowany jest „model składników”. Te modele składników tworzą razem model zespolony. Model zespolony ocenia nowe rekordy przy pomocy reguły łączenia; dostępne reguły zależą od poziomu pomiaru celu.
- **Utworzyć model dla dużych zbiorów danych.** Metoda ta tworzy model zespolony przez podział zbioru danych na oddzielne bloki danych. Wybierz tę opcję, jeśli Twój zbiór danych jest zbyt duży do utworzenia któregokolwiek z powyższych modeli, lub aby utworzyć model przyrostowy. Tworzenie tej opcji może być szybsze, ale ocena może potrwać dłużej niż w przypadku standardowego modelu.

Jeśli istnieje wiele zmiennych przewidywanych, ta metoda spowoduje utworzenie modelu standardowego bez względu na wybrany cel.

Podstawowe



Rysunek 33. Ustawienia Podstawowe

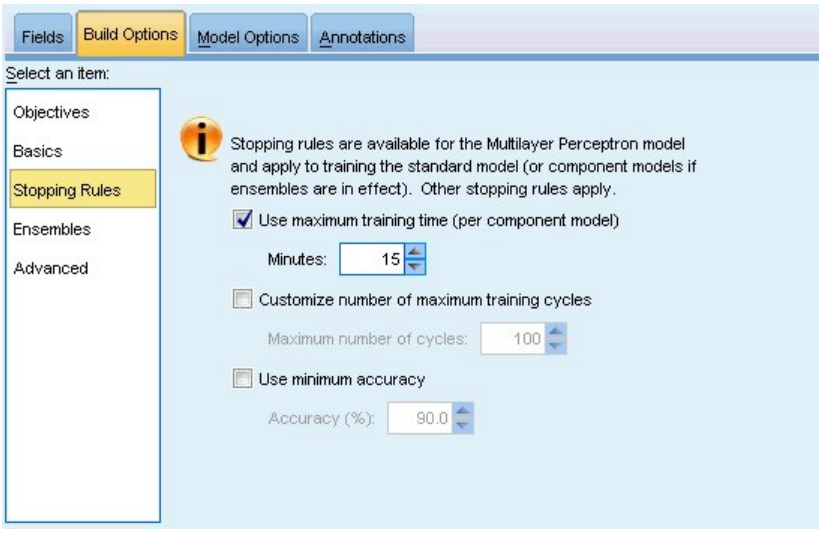
Model sieci neuronowej. Typ modelu określa sposób, w jaki sieć łączy predyktory ze zmiennymi przewidywanymi za pośrednictwem warstw ukrytych. **Perceptron wielowarstwowy (MLP)** umożliwia bardziej złożone powiązania, a możliwą konsekwencją jest wydłużenie czasu uczenia i oceny. **Radialna funkcja bazowa (RBF)** może skracać czas uczenia i czas oceny, a możliwą konsekwencją jest obniżenie jakości predykcji w porównaniu do MLP.

Warstwy ukryte. Warstwy ukrytej sieci neuronowej zawierają jednostki, które nie mogą być obserwowane. Wartością każdej jednostki ukrytej jest pewna funkcja predyktorów; dokładna forma funkcji jest zależna częściowo od typu sieci. Perceptron wielowarstwowy może obejmować jedną lub dwie warstwy ukryte; radialna funkcja bazowa może obejmować jedną warstwę ukrytą.

- **Automatycznie wyliczona liczba neuronów.** Ta opcja umożliwia zbudowanie sieci z jedną warstwą ukrytą i oblicza „najlepszą” liczbę jednostek w warstwie ukrytej.
- **Ustalona liczba neuronów.** Ta opcja umożliwia określenie liczby jednostek w każdej warstwie ukrytej. Pierwsza warstwa ukryta musi zawierać co najmniej jedną jednostkę. Określenie 0 jednostek dla drugiej warstwy ukrytej powoduje zbudowanie perceptronu wielowarstwowego z pojedynczą warstwą ukrytą.

  Uwaga: Wartości należy wybierać w taki sposób, aby liczba węzłów nie przekraczała liczby predyktorów ilościowych powiększonej o łączną liczbę kategorii wśród wszystkich predyktorów jakościowych (flaga, nominalne i porządkowe).

Reguły zatrzymujące



Rysunek 34. Ustawienia Reguły zatrzymujące

Są to reguły określające okoliczności zatrzymania uczenia sieci mające postać wielowarstwowych perceptronów; te ustawienia są ignorowane, gdy używany jest algorytm radialnej funkcji bazowej. Uczenie jest kontynuowane przez co najmniej jeden cykl (przekazywanie danych) i może zostać zatrzymane zgodnie z poniższymi kryteriami.

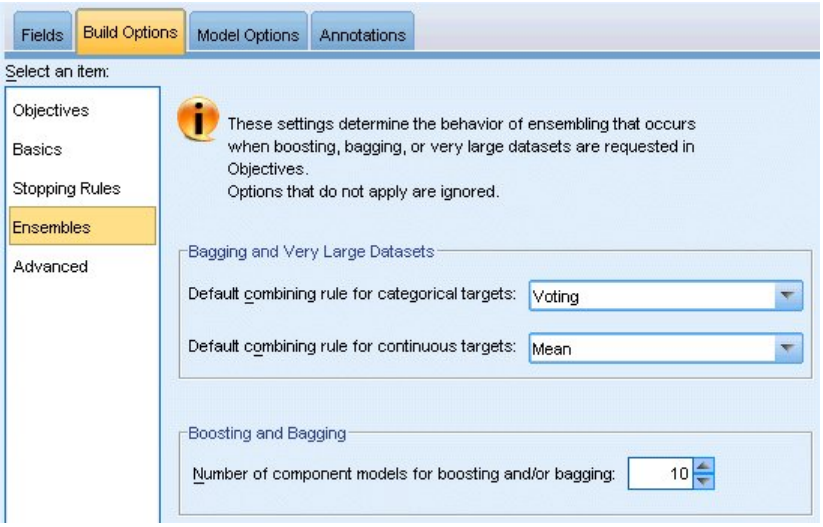
Użyj maksymalnego czasu uczenia (na model zespolony). Ta opcja umożliwia określenie maksymalnej liczby minut, przez jaką będzie działał algorytm. Określ wartość większą od 0. Gdy zostanie zbudowany model zespolony, będzie to czas uczenia dozwolony dla każdego modelu zespolonego z zespołu. Należy zwrócić uwagę na to, że w celu ukończenia bieżącej epoki uczenie może trwać trochę dłużej niż podany czas.

Maksymalna liczba epok uczenia. Maksymalna dozwolona liczba epok uczenia. Jeśli maksymalna liczba epok zostanie przekroczona, uczenie zostanie zatrzymane. Podaj liczbę całkowitą większą od 0.

Użyj minimalnej dokładności. W przypadku wyboru tej opcji uczenie będzie kontynuowane do czasu osiągnięcia określonej dokładności. To może nigdy nie nastąpić, ale użytkownik może przerwać uczenie w dowolnym momencie i zachować sieć z największą osiągniętą dokładnością.

Algorytm uczący również zostanie zatrzymany, jeśli błąd w zbiorze zabezpieczającym przed przeuczeniem nie ulegnie zmniejszeniu po każdej epoce, jeśli względna zmiana błędu uczenia jest niewielka lub jeśli współczynnik bieżącego błędu uczenia jest niewielki w porównaniu do błędu początkowego.

Zespoły



Rysunek 35. Ustawienia Zespoły

Ustawienia te determinują zachowanie tworzenia zespołów, które występuje, gdy w Celach pożądane jest wspomaganie, agregacja metodą bootstrap lub bardzo duże zbiory danych. Opcje, które nie mają zastosowania do wybranego celu są ignorowane.

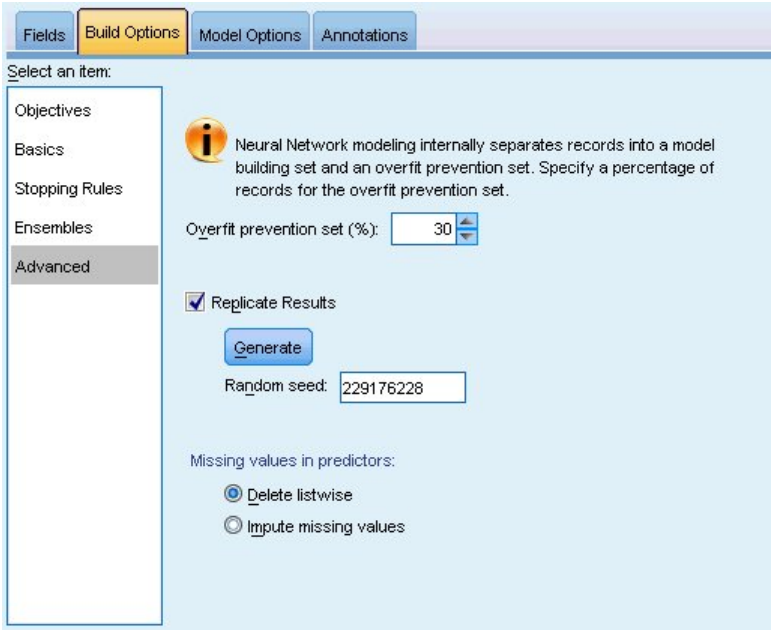
Agregacja metodą bootstrap i bardzo duże zbiory danych. Podczas wybierania zestawu jest to reguła służąca do łączenia przewidywanych wartości z modeli podstawowych w celu wyliczenia wartości oceny zestawu.

- **Domyślna reguła zespolenia dla przewidywanych zmiennych jakościowych.** Przewidywane wartości zestawu dla przewidywanych zmiennych jakościowych mogą być połączone przy pomocy głosowania, największego prawdopodobieństwa lub największego, średniego prawdopodobieństwa. **Głosowanie** wybiera kategorię, która ma największe prawdopodobieństwo, najczęściej wśród modeli podstawowych. **Największe prawdopodobieństwo** wybiera kategorię, która uzyskuje największe, pojedyncze prawdopodobieństwo wśród modeli podstawowych. **Największe średnie prawdopodobieństwo** wybiera kategorię z najwyższą wartością, gdy prawdopodobieństwa kategorii wśród modeli podstawowych są uśrednione.
- **Domyślna reguła zespolenia dla docelowych wartości ilościowych.** Przewidywane wartości zestawu dla jakościowych zmiennych docelowych można połączyć przy pomocy średniej lub mediany przewidywanych wartości z modeli podstawowych.

Należy zwrócić uwagę, że gdy celem jest zwiększenie dokładności modelu, wybory reguły łączenia są ignorowane. Wzmocnienie zawsze wykorzystuje głos ważonej większości do oceny jakościowych zmiennych docelowych i ważonej mediany do oceny jakościowych zmiennych docelowych.

Boosting i agregacja bootstrapowa. Podaj liczbę modeli podstawowych do utworzenia, gdy celem jest zwiększenie dokładności lub stabilności modelu; dla agregacji metodą bootstrap jest to liczba prób agregacji metodą bootstrap. Powinna to być dodatnia liczba całkowita.

Zaawansowane



Rysunek 36. Ustawienia Zaawansowane

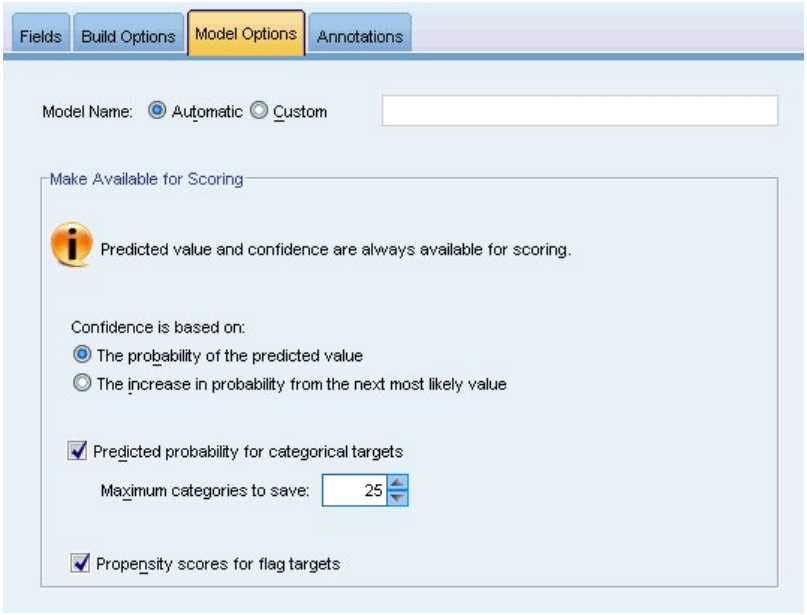
Ustawienia zaawansowane zapewniają kontrolę nad opcjami, które nie należą do innych grup ustawień.

Zbiór zabezpieczający przed przeuczeniem. Metoda sieci neuronowej wewnętrznie rozdziela rekordy między podzbiór budowania modelu oraz zbiór zabezpieczający przed przeuczeniem, który jest niezależnym zbiorem rekordów danych używanym do śledzenia błędów podczas uczenia i zapobiegania modelowaniu przez metodę zmienności prawdopodobieństwa w danych. Należy określić procent rekordów. Domyślną wartością jest 30.

Replikacja wyników. Ustawienie wartości początkowej generatora liczb losowych umożliwia powielenie analizy. Podaj liczbę całkowitą lub kliknij przycisk **Generuj**, co spowoduje utworzenie pseudolosowej liczby całkowitej między 1 a 2147483647, włącznie. Domyślnie analizy są replikowane z wartością startową generatora 229176228.

Braki danych w predyktorach. Ta opcja określa sposób postępowania z brakami danych. **Usuń obserwacjami** — usuwa z budowania modelu rekordy z brakującymi wartościami w predyktorach. **Uwzględnij braki danych** — powoduje zastąpienie brakujących wartości w predyktorach i użycie tych rekordów w analizie. Zmienne ilościowe podstawiają średnią z minimalnej i maksymalnej wartości obserwowanej; zmienne jakościowe podstawiają najczęściej występującą kategorię. Należy zwrócić uwagę na to, że rekordy z brakującymi wartościami w dowolnej innej zmiennej określonej na karcie Zmienne są zawsze usuwane z budowania modelu.

Opcje modelu



Rysunek 37. Karta Opcje modelu

Nazwa modelu. Można automatycznie generować nazwę modelu na podstawie zmiennych docelowych lub podać nazwę użytkownika. Automatycznie wygenerowana nazwa jest nazwą zmiennej docelowej. Jeśli istnieje więcej niż jedna zmienna przewidywana, nazwa modelu składa się z listy nazw zmiennych połączonych ampersandami. Na przykład, jeśli zmienne przewidywane to *field1 field2 field3*, nazwa modelu będzie miała postać: *field1 & field2 & field3*.

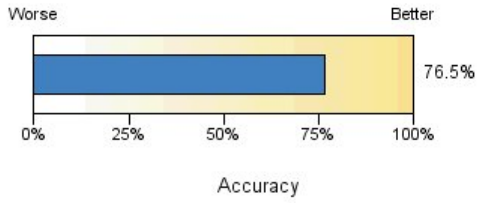
Udostępnij do oceniania. Podczas oceniania modelu powinny być wygenerowane zaznaczone elementy z tej grupy. Przy ocenie modelu zawsze obliczana jest wartość przewidywana (dla wszystkich zmiennych przewidywanych) i ufność (dla jakościowych zmiennych przewidywanych). Obliczona ufność może być oparta na prawdopodobieństwie przewidywanej wartości (najwyższe przewidywane prawdopodobieństwo) lub różnicy między najwyższym przewidywanym prawdopodobieństwem a drugim co do wysokości przewidywanym prawdopodobieństwem.

- **Przewidywane prawdopodobieństwo dla przewidywanych zmiennych jakościowych.** Generuje przewidywane prawdopodobieństwa dla jakościowych zmiennych przewidywanych. Dla każdej kategorii tworzona jest jedna zmienna.
- **Oceny skłonności (ważne tylko dla przewidywanych zmiennych typu flaga).** W przypadku modeli z przewidywaną zmienną typu flaga (zwracających predykcje tak lub nie) można wyliczyć oceny skłonności definiujące prawdopodobieństwo prawdziwego wyniku określonego dla zmiennej przewidywanej. Model generuje surowe oceny skłonności; jeśli stosowane są podzbiory, model generuje także skorygowane oceny skłonności na podstawie podzbioru testowego.

Podsumowanie modelu

Model Summary

Target	Previously defaulted
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	4



Rysunek 38. Widok Podsumowanie modelu sieci neuronowych

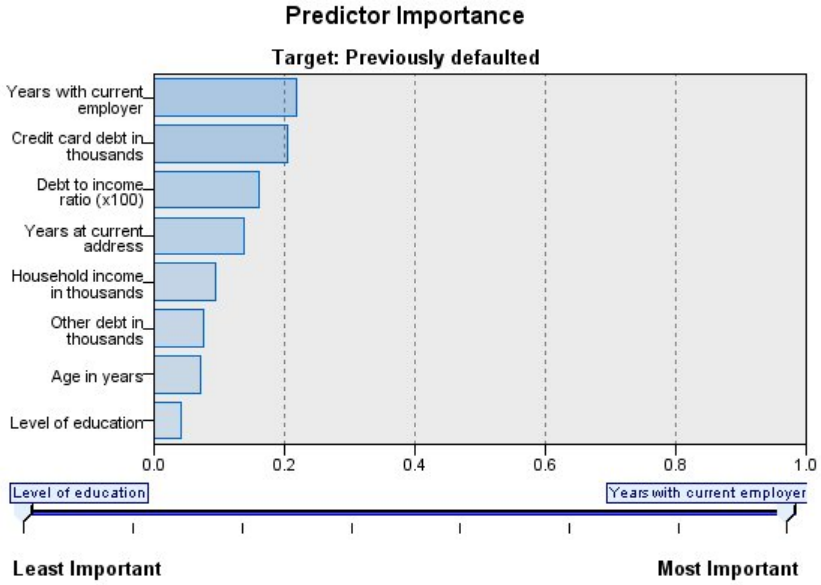
Widok Podsumowanie modelu stanowi obraz stanu, dostępne na pierwszy rzut oka podsumowanie dokładności predykcyjnej lub dokładności klasyfikacji sieci neuronowej.

Podsumowanie modelu. Tabela przedstawia: zmienną przewidywaną; typ uczonej sieci neuronowej; regułę zatrzymującej, która zatrzymała uczenie (jest widoczna, jeśli uczona była sieć w postaci perceptrona wielowarstwowego), a także liczbę neuronów w każdej ukrytej warstwie sieci.

Jakość sieci neuronowej. Na wykresie przedstawiono dokładność modelu finalnego, który jest przedstawiony w większym i lepszym formacie. Dla jakościowej zmiennej przewidywanej dokładność jest zwykłą wartością procentową rekordów, dla których przewidywana wartość jest zgodna z zaobserwowaną wartością. W przypadku ilościowej zmiennej przewidywanej dokładność jest podawana jako wartość R^2.

Wiele zmiennych przewidywanych. Jeśli istnieje wiele zmiennych przewidywanych, wówczas każda zmienna przewidywana jest wyświetlana w wierszu **Zmienna przewidywana** w tabeli. Dokładność wyświetlana w wykresie jest średnią z dokładności poszczególnych zmiennych przewidywanych.

Ważność predyktora

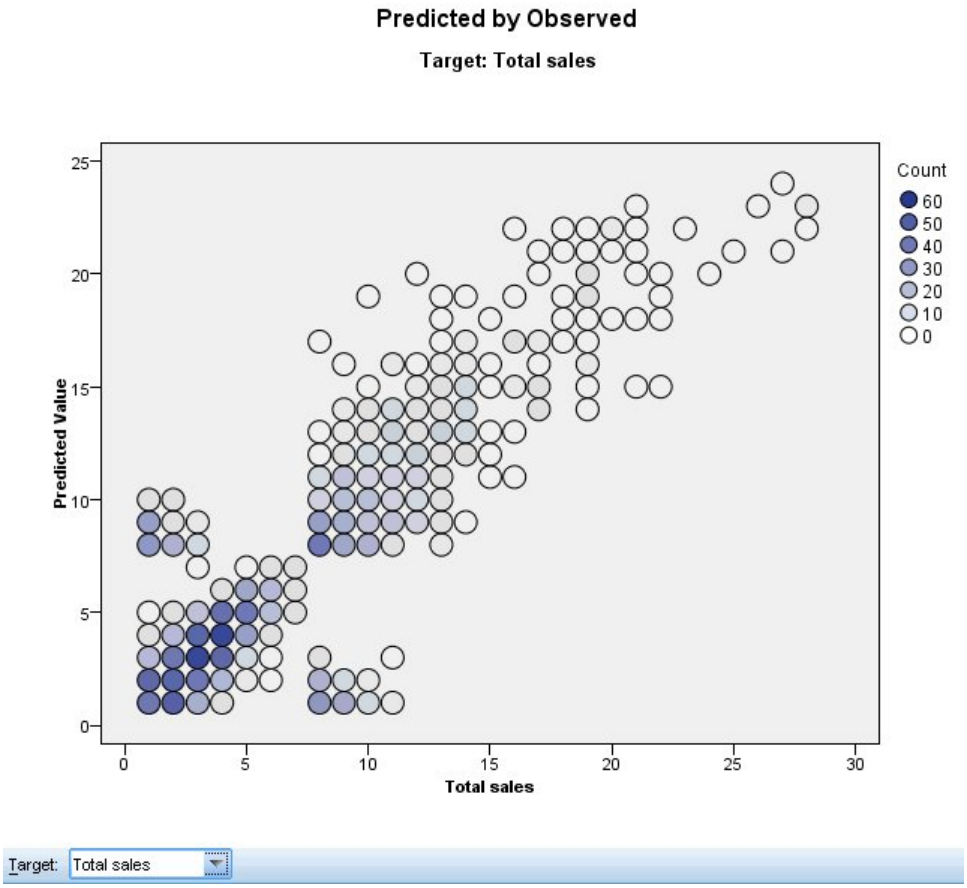


Rysunek 39. Widok Ważność predyktorów

Zazwyczaj działania modelujące mają koncentrować się na zmiennych predyktorów, które są najważniejsze, a opuszczane lub ignorowane mają być te zmienne, które są najmniej ważne. Wykres ważności predyktorów pomaga osiągnąć ten cel przez wskazanie względnej ważności każdego predyktora przy szacowaniu modelu. Ponieważ wartości są względne, suma wartości wszystkich wyświetlanych predyktorów wynosi 1,0. Ważność predyktora nie jest powiązana z dokładnością modelu. Jest powiązana z ważnością każdego predyktora przy prognozach, a nie z tym, czy taka prognoza jest dokładna.

Wiele zmiennych przewidywanych. Jeśli istnieje wiele zmiennych przewidywanych, wówczas każda zmienna przewidywana jest wyświetlana w osobnym wykresie i istnieje lista rozwijana **Zmienna przewidywana**, która kontroluje wyświetlane zmienne przewidywane.

Przewidywane według obserwowanych

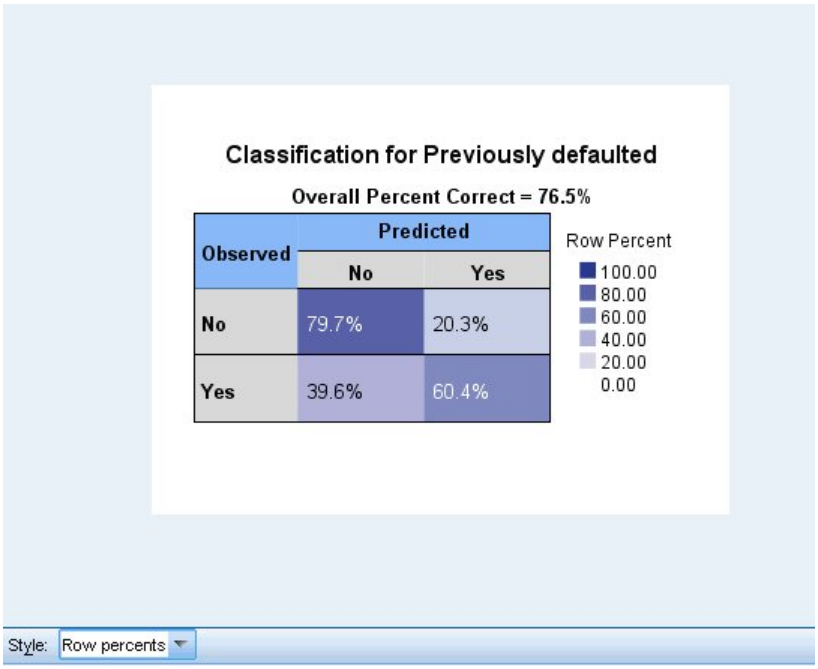


Rysunek 40. Widok Przewidywane według obserwowanych

W przypadku ilościowych zmiennych przewidywanych ten widok przedstawia wykres rozrzutu z kategoryzacją przewidywanych wartości na osi pionowej przez obserwowane wartości na osi poziomej.

Wiele zmiennych przewidywanych. Jeśli istnieje wiele ilościowych zmiennych przewidywanych, wówczas każda zmienna przewidywana jest wyświetlana w osobnym wykresie i istnieje lista rozwijana **Zmienna przewidywana**, która kontroluje wyświetlane zmienne przewidywane.

Klasyfikacja



Rysunek 41. Widok Klasyfikacja, styl procentu w wierszach

W przypadku jakościowych zmiennych przewidywanych wyświetlana jest klasyfikacja krzyżowa wartości obserwowanych względem przewidywanych w mapie natężeń, a dodatkowo przedstawiany jest ogólny procent poprawnych.

Style tabel. Istnieje kilka różnych stylów wyświetlania, które są dostępne z poziomu listy rozwijanej **Styl**.

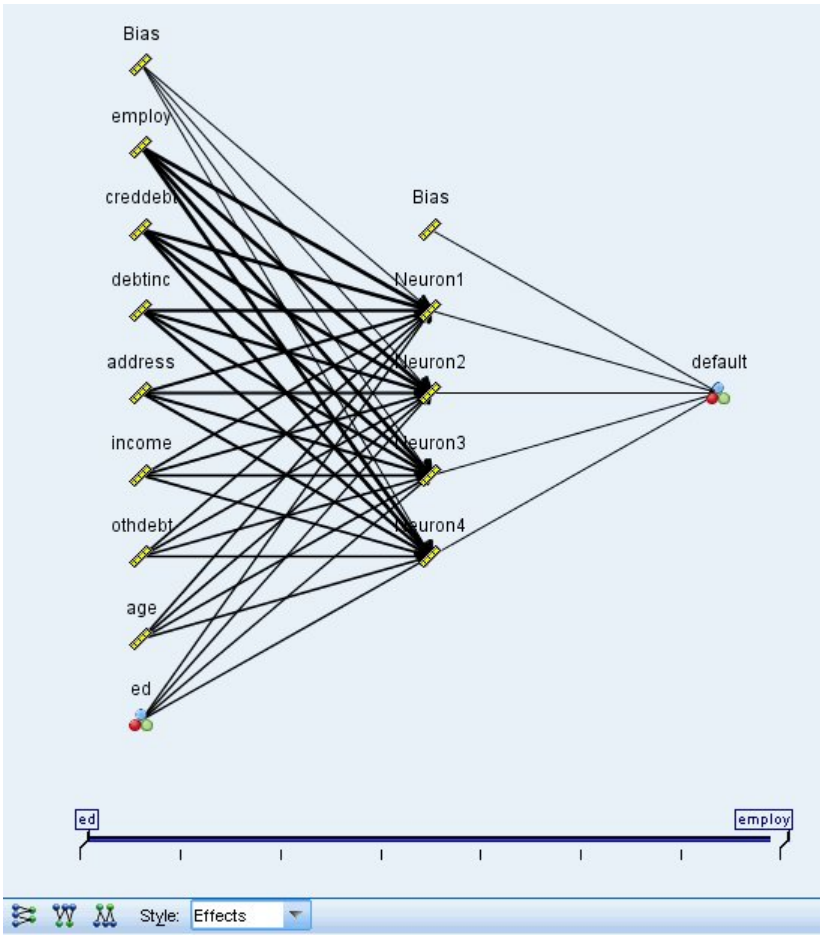
- **Procent w wierszu.** Wyświetla procent w wierszu (liczby komórek wyrażone jako procent sum z wierszy) w komórkach. Jest to ustawienie domyślne.
- **Liczby komórek.** Wyświetla liczby komórek w komórkach. Cieniowanie mapy komórek jest w dalszym ciągu wyrażone jako procent w wierszu.
- **Mapa natężeń.** W komórkach nie są wyświetlane żadne wartości, tylko cieniowanie.
- **Skompresowane.** Nagłówki kolumn ani wierszy nie są wyświetlane. Nie są wyświetlane również wartości w komórkach. Taki styl może być użyteczny, gdy zmienna przewidywana zawiera wiele kategorii.

Braki danych. Jeśli w jakichkolwiek rekordach brakuje wartości w zmiennej przewidywanej, wówczas takie rekordy są wyświetlane w wierszu **(Braki danych)**, który jest wyświetlany pod poprawnymi wierszami. Rekordy z brakami danych nie zwiększają ogólnego procentu poprawnych.

Wiele zmiennych przewidywanych. Jeśli istnieje wiele przewidywanych zmiennych jakościowych, wówczas każda zmienna przewidywana jest wyświetlana w osobnej tabeli i istnieje lista rozwijana **Zmienna przewidywana**, która kontroluje wyświetlane zmienne przewidywane.

Duże tabele. Jeśli wyświetlana zmienna przewidywana zawiera ponad 100 kategorii, żadna tabela nie jest wyświetlana.

Sieć



Rysunek 42. Widok Sieć, dane wejściowe po lewej stronie, styl efektów

Jest to graficzne odzwierciedlenie sieci neuronowej.

Style wykresu. Dostępne są dwa różne style wyświetlania, które są dostępne z poziomu listy rozwijanej **Styl**.

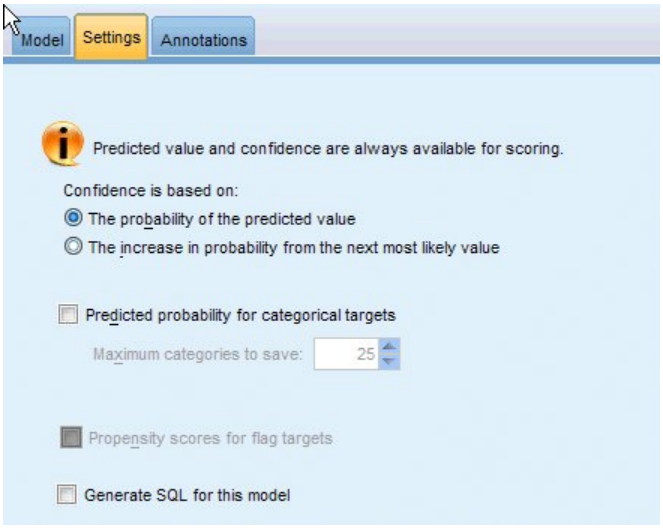
- **Efekty.** Każdy predyktor i zmienna przewidywana są wyświetlane jako osobny węzeł w diagramie bez względu na to, czy skala pomiaru jest ilościowa, czy jakościowa. Jest to ustawienie domyślne.
- **Współczynniki.** Wyświetlanych jest wiele węzłów wskaźnikowych dla predyktorów jakościowych i jakościowych zmiennych przewidywanych. Linie łączące w diagramie w stylu Współczynniki mają kolory zależne od oszacowanej wartości wagi synaptycznej.

Orientacja diagramu. Domyślnie diagram struktury sieci jest ułożony w taki sposób, że dane wejściowe są widoczne po lewej stronie, a zmienne przewidywane po prawej stronie. Korzystając z elementów sterujących na pasku narzędzi użytkownik może zmienić orientację w taki sposób, że dane wejściowe będą wyświetlane u góry, a zmienne przewidywane u dołu albo dane wejściowe u dołu, a zmienne przewidywane u góry.

Ważność predyktora. Linie łączące na diagramie są ważone na podstawie ważności predyktorów, przy czym grubsza linia odpowiada większej ważności. Na pasku narzędzi dostępny jest suwak Ważność predyktorów, który kontroluje predyktory widoczne na diagramie sieci. Nie zmienia to modelu, ale pozwala po prostu na skoncentrowanie się na najistotniejszych predyktorach.

Wiele zmiennych przewidywanych. Jeśli istnieje wiele zmiennych przewidywanych, wówczas wszystkie są wyświetlane na wykresie.

Ustawienia



Rysunek 43. Karta Ustawienia

Podczas oceniania modelu powinny być wygenerowane zaznaczone elementy z tej karty. Przy ocenie modelu zawsze obliczana jest wartość przewidywana (dla wszystkich zmiennych przewidywanych) i ufność (dla jakościowych zmiennych przewidywanych). Obliczona ufność może być oparta na prawdopodobieństwie przewidywanej wartości (najwyższe przewidywane prawdopodobieństwo) lub różnicy między najwyższym przewidywanym prawdopodobieństwem a drugim co do wysokości przewidywanym prawdopodobieństwem.

- **Przewidywane prawdopodobieństwo dla przewidywanych zmiennych jakościowych.** Generuje przewidywane prawdopodobieństwa dla jakościowych zmiennych przewidywanych. Dla każdej kategorii tworzona jest jedna zmienna.
- **Oceny skłonności (ważne tylko dla przewidywanych zmiennych typu flaga).** W przypadku modeli z przewidywaną zmienną typu flaga (zwracających predykcje tak lub nie) można wyliczyć oceny skłonności definiujące prawdopodobieństwo prawdziwego wyniku określonego dla zmiennej przewidywanej. Model generuje surowe oceny skłonności; jeśli stosowane są podzbiory, model generuje także skorygowane oceny skłonności na podstawie podzbioru testowego.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.

Przeprowadź ocenę, wykorzystując natywny kod SQL Jeśli ta opcja jest wybrana, generowany jest natywny kod SQL w celu oceny modelu w bazie danych.

Uwaga: Ta opcja może szybciej zwracać wyniki, ale rozmiar i złożoność natywnego kodu SQL wzrastają wraz ze wzrostem złożoności modelu.

Przeprowadź ocenę poza bazą danych Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Rozdział 9. Lista decyzyjna

Modele typu Decision List ujawniają podgrupy lub **segmenty** wykazujące wyższe lub niższe prawdopodobieństwo danego wyniku binarnego (tak albo nie) względem całej próby. Można na przykład wyszukać klientów, których prawdopodobieństwo odejścia jest najmniejsze, lub którzy z największym prawdopodobieństwem pozytywnie zareagują na kampanię. Decision List Viewer zapewnia użytkownikowi pełną kontrolę nad modelem, umożliwiając edytowanie segmentów, dodawanie własnych reguł biznesowych, określanie sposobu oceny każdego segmentu i modyfikowanie modelu na różne inne sposoby w celu zoptymalizowania odsetka trafień we wszystkich segmentach. Dlatego narzędzie to szczególnie dobrze nadaje się do generowania list mailingowych lub wybierania rekordów, do których powinna być skierowana konkretna kampania. Można wykorzystać wiele **zadań eksploracji**, by łącznie zastosować różne techniki modelowania — np. ujawnić najlepsze i najgorsze segmenty w jednym modelu i uwzględnić lub wykluczyć takie segmenty na etapie oceniania.

Segmenty, reguły i warunki

Model składa się z listy segmentów, a każdy segment jest zdefiniowany przez regułę, która wybiera dopasowane rekordy. Jedna reguła może składać się z wielu warunków, na przykład:

```
RFM_SCORE > 10 and
MONTHS_CURRENT <= 9
```

Reguły stosowane są w kolejności, w jakiej są wymienione, przy czym pierwsza spełniona reguła determinuje wynik dla danego rekordu. Rozpatrywane niezależnie reguły lub warunki mogą zachodzić na siebie, jednak kolejność reguł w pełni je ujednoznacznia. Jeśli żadna reguła nie jest spełniona, rekord jest przypisywany do pozostałości.

Pełna kontrola nad ocenianiem

Decision List Viewer umożliwia przeglądanie, modyfikowanie i reorganizowanie segmentów oraz włączanie i wykluczanie ich do/z oceniania. Możemy na przykład wykluczyć jedną grupę klientów z przyszłych ofert i uwzględnić inną grupę, by natychmiast zobaczyć, jak wpłynie to na ogólny wskaźnik trafień. Modele typu Decision List zwracają ocenę *Tak* dla segmentów uwzględnionych i *$null$* dla wszystkich pozostałych, w tym reszt. Dzięki umożliwieniu tak bezpośredniej kontroli nad procesem oceniania modele typu Decision List doskonale nadają się do generowania list mailingowych i są szeroko stosowane w zarządzaniu relacjami z klientami, w tym w telefonicznych centrach obsługi i zastosowaniach marketingowych.

Zadania eksploracji, miary i wybory

Procesem modelowania steruje się za pośrednictwem **zadań eksploracji**. Każde zadanie eksploracji zasadniczo inicjuje nowy przebieg modelowania i zwraca nowy zestaw alternatywnych modeli do wyboru. Domyślne zadanie oparte jest na początkowej specyfikacji określonej przez użytkownika w węźle Decision List, jednak można zdefiniować dowolną liczbę innych zadań. Zadania można też uruchamiać iteracyjnie, np. uruchomić wyszukiwanie wysokiego prawdopodobieństwa na całym zbiorze uczącym, a następnie uruchomić wyszukiwanie niskiego prawdopodobieństwa na pozostałości, aby wyeliminować słabe segmenty.

Wybory danych

Można definiować wybory danych i własne miary modelu na potrzeby budowania i oceny jakości modelu. Na przykład w zadaniu eksploracji można określić wybór danych, aby dopasować model do konkretnego regionu, i utworzyć własną miarę, aby ocenić działanie tego modelu na danych z całego kraju. W odróżnieniu od zadań eksploracji miary nie zmieniają bazowego modelu, lecz pozwalają w inny sposób spojrzeć na jakość jego działania.

Wykorzystanie wiedzy biznesowej użytkownika

Decision List Viewer umożliwia użytkownikowi wykorzystanie wiedzy biznesowej bezpośrednio w modelu poprzez optymalizację lub uzupełnianie segmentów ujawnionych przez algorytm. Można edytować segmenty wygenerowane przez model lub dodawać nowe segmenty bazujące na regułach określonych przez użytkownika. Następnie można zastosować te zmiany i wyświetlić podgląd wyników.

Gdy potrzebne są pogłębione analizy, dynamiczne powiązanie z programem Excel umożliwia wyeksportowanie danych do tego programu i wykorzystanie ich do utworzenia prezentacji, obliczenia własnych miar, np. złożonych miar zysku i zwrotu z inwestycji. Decision List Viewer może wyświetlać te miary użytkownika w trakcie budowania modelu.

Przykład. Dział marketingowy instytucji finansowej chce przyczynić się do zwiększenia jej zysków, organizując kampanię, w której oferty będą precyzyjnie dopasowane do charakterystyki poszczególnych klientów. W takim scenariuszu można wykorzystać model Lista decyzyjna, aby na podstawie poprzednich promocji określić cechy klientów, którzy z największym prawdopodobieństwem pozytywnie zareagują na ofertę, i wygenerować listę mailingową na podstawie wyników.

Wymagania. Potrzebna jest jedna jakościowa zmienna przewidywana z typem pomiaru *Flaga* lub *Nominalna*, która będzie oznaczała binarny przewidywany wynik (tak/nie), oraz co najmniej jedna zmienna wejściowa. Gdy zmienna przewidywana jest *Nominalna*, należy ręcznie wybrać jedną wartość, która będzie oznaczała **trafienie** lub **reakcję**; wszystkie pozostałe wartości będą traktowane jako **brak trafienia**. Opcjonalnie można też określić zmienną częstości. Ilościowe zmienne daty/czasu są ignorowane. Wejściowe ilościowe zmienne liczbowe są automatycznie kategoryzowane przez algorytm w sposób określony na karcie Zaawansowane węzła modelowania. Aby móc bardziej precyzyjnie sterować kategoryzacją, należy przed węzłem modelowania dodać węzeł kategoryzacji i wykorzystać zmienną kategoryzowaną jako predyktor o poziomie pomiaru *Porządkowa*.

Opcje modelu Lista decyzyjna

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Utwórz modele rozdzielone. Tworzy osobny model dla każdej możliwej wartości zmiennych wejściowych, jakie są określone jako zmienne podziału. Więcej informacji można znaleźć w “Budowanie modeli rozdzielonych” na stronie 28.

Dominanta. Określa metodę używaną do budowy modelu.

- **Model.** Automatycznie generuje model na palecie modeli podczas wykonywania węzła. Wynikowy model można dodawać do strumieni w celu przeprowadzania oceny, ale nie można go już dalej edytować.
- **Drzewo interakcyjne.** otwiera interaktywne okno modelowania (wyników) programu Decision List Viewer, w którym można wybierać różne alternatywy i wielokrotnie stosować algorytm z różnymi ustawieniami, aby stopniowo rozbudowywać lub modyfikować model. Więcej informacji można znaleźć w temacie “Decision List Viewer” na stronie 155.
- **Użyj informacji o zapisanej sesji interaktywnej.** Uruchamia sesję interaktywną z zapisanymi wcześniej ustawieniami. Ustawienia interaktywne można zapisać z programu Decision List Viewer, używając menu Utwórz (w celu utworzenia modelu lub węzła modelowania) lub z menu Plik (w celu zaktualizowania węzła, z którego sesja została uruchomiona).

Wartość przewidywana. Określa wartość zmiennej przewidywanej, która oznacza wynik modelowania. Na przykład, jeśli zmienna przewidywana churn (odejście) jest zakodowana tak, że 0 = nie i 1 = tak, należy określić 1, aby wskazać reguły zwracające rekordy obarczone prawdopodobieństwem odejścia.

Wyszukaj segmenty z. Wskazuje, czy podczas wyszukiwania zmiennej przewidywanej należy szukać wystąpień z **Wysokim prawdopodobieństwem**, czy z **Niskim prawdopodobieństwem**. Znalezienie i wykluczenie takich wystąpień może pomóc w udoskonaleniu modelu, zwłaszcza jeśli pozostałe wystąpienia mają niskie prawdopodobieństwo.

Maksymalna liczba segmentów. Określa maksymalną liczbę segmentów, jaka ma być zwracana. Tworzonych jest *N* najlepszych segmentów, przy czym najlepszy segment to taki, który ma najwyższe prawdopodobieństwo lub, jeśli więcej niż jeden model ma takie samo prawdopodobieństwo, taki, który ma największe pokrycie. Minimalne dozwolone ustawienie to 1; nie ma ustawienia maksymalnego.

Minimalna wielkość segmentu. Dwa poniższe ustawienia określają minimalną wielkość segmentu. Priorytet ma większa z dwóch wartości. Na przykład, jeśli z wartości procentowej wynika liczba większa od wartości bezwzględnej, to obowiązuje wartość procentowa.

- **Jako procent poprzedniego segmentu (%).** Określa minimalną wielkość grupy jako odsetek rekordów. Minimalne dozwolone ustawienie to 0; maksymalne dozwolone ustawienie to 99,9.
- **Jako wartość bezwzględna (N).** Określa minimalną wielkość grupy jako bezwzględną liczbę rekordów. Minimalne dozwolone ustawienie to 1; nie ma ustawienia maksymalnego.

Reguły segmentacyjne.

Maksymalna liczba atrybutów. Określa maksymalną liczbę warunków na jedną regułę segmentacyjną. Minimalne dozwolone ustawienie to 1; nie ma ustawienia maksymalnego.

- **Ponowne użycie atrybutu.** Gdy ta opcja jest włączona, w każdym cyklu mogą być brane pod uwagę wszystkie atrybuty, nawet te, które zostały już wykorzystane w poprzednich cyklach. Warunki dla segmentu budowane są w cyklach, a w każdym cyklu dodawany jest nowy warunek. Liczbę cykli definiuje się za pośrednictwem ustawienia **Maksymalna liczba atrybutów**.

Przedział ufności dla nowych kryteriów (%). Określa poziom ufności używany do testowania istotności segmentów. To ustawienie ma istotny wpływ na liczbę zwracanych segmentów (o ile w ogóle są zwracane) oraz liczbę warunków przypadającą na jedną regułę segmentacyjną. Im wyższa wartość, tym mniejszy zwracany zestaw wynikowy. Minimalne dozwolone ustawienie to 50; maksymalne dozwolone ustawienie to 99,9.

Opcje zaawansowane węzła Lista decyzyjna

Opcje zaawansowane umożliwiają precyzyjne dostosowanie procesu budowania modelu.

Metoda kategoryzacji. Metoda używana do kategoryzacji zmiennych ilościowych (równa liczebność lub równa szerokość).

Liczba przedziałów. Liczba przedziałów (kategorii), jaka ma być tworzona dla zmiennych ilościowych. Minimalne dozwolone ustawienie to 2; nie ma ustawienia maksymalnego.

Liczba wyników modelu. Maksymalna liczba wyników modelu z jednego cyklu, jaka może być wykorzystana w następnym cyklu. Minimalne dozwolone ustawienie to 1; nie ma ustawienia maksymalnego.

Wynikowa szerokość reguły. Maksymalna liczba wyników reguły na jeden cykl, jaka może być użyta w następnym cyklu. Minimalne dozwolone ustawienie to 1; nie ma ustawienia maksymalnego.

Czynnik łączenia w kategorie. Minimalna wielkość, o jaką segment musi powiększyć się w wyniku scalenia z sąsiadem. Minimalne dozwolone ustawienie to 1,01; nie ma ustawienia maksymalnego.

- **Zezwól na brakujące wartości w warunkach.** Prawda zezwala na test IS MISSING w regułach.
- **Odrzuć wyniki pośrednie.** Ustawienie Prawda powoduje, że zwracane są tylko ostateczne wyniki procesu wyszukiwania. Ostateczny wynik to wynik, który nie zostałby już w żaden sposób udoskonalony w dalszym procesie wyszukiwania. Ustawienie Fałsz powoduje, że zwracane są także wyniki pośrednie.

Maksymalna liczba alternatyw. Określa maksymalną liczbę alternatyw, która może być zwrócona po uruchomieniu zadania eksploracji. Minimalne dozwolone ustawienie to 1; nie ma ustawienia maksymalnego.

Należy zwrócić uwagę, że zadanie eksploracji zwróci tylko rzeczywiście dostępną liczbę alternatyw, ale nie więcej od określonego maksimum. Na przykład, jeśli maksimum ustawione jest na 100, a znalezione zostaną tylko 3 alternatywy, to zwrócone zostaną tylko te 3 alternatywy.

Model użytkowy Lista decyzyjna

Model składa się z listy **segmentów**, a każdy segment jest zdefiniowany przez **regułę**, która wybiera dopasowane rekordy. Można w prosty sposób przeglądać lub modyfikować segmenty przed wygenerowaniem modelu i wybrać segmenty do uwzględnienia lub wykluczenia. Modele Lista decyzyjna używane do oceny zwracają *Tak* dla segmentów uwzględnionych i *$null$* dla wszystkich pozostałych, w tym reszt. Dzięki umożliwieniu tak bezpośredniej kontroli nad procesem oceniania modele list decyzyjnych doskonale nadają się do generowania list mailingowych i są szeroko stosowane w zarządzaniu relacjami z klientami, w tym w telefonicznych centrach obsługi i zastosowaniach marketingowych.

Po uruchomieniu strumienia zawierającego model Lista decyzyjna węzeł dodaje trzy nowe zmienne zawierające ocenę: albo *1* (czyli *Tak*) dla zmiennych uwzględnionych, albo *$null$* dla zmiennych wykluczonych, prawdopodobieństwo (współczynnik trafień) dla segmentu, do którego należy rekord, oraz numer identyfikacyjny segmentu. Nazwy nowych zmiennych są tworzone na podstawie nazwy zmiennej przewidywanej, do której dodawany jest przedrostek *$D-* w przypadku oceny, *$DP-* w przypadku prawdopodobieństwa i *$DI-* w przypadku identyfikatora segmentu.

Model jest oceniany na podstawie wartości przewidywanej określonej w momencie budowania modelu. Można ręcznie wykluczać segmenty, tak aby były oceniane jako *$null$*. Na przykład, jeśli uruchomimy wyszukiwanie rekordów o niskim prawdopodobieństwie, znalezione „niskie" segmenty otrzymają ocenę *Tak*, jeśli ich ręcznie nie wykluczymy. W razie potrzeby oceny null można przekodować na *Nie* za pomocą węzła wyliczania lub wypełniania.

PMML

Model Lista decyzyjna można także zapisać jako model zestawu reguł w języku PMML z kryterium wyboru „pierwszego trafienia". Jednak wszystkie reguły mają tę samą oczekiwaną ocenę. Aby umożliwić zmiany zmiennej przewidywanej lub wartości przewidywanej, można w jednym pliku zapisać wiele modeli zestawów reguł i stosować je po kolei, przy czym obserwacje niedopasowane przez pierwszy model będą przekazywane do następnego i tak dalej. Nazwa algorytmu *DecisionList* sygnalizuje to niestandardowe zachowanie i tylko modele zestawów reguł o tej nazwie są rozpoznawane jako modele Lista decyzyjna i tak oceniane.

Ustawienia modelu użytkowego Lista decyzyjna

Karta Ustawienia modelu Lista decyzyjna umożliwia uzyskiwanie ocen skłonności i włączenie lub wyłączenie optymalizacji kodu SQL. Karta jest dostępna tylko po dodaniu modelu użytkowego do strumienia.

Wylicz surowe oceny skłonności. W przypadku modeli z przewidywaną zmienną typu flaga (zwracających predykcje tak lub nie) można wyliczyć oceny skłonności definiujące prawdopodobieństwo prawdziwego wyniku określonego dla zmiennej przewidywanej. Stanowią one uzupełnienie pozostałych współczynników ufności i wartości predykcyjnych, jakie mogą zostać wygenerowane podczas oceniania.

Wylicz skorygowane oceny skłonności. Surowe oceny skłonności bazują na danych uczących i mogą być zbyt optymistyczne z uwagi na tendencję wielu modeli do przeuczania tych danych. Opcja wyliczania skłonności skorygowanych próbuje skompensować tę tendencję na drodze oceny wydajności modelu w teście lub w podzbiorze walidacyjnym. Opcja ta wymaga, aby zmienna dzieląca na podzbiory była zdefiniowana w strumieniu oraz aby skorygowane oceny skłonności były włączone w węźle modelowania przed przystąpieniem do generowania modelu.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę, wykorzystując natywny kod SQL** Jeśli ta opcja jest wybrana, generowany jest natywny kod SQL w celu oceny modelu w bazie danych.

  Uwaga: Ta opcja może szybciej zwracać wyniki, ale rozmiar i złożoność natywnego kodu SQL wzrastają wraz ze wzrostem złożoności modelu.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Decision List Viewer

Łatwy w obsłudze i zorientowany zadaniowo interfejs graficzny programu Decision List Viewer ułatwia budowanie modelu, separując użytkownika od niskopoziomowych szczegółów techniki eksploracji danych, a pozwalając poświęcić całą uwagę na elementy analizy wymagające interwencji użytkownika, takie jak ustalanie celów, wybór grup docelowych, analizowanie wyników i wybór optymalnego modelu.

Panel Model roboczy

W panelu Model roboczy wyświetlany jest bieżący model, w tym zadania eksploracji i inne działania mające zastosowanie do modelu roboczego

Identyfikator. Numer kolejny segmentu. Segmenty modelu są obliczane w kolejności wynikającej z identyfikatorów.

Reguły segmentacyjne. Nazwa segmentu i zdefiniowane warunki segmentu. Domyślnie nazwa segmentu jest identyczna z nazwą zmiennej lub składa się z szeregu nazw zmiennych używanych w warunkach, rozdzielonych przecinkami.

Ocena. Określa zmienną, która ma być przewidywana i której wartość z założenia jest powiązana z wartościami pozostałych zmiennych (predyktorów).

Uwaga: Wyświetlanie pozostałych opcji można włączać i wyłączać w oknie dialogowym “Organizacja miar modelu” na stronie 165.

Pokrycie. Wykres kołowy przedstawia w formie wizualnej pokrycie poszczególnych segmentów względem całego pokrycia.

Pokrycie (n). Lista pokryć poszczególnych segmentów względem całego pokrycia.

Częstość. Lista liczb trafień względem pokrycia. Na przykład, gdy pokrycie wynosi 79, a częstość wynosi 50, to w wybranym segmencie mamy 50 trafień na 79 klientów.

Prawdopodobieństwo. Określa prawdopodobieństwo segmentu. Na przykład, gdy pokrycie wynosi 79, a częstość wynosi 50, to prawdopodobieństwo segmentu wynosi 63,29% (50 podzielone na 79).

Błąd. Określa błąd segmentu.

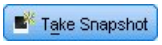
Na dole panelu podane jest pokrycie, częstość i prawdopodobieństwo całego modelu.

Pasek narzędzi modelu roboczego

W panelu Model roboczy dostępny jest pasek narzędzi z funkcjami opisanymi poniżej.

Uwaga: Niektóre funkcje można też wywołać, klikając segment modelu prawym przyciskiem myszy.

Tabela 9. Przyciski na pasku narzędzi panelu Model roboczy

Przycisk na pasku narzędzi	Opis
	Otwiera okno dialogowe Generuj nowy model zawierające opcje służące do tworzenia nowego modelu użytkowego.
	Zapisuje bieżący stan sesji interaktywnej. Węzeł modelowania Lista decyzyjna jest aktualizowany z uwzględnieniem bieżących ustawień, w tym zadań eksploracji, obrazów stanu modelu, wyborów danych i miar użytkownika. Aby przywrócić sesję do zapisanego stanu, zaznacz pole wyboru **Użyj informacji o zapisanej sesji interaktywnej** na karcie Model węzła modelowania i kliknij przycisk **Uruchom**.
	Wyświetla okno dialogowe Organizuj miary modelu. Więcej informacji można znaleźć w temacie “Organizacja miar modelu” na stronie 165.
	Wyświetla okno dialogowe Przygotuj wybory danych. Więcej informacji można znaleźć w temacie “Przygotowywanie wyborów danych” na stronie 160.
	Wyświetla kartę Obrazy stanu. Więcej informacji można znaleźć w temacie “Karta Obrazy stanu” na stronie 157.
	Wyświetla kartę Alternatywne modele. Więcej informacji można znaleźć w temacie “Karta Alternatywne modele” na stronie 157.
	Zapisuje bieżący stan struktury modelu („obraz stanu”). Obrazy stanu widoczne są na karcie Obrazy stanu i często używane są do porównywania modelu.
	Otwiera okno dialogowe Wstaw segment zawierające opcje służące do tworzenia nowych segmentów modelu.
	Otwiera okno dialogowe Reguły segmentacyjne z opcjami umożliwiającymi dodawanie warunków do segmentów modelu i zmianę wcześniej zdefiniowanych warunków segmentów modelu.
	Przenosi zaznaczony segment na wyższą pozycję w hierarchii modelu.
	Przenosi zaznaczony segment na niższą pozycję w hierarchii modelu.
	Usuwa zaznaczony segment.
	Przełącza stan uwzględnienia zaznaczonego segmentu w modelu. Gdy segment nie jest uwzględniony (tj. jest wykluczony), jego wyniki są dodawane do pozostałości. Operacja ta nie jest równoznaczna z usunięciem segmentu, ponieważ pozostawia otwartą drogę do jego reaktywacji.

Karta Alternatywne modele

Po kliknięciu opcji **Znajdź segmenty** na karcie Modele alternatywne panelu modelu roboczego pojawiają się alternatywne wyniki eksploracji wybranego modelu lub segmentu.

Aby jeden z modeli alternatywnych stał się modelem roboczym, zaznacz ten model alternatywny i kliknij przycisk **Wczytaj**; model alternatywny pojawi się w panelu Model roboczy.

Uwaga: Karta Modele alternatywne jest wyświetlana tylko wtedy, gdy określono wartość **Maksymalna liczba alternatyw** na karcie Zaawansowane węzła modelowania Lista decyzyjna w celu utworzenia więcej niż jednego modelu alternatywnego.

Dla każdego utworzonego modelu alternatywnego wyświetlane są określone informacje:

Nazwa. Każdy model alternatywny ma numer kolejny. Pierwszy model alternatywny zwykle zawiera najlepsze wyniki.

Zmienna przewidywana. Określa wartość przewidywaną. Na przykład 1, co oznacza „prawda".

Liczba segmentów. Liczba reguł segmentacyjnych użytych w modelu alternatywnym.

Pokrycie. Pokrycie modelu alternatywnego.

Częstość. Lista liczb trafień względem pokrycia.

Prawdopodobieństwo. Procentowe prawdopodobieństwo modelu alternatywnego.

Uwaga: Wyniki alternatywne nie są zapisywane z modelem; wyniki są ważne tylko w trakcie aktywnej sesji.

Karta Obrazy stanu

Obraz stanu jest zapisem stanu modelu w konkretnym momencie. Możemy na przykład utworzyć obraz stanu modelu, gdy chcemy wczytać inny model alternatywny do panelu Model roboczy, ale nie chcemy tracić wyników pracy z bieżącym modelem. Karta Obrazy stanu zawiera listę wszystkich obrazów stanu utworzonych ręcznie i będących zapisami dowolnej liczby stanów modelu roboczego.

Uwaga: Obrazy stanu są zapisywane razem z modelem. Zaleca się utworzenie obrazu stanu po wczytaniu pierwszego modelu. Obraz stanu zachowa oryginalną strukturę modelu, tak aby zawsze można było wrócić do jego pierwotnego stanu. Wygenerowana nazwa obrazu stanu ma postać znacznika czasu utworzenia.

Tworzenie obrazu stanu modelu

1. Wybierz odpowiedni model/model alternatywny, który ma być wyświetlany w panelu Model roboczy.
2. Dokonaj wszelkich niezbędnych zmian w modelu roboczym.
3. Kliknij przycisk **Wykonaj obraz stanu**. Nowy obraz stanu pojawi się na karcie Obrazy stanu.

   Nazwa. Nazwa obrazu stanu. Można zmienić nazwę obrazu stanu, klikając ją dwukrotnie.

   Zmienna przewidywana. Określa wartość przewidywaną. Na przykład 1, co oznacza „prawda".

   Liczba segmentów. Liczba reguł segmentacyjnych użytych w modelu.

   Pokrycie. Pokrycie modelu.

   Częstość. Lista liczb trafień względem pokrycia.

   Prawdopodobieństwo. Procentowe prawdopodobieństwo modelu.
4. Aby jeden z obrazów stanu stał się modelem roboczym, zaznacz ten obraz stanu i kliknij przycisk **Wczytaj**; obraz stanu pojawi się w panelu Model roboczy.
5. Można usunąć obraz stanu, klikając przycisk **Usuń** lub klikając obraz stanu prawym przyciskiem myszy i wybierając z menu polecenie **Usuń**.

Praca z Decision List Viewer

Budowa modelu, który najtrafniej przewidzi reakcje i zachowania klientów, jest procesem wieloetapowym. Po uruchomieniu przeglądarki list decyzyjnych Decision List Viewer model roboczy wypełniany jest zdefiniowanymi segmentami i miarami, po czym staje się gotowy do uruchomienia zadania eksploracji, modyfikowania segmentów/miar i tworzenia nowego modelu lub węzła modelowania.

Można dodawać reguły segmentacyjne, aż do uzyskania zadowalającego modelu. Reguły segmentacyjne można dodawać do modelu, uruchamiając zadania eksploracji lub korzystając z funkcji **Edytuj regułę segmentacyjną**.

W procesie budowania modelu można oceniać jego działanie, walidując model względem danych pomiarowych, wizualizując go na wykresie lub generując niestandardowe miary w programie Excel.

Po uzyskaniu pewności co do jakości modelu można wygenerować nowy model i umieścić go w obszarze roboczym programu IBM SPSS Modeler lub na palecie Model.

Zadania eksploracji

Zadanie eksploracji jest to zbiór parametrów określający sposób generowania nowych reguł. Niektóre z tych parametrów można wybierać, elastycznie adaptując modele do nowej sytuacji. Zadanie składa się z szablonu (typu) zadania, zmiennej przewidywanej i zbioru danych eksploracji.

W kolejnych sekcjach omówiono różne operacje związane z zadaniami eksploracji

- "Uruchamianie zadań eksploracji"
- "Tworzenie i edycja zadania eksploracji" na stronie 159
- "Przygotowywanie wyborów danych" na stronie 160

Uruchamianie zadań eksploracji: Decision List Viewer umożliwia ręczne dodawanie reguł segmentacyjnych do modelu poprzez uruchamianie zadań eksploracji lub kopiowanie i wklejanie reguł segmentacyjnych między modelami. Zadanie eksploracji zawiera informacje o sposobie generowania nowych reguł segmentacji (parametry eksploracji danych, takie jak strategia wyszukiwania, atrybuty źródłowe, liczba wyników, przedział ufności itd.), zachowania klientów, które chcemy przewidywać, i dane podlegające analizie. Celem zadania eksploracji jest znalezienie jak najlepszych reguł segmentacyjnych.

Aby wygenerować regułę segmentacyjną modelu poprzez uruchomienie zadania eksploracji:

1. Kliknij wiersz **Pozostałość**. Jeśli w panelu Model roboczy są już wyświetlone segmenty, można także wybrać jeden z segmentów, aby znaleźć dodatkowe reguły na jego podstawie. Po wybraniu pozostałości lub segmentu użyj jednej z poniższych metod, aby wygenerować model lub modele alternatywne:

- Z menu Narzędzia wybierz opcję **Znajdź segmenty**.
- Prawym przyciskiem myszy kliknij wiersz **Pozostałość** lub segment i wybierz polecenie **Znajdź segmenty**.
- Kliknij przycisk **Znajdź segmenty** w panelu Model roboczy.

Podczas wykonywania zadania u dołu obszaru roboczego wyświetlany jest wskaźnik postępu, a po zakończeniu zadania pojawia się stosowna informacja. Czas wykonywania zadania eksploracji zależy od jego złożoności i wielkości zbioru danych. Jeśli wyniki zawierają tylko jeden model, to pojawi się on w panelu Model roboczy od razu po zakończeniu wykonywania zadania; jeśli jednak wyniki zawierają więcej niż jeden model, to są one wyświetlane na karcie Alternatywne modele.

Uwaga: Wynikiem zadania będzie utworzenie modeli, nieutworzenie modeli albo niepowodzenie.

Proces znajdowania nowych reguł segmentacyjnych można powtarzać do czasu, aż do modelu nie będą już dodawane żadne nowe reguły. Będzie to oznaczać, że znaleziono wszystkie istotne grupy klientów.

Zadanie eksploracji można uruchomić na dowolnym istniejącym segmencie modelu. Jeśli wynik zadania nie jest zgodny z oczekiwaniami, można uruchomić kolejne zadanie eksploracji na tym samym segmencie. Spowoduje to

znalezienie dodatkowych reguł na podstawie wybranego segmentu. Segmenty znajdujące się „poniżej" wybranego (tj. dodane do modelu później niż on) zostaną zastąpione przez nowe segmenty, ponieważ każdy segment zależy od swoich poprzedników.

Tworzenie i edycja zadania eksploracji: Zadanie eksploracji to mechanizm wyszukiwania zbioru reguł składających się na model danych. Obok kryteriów wyszukiwania zdefiniowanych w wybranym szablonie zadanie definiuje także zmienną przewidywaną (właściwe pytanie stanowiące powód analizy, np. ilu klientów prawdopodobnie odpowie na mailing) oraz zbiory danych, które mają być używane. Celem zadania eksploracji jest znalezienie jak najlepszych modeli.

Tworzenie zadania eksploracji

Aby utworzyć zadanie eksploracji:

1. Wybierz segment, z którego chcesz eksplorować dodatkowe warunki.
2. Kliknij przycisk **Ustawienia**. Zostanie otwarte okno dialogowe tworzenia/edycji zadania eksploracji. Okno to zawiera opcje służące do definiowania zadania eksploracji.
3. Wprowadź potrzebne zmiany i kliknij przycisk **OK**, aby wrócić do panelu Model roboczy. Decision List Viewer użyje tych ustawień jako domyślnych przy uruchamianiu każdego zadania, dopóki nie zostaną wybrane ustawienia alternatywne.
4. Kliknij opcję **Znajdź segmenty**, aby uruchomić zadanie eksploracji na wybranym segmencie.

Edytowanie zadania eksploracji

Okno dialogowe tworzenia/edycji zadania eksploracji zawiera opcje służące do definiowania nowego zadania eksploracji lub edytowania istniejącego.

Większość parametrów zadań eksploracji jest podobna do parametrów węzła Lista decyzyjna. Różnice omówiono poniżej. Więcej informacji można znaleźć w temacie "Opcje modelu Lista decyzyjna" na stronie 152.

Wczytaj ustawienia: Po utworzeniu więcej niż jednego zadania eksploracji, wybierz żądane zadanie.

Nowe... Kliknij, aby utworzyć nowe zadanie eksploracji na podstawie ustawień obecnie wyświetlanego zadania.

Zmienna przewidywana

Zmienne przewidywane: Zmienna przewidywana, której wartość z założenia jest powiązana z wartościami pozostałych zmiennych (predyktorów).

Wartość przewidywana. Określa wartość zmiennej przewidywanej, która oznacza wynik modelowania. Na przykład, jeśli zmienna przewidywana churn (odejście) jest zakodowana tak, że 0 = nie i 1 = tak, należy określić 1, aby wskazać reguły zwracające rekordy obarczone prawdopodobieństwem odejścia.

Ustawienia proste

Maksymalna liczba alternatyw. Określa maksymalną liczbę modeli alternatywnych, która zostanie wyświetlona po wykonaniu zadania eksploracji. Minimalne dozwolone ustawienie to 1; nie ma ustawienia maksymalnego.

Ustawienia zaawansowane

Edytuj... Otwiera okno dialogowe edycji parametrów zaawansowanych, w którym można określić zaawansowane ustawienia. Więcej informacji można znaleźć w temacie "Edytowanie parametrów zaawansowanych" na stronie 160.

Dane

Utwórz wybór. Decision List Viewer będzie analizować określoną tutaj miarę ewaluacyjną w celu znalezienia nowych reguł. Wymienione miary ewaluacyjne tworzy/edytuje się w oknie dialogowym Przygotuj wybory danych.

Dostępne zmienne. Udostępnia opcje wyświetlania wszystkich zmiennych lub ręcznego wyboru zmiennych do wyświetlanie.

Edytuj... Jeśli wybrana jest opcja **Użytkownika**, powoduje otwarcie okna dialogowego **Zmodyfikuj dostępne zmienne** służącego do wyboru zmiennych, które będą dostępne jako atrybuty segmentów znajdowanych przez zadanie eksploracji. Więcej informacji można znaleźć w temacie "Dostosowywanie dostępnych zmiennych".

Edytowanie parametrów zaawansowanych: Okno dialogowe edycji parametrów zaawansowanych udostępnia następujące opcje konfiguracji.

Metoda kategoryzacji. Metoda używana do kategoryzacji zmiennych ilościowych (równa liczebność lub równa szerokość).

Liczba przedziałów. Liczba przedziałów (kategorii), jaka ma być tworzona dla zmiennych ilościowych. Minimalne dozwolone ustawienie to 2; nie ma ustawienia maksymalnego.

Liczba wyników modelu. Maksymalna liczba wyników modelu z jednego cyklu, jaka może być wykorzystana w następnym cyklu. Minimalne dozwolone ustawienie to 1; nie ma ustawienia maksymalnego.

Wynikowa szerokość reguły. Maksymalna liczba wyników reguły na jeden cykl, jaka może być użyta w następnym cyklu. Minimalne dozwolone ustawienie to 1; nie ma ustawienia maksymalnego.

Czynnik łączenia w kategorie. Minimalna wielkość, o jaką segment musi powiększyć się w wyniku scalenia z sąsiadem. Minimalne dozwolone ustawienie to 1,01; nie ma ustawienia maksymalnego.

- **Zezwól na brakujące wartości w warunkach.** Prawda zezwala na test IS MISSING w regułach.
- **Odrzuć wyniki pośrednie.** Ustawienie Prawda powoduje, że zwracane są tylko ostateczne wyniki procesu wyszukiwania. Ostateczny wynik to wynik, który nie zostałby już w żaden sposób udoskonalony w dalszym procesie wyszukiwania. Ustawienie Fałsz powoduje, że zwracane są także wyniki pośrednie.

Dostosowywanie dostępnych zmiennych: Okno dialogowe dostosowywania dostępnych zmiennych umożliwia wybranie zmiennych, które będą dostępne jako atrybuty segmentów znajdowanych przez zadanie eksploracji.

Dostępne. Lista zmiennych, które obecnie są dostępne jako atrybuty segmentów. Aby usunąć zmienne z listy, zaznacz je i kliknij przycisk **Usuń>>**. Zaznaczone zmienne zostaną przeniesione z listy dostępnych na listę niedostępnych.

Niedostępne. Lista zmiennych, które obecnie nie są dostępne jako atrybuty segmentów. Aby umieścić zmienne na liście dostępnych, zaznacza je i kliknij przycisk **<<Dodaj**. Zaznaczone zmienne zostaną przeniesione z listy niedostępnych na listę dostępnych.

Przygotowywanie wyborów danych: Przygotowując odpowiednio wybory danych (zbiór danych eksploracji), można określić miary ewaluacyjne, które Decision List Viewer będzie analizować w celu znalezienia nowych reguł, i wybrać dane, które będą podstawą dla miar.

Aby przygotować wybory danych:

1. Z menu Narzędzia wybierz opcję **Przygotuj wybory danych** lub kliknij prawym przyciskiem myszy i wybierz tę samą opcję. Zostanie otwarte okno dialogowe Przygotuj wybory danych.

   Uwaga: Okno dialogowe Przygotuj wybory danych umożliwia także edytowanie lub usuwanie istniejących wyborów danych.
2. Kliknij przycisk **Dodaj nowy wybór danych**. Do istniejącej tabeli zostanie dodany nowy wybór danych.
3. Kliknij opcję **Nazwa** i wprowadź nazwę wyboru.
4. Kliknij opcję **Partycja** i wybierz typ podzbioru.

5. Kliknij opcję **Warunek** i wybierz odpowiednią opcję warunku. Wybranie opcji **Określ** spowoduje otwarcie okna dialogowe Określ warunek wyboru z opcjami służącymi do definiowania warunków dla zmiennych.
6. Zdefiniuj odpowiedni warunek i kliknij przycisk **OK**.

Wybory danych są dostępne na liście rozwijanej Utwórz wybór w oknie dialogowym tworzenia/edycji zadania eksploracji. Lista ta umożliwia wybranie miary ewaluacyjnej, która ma być używana w konkretnym zadaniu eksploracji.

Reguły segmentacyjne

Aby znaleźć reguły segmentacyjne, należy uruchomić zadanie eksploracji na podstawie szablonu zadania. Można ręcznie dodawać reguły segmentacyjne do modelu, korzystając z funkcji Wstaw segment lub Edytuj regułę segmentacyjną.

Ewentualne wyniki eksploracji w poszukiwaniu nowych reguł segmentacyjnych wyświetlane są na karcie Przegląd w oknie dialogowym Lista interaktywna. Można szybko dopracować model, wybierając jeden z wyników alternatywnych w oknie dialogowym albumu z modelami i klikając przycisk **Wczytaj**. W ten sposób można eksperymentować z różnymi wynikami, aż do zbudowania modelu, który będzie dokładnie opisywał optymalną grupę przewidywaną.

Wstawianie segmentów: Można ręcznie dodawać reguły segmentacyjne do modelu, korzystając z funkcji Wstaw segment.

Aby dodać warunek reguły segmentacyjnej do modelu:

1. W oknie dialogowym Lista interaktywna wybierz miejsce, w którym chcesz dodać nowy segment. Nowy segment zostanie wstawiony bezpośrednio nad wybranym.
2. W menu Edycja wybierz polecenie **Wstaw segment** lub wywołaj tę opcję, klikając segment prawym przyciskiem myszy.

   Zostanie otwarte okno wstawiania segmentu, w którym można wstawiać nowe warunki reguł segmentacyjnych.
3. Kliknij przycisk **Wstaw**. Zostanie otwarte okno wstawiania warunku, w którym można zdefiniować atrybuty dla nowego warunku reguły.
4. Wybierz zmienną i operator z list rozwijanych.

   Uwaga: W wypadku wybrania operatora **Poza** wybrany warunek będzie działał jako warunek wykluczający i w oknie dialogowym wstawiania reguły będzie wyświetlany w kolorze czerwonym. Na przykład, gdy warunek region = 'TOWN' wyświetlany jest w kolorze czerwonym, wartość TOWN jest wykluczona z zestawu wyników.
5. Wprowadź jedną lub wiele wartości lub kliknij ikonę **Wstaw wartość**, aby wyświetlić okno dialogowe wstawiania wartości. Okno to umożliwia wybranie wartości zdefiniowanej dla wybranej zmiennej. Na przykład zmienna **married** (w związku małżeńskim) może przyjmować wartości **yes** (tak) i **no** (nie).
6. Kliknij przycisk **OK**, aby wrócić do okna dialogowego wstawiania segmentu. Ponownie kliknij przycisk **OK**, aby dodać utworzony segment do modelu.

Nowy segment pojawi się w określonym miejscu w modelu.

Edytowanie reguł segmentacyjnych: Funkcja edytowanie reguł segmentacyjnych umożliwia dodawanie, zmienianie i usuwanie warunków reguł segmentacyjnych.

Aby zmienić warunek reguły segmentacyjnej:

1. Zaznacz segment modelu, który chcesz edytować.
2. Z menu Edycja wybierz polecenie **Edytuj regułę segmentacyjną** lub kliknij regułę prawym przyciskiem myszy, aby wywołać te opcję.

   Zostanie otwarte okno dialogowe edycji reguły segmentacyjnej.
3. Wybierz odpowiedni warunek i kliknij przycisk **Edytuj**.

   Zostanie otwarte okno edytowania warunku, w którym można zdefiniować atrybuty dla wybranego warunku reguły.

4. Wybierz zmienną i operator z list rozwijanych.

   Uwaga: W wypadku wybrania operatora **Poza** wybrany warunek będzie działał jako warunek wykluczający i w oknie dialogowym edytowania reguły będzie wyświetlany w kolorze czerwonym. Na przykład, gdy warunek region = 'TOWN' wyświetlany jest w kolorze czerwonym, wartość TOWN jest wykluczona z zestawu wyników.
5. Wprowadź jedną lub wiele wartości lub kliknij przycisk **Wstaw wartość**, aby wyświetlić okno dialogowe wstawiania wartości. Okno to umożliwia wybranie wartości zdefiniowanej dla wybranej zmiennej. Na przykład zmienna **married** (w związku małżeńskim) może przyjmować wartości **yes** (tak) i **no** (nie).
6. Kliknij przycisk **OK**, aby wrócić do okna dialogowego edytowania reguły segmentacyjnej. Ponownie kliknij przycisk **OK**, aby wrócić do modelu roboczego.

Wybrany segment zostanie wyświetlony ze zmienionymi warunkami reguły.

Usuwanie warunków reguły segmentacyjnej: **Aby usunąć warunek reguły segmentacyjnej:**

1. Zaznacz segment modelu zawierający warunki reguł, które chcesz usunąć.
2. Z menu Edycja wybierz polecenie **Edytuj regułę segmentacyjną** lub kliknij segment prawym przyciskiem myszy, aby wywołać te opcję.

   Zostanie otwarte okno dialogowe edytowania reguły segmentacyjnej, które umożliwia usunięcie jednego lub wielu warunków reguły.
3. Wybierz odpowiedni warunek reguły i kliknij przycisk **Usuń**.
4. Kliknij przycisk **OK**.

Usunięcie jednego lub wielu warunków reguły segmentacyjnej powoduje odświeżenie metryk miar na panelu Model roboczy.

Kopiowanie segmentów: Decision List Viewer udostępnia wygodny sposób kopiowania segmentów. Aby zastosować segment z jednego modelu w innym, wystarczy skopiować (lub wyciąć) ten segment z jednego modelu i wkleić do drugiego. Można także skopiować segment z modelu wyświetlanego na panelu podglądu modeli alternatywnych i wkleić go do modelu wyświetlanego w panelu modelu roboczego. Przy wykonywaniu tych operacji wycinania, kopiowania i wklejania wykorzystywany jest schowek systemowy, w którym dane są tymczasowo umieszczane. Oznacza to, że schowek zawiera kopię warunków i zmiennej przewidywanej. Decision List Viewer nie ma wyłączności na dostęp do schowka — jego zawartość można też wklejać do innych aplikacji. Na przykład warunki i zmienna przewidywana po wklejaniu do edytora tekstu przyjmą postać zapisu XML.

Aby skopiować lub wyciąć segmenty modelu:

1. Zaznacz segment modelu, którego chcesz używać w innym modelu.
2. Z menu Edycja wybierz polecenie **Kopiuj** (lub **Wytnij**) albo kliknij segment modelu prawym przyciskiem myszy i wybierz polecenie **Kopiuj** lub **Wytnij**.
3. Otwórz model, do którego chcesz wkleić segment.
4. Zaznacz jeden z segmentów modelu i kliknij przycisk **Wklej**.

Uwaga: Zamiast poleceń **Wklej**, **Wytnij** i **Wklej** można także używać kombinacji klawiszy: **Ctrl+X**, **Ctrl+C** i **Ctrl+V**.

Skopiowany (lub wycięty) segment zostanie wstawiony nad uprzednio zaznaczonym segmentem modelu. Miary wklejonego segmentu i wszystkich segmentów pod nim zostaną przeliczone.

Uwaga: Oba modele w tej procedurze muszą bazować na tym samym szablonie i zawierać tę samą zmienną przewidywaną. W przeciwnym razie zostanie wyświetlony komunikat o błędzie.

Alternatywne modele: Gdy istnieje więcej niż jeden wynik, na karcie Alternatywne modele wyświetlane są wyniki poszczególnych zadań eksploracji. Każdy wynik zawiera te warunki dotyczące wybranych danych, które są najlepiej dopasowane do zmiennej przewidywanej, a także wszelkie „wystarczająco dobre" alternatywy. Łączna liczba wyświetlonych modeli alternatywnych zależy od kryterium wyszukiwania użytego w procesie analizy.

Aby przeglądać modele alternatywne:

1. Kliknij jeden z modeli alternatywnych na karcie Alternatywne modele. W panelu przeglądu modeli alternatywnych zostaną wyświetlone segmenty modelu alternatywnego lub zastąpią one segmenty bieżącego modelu.
2. Aby pracować z modelem alternatywnym w panelu modelu roboczego, zaznacz model i kliknij przycisk **Wczytaj** w panelu przeglądu modeli alternatywnych, ewentualnie kliknij nazwę modelu alternatywnego na karcie Alternatywne modele i wybierz polecenie **Wczytaj**.

Uwaga: Modele alternatywne nie są zapisywane po wygenerowaniu nowego modelu.

Dostosowywanie modelu

Dane z reguły nie są statyczne. Klienci przeprowadzają się, zawierają związki małżeńskie i zmieniają pracę. Produkty tracą zainteresowanie rynku i stają się przestarzałe.

Decision List Viewer umożliwia użytkownikom biznesowym elastyczne adaptowanie modeli do nowych sytuacji. Model można zmieniać, edytując, usuwając lub dezaktywując konkretne segmenty modelu, a także zmieniając priorytety segmentów.

Określanie priorytetu segmentów: Regułom modelu można nadawać dowolne rangi. Domyślnie segmenty modelu są wyświetlane w kolejności odzwierciedlającej ich priorytety, przy czym pierwszy segment ma najwyższy priorytet. Zmiana priorytetu jednego lub wielu segmentów powoduje odpowiednią zmianę modelu. Można modyfikować model, przenosząc segmenty na pozycje odpowiadające wyższym lub niższym priorytetom.

Aby określić priorytety segmentów modelu:

1. Zaznacz segment modelu, którego priorytet chcesz zmienić.
2. Kliknij jeden z dwóch przycisków ze strzałkami na pasku narzędzi panelu modelu roboczego, aby przenieść zaznaczony segment modelu na wyższą lub niższą pozycję na liście.

Po zmianie priorytetu wszystkie dotychczasowe wyniki ocen są ponownie obliczane i wyświetlane są nowe wartości.

Usuwanie segmentów: Aby usunąć jeden lub wiele segmentów:

1. Zaznacz segment modelu.
2. Z menu Edycja wybierz polecenie **Usuń segment** lub kliknij przycisk usuwania na pasku narzędzi panelu modelu roboczego.

Miary zmodyfikowanego modelu zostaną ponownie obliczone, a model odpowiednio się zmieni.

Wykluczanie segmentów: Ponieważ w analizie interesują nas konkretne grupy, działania biznesowe podejmować będziemy zwykle tylko w odniesieniu do wybranych segmentów modelu. Wdrażając model, można wykluczyć niektóre jego segmenty. Wykluczone segmenty są oceniane jako wartości null. Wykluczenie segmentu nie oznacza, że nie będzie on używany; oznacza natomiast, że wszystkie rekordy spełniające jego regułę będą wykluczone z listy mailingowej. Reguła nadal jest stosowana, ale w inny sposób.

Aby wykluczyć konkretne segmenty modelu:

1. Zaznacz segment na panelu Model roboczy.
2. Kliknij przycisk **Zmień wykluczenie segmentów** na pasku narzędzi w panelu Model roboczy. W wybranej kolumnie Zmienna przewidywana wybranego segmentu wyświetlane jest teraz słowo **Wykluczono**.

Uwaga: W odróżnieniu od segmentów usuniętych, segmenty wykluczone pozostają dostępne do wykorzystania w ostatecznym modelu. Wykluczone segmenty wpływają na wyniki prezentowane na wykresie.

Zmiana wartości przewidywanej: Okno dialogowe zmiany wartości przewidywanej umożliwia zmianę wartości przewidywanej dla bieżącej zmiennej przewidywanej.

Obrazy stanu i wyniki sesji z wartością zmiennej przewidywanej inną niż obowiązująca w modelu roboczym są oznaczone żółtym tłem odpowiednich wierszy tabeli. Wyróżnione w ten sposób obrazy stanu/wyniki sesji są zdezaktualizowane.

W oknie dialogowym tworzenia/edycji zadania eksploracji wyświetlana jest wartość przewidywana bieżącego modelu roboczego. Wartość przewidywana nie jest zapisywana razem z zadaniem eksploracji. Jest ona przejmowana z modelu roboczego.

Gdy użytkownik nada zapisanemu modelowi status modelu roboczego, i ten nowy model ma inną wartość przewidywaną niż bieżący model roboczy (np. zmienioną poprzez edycję wyniku alternatywnego lub kopii obrazu stanu), wartość przewidywana zapisanego modelu zmieni się na wartość przewidywaną z modelu roboczego (wartość przewidywana wyświetlana w panelu Model roboczy nie zmienia się). Metryki modelu są ponownie poddawane ewaluacji z uwzględnieniem nowej wartości przewidywanej.

Generowanie nowego modelu

Okno dialogowe Generuj nowy model zawiera opcje umożliwiające nadanie modelowi nazwy i wybranie miejsca, w którym zostanie utworzony.

Nazwa modelu. Wybierz opcję **Użytkownika**, aby zmodyfikować automatycznie wygenerowaną nazwę lub utworzyć dla węzła unikalną nazwę, która ma być wyświetlana w obszarze roboczym strumienia.

Utwórz węzeł na. Wybranie opcji **Obszar roboczy** spowoduje, że nowy model zostanie umieszczony w obszarze roboczym; wybranie opcji **Paleta modeli** spowoduje umieszczenie modelu na palecie modeli; wybranie opcji **Łącznie** spowoduje umieszczenie modelu zarówno w obszarze roboczym, jak i na palecie modeli.

Dołącz stan sesji interaktywnej. Gdy ta opcja jest włączona, w wygenerowanym modelu zachowywany jest stan sesji interaktywnej. Gdy później z modelu zostanie wygenerowany węzeł modelowania, stan ten zostanie przejęty i posłuży do zainicjowania sesji interaktywnej. Ocena nowych danych przez model odbywa się tak samo, niezależnie od stanu tej opcji. Gdy opcja nie jest wybrana, model nadal jest w stanie utworzyć węzeł budowy, ale będzie to węzeł o bardziej ogólnym charakterze, który uruchomi sesję interaktywną od nowa zamiast kontynuować dotychczasową sesję. Jeśli zmienisz ustawienia węzła, ale wykonasz go z zapisanym stanem, to zmienione ustawienia będą ignorowane na rzecz tych, które zapisane zostały w stanie.

Uwaga: Jedynymi metrykami pozostałymi w modelu są metryki standardowe. Pozostałe metryki są zachowywane w stanie interaktywnym. Wygenerowany model nie odzwierciedla zapisanego stanu interaktywnego zadania eksploracji. Decision List Viewer po uruchomieniu wyświetla ustawienia pierwotnie wybrane w Przeglądarce.

Więcej informacji można znaleźć w temacie “Ponowne generowanie węzła modelowania” na stronie 48.

Ocena jakości modelu

Warunkiem powodzenia modelowania jest przeprowadzenie dokładnej oceny jakości modelu przed wdrożeniem go w środowisku produkcyjnym. Decision List Viewer oferuje szereg miar statystycznych i biznesowych, których można użyć do oceny wpływu modelu na rzeczywiste wyniki. Należą do nich wykresy korzyści oraz mechanizmy współdziałania z programem Excel. Użytkownicy mogą dzięki nim symulować koszty i korzyści charakterystyczne dla różnych scenariuszy.

Jakość modelu należy oceniać następującymi metodami:

- Używając predefiniowanych miar statystycznych i biznesowych, które udostępnia Decision List Viewer (prawdopodobieństwo, częstość).
- Oceniając miary zaimportowane z programu Microsoft Excel.
- Wizualizując model na wykresie korzyści.

Organizacja miar modelu: Decision List Viewer udostępnia opcje służące do definiowania miar obliczanych i wyświetlanych w formie kolumn. Do każdego segmentu może być przypisane domyślne pokrycie, częstość, prawdopodobieństwo i błąd. Miary te są przedstawione w formie kolumn. Można także tworzyć nowe miary; one również będą wyświetlane w postaci kolumn.

Definiowanie miar modelu

Aby dodać miarę do modelu lub zdefiniować istniejącą miarę:

1. Z menu Narzędzia wybierz polecenie **Organizuj miary modelu** lub kliknij model prawym przyciskiem myszy i wybierz to samo polecenie. Zostanie otwarte okno dialogowe Organizuj miary modelu.
2. Kliknij przycisk **Dodaj nową miarę modelu** (na prawo od kolumny Przedstaw). Nowa miara zostanie wyświetlona w tabeli.
3. Podaj nazwę miary i wybierz odpowiedni typ, opcję wyświetlania i wybór. Kolumna Przedstaw zawiera informację o tym, czy miara będzie wyświetlana w modelu roboczym. Definiując istniejącą miarę, wybierz odpowiednią metrykę i wybór oraz wskaż, czy miara będzie wyświetlana w modelu roboczym.
4. Kliknij przycisk **OK**, aby wrócić do obszaru roboczego Decision List Viewer. Jeśli kolumna Przedstaw dla nowej miary była zaznaczona, nowa miara zostanie wyświetlona w modelu roboczym.

Metryki użytkownika w programie Excel

Więcej informacji można znaleźć w temacie “Ocena jakości w programie Excel”.

Odświeżanie miar: W niektórych przypadkach — na przykład po zastosowaniu istniejącego modelu do nowego zbioru klientów — konieczne jest ponowne obliczenie miar modelu.

Aby ponownie obliczyć (odświeżyć) miary modelu:

Z menu Edycja wybierz polecenie **Odśwież wszystkie miary**.

lub

Naciśnij klawisz F5.

Wszystkie miary zostaną obliczone od nowa, a nowe wartości zostaną wyświetlone w modelu roboczym.

Ocena jakości w programie Excel: Decision List Viewer może współpracować z programem Microsoft Excel, dzięki czemu można wykorzystać własne obliczenia wartości i wzory na zysk w procesie budowania modelu, by symulować w ten sposób różne koszty/korzyści w różnych scenariuszach. Powiązanie z programem Excel umożliwia wyeksportowanie danych do tego programu i wykorzystanie ich do utworzenia prezentacji, obliczenia własnych miar, np. złożonych miar zysku i zwrotu z inwestycji. Decision List Viewer może wyświetlać te miary użytkownika w trakcie budowania modelu.

Uwaga: Aby użytkownicy biznesowi mogli pracować z arkuszem kalkulacyjnym programu Excel, ekspert od analiz CRM musi zdefiniować dane konfiguracyjne, dzięki którym Decision List Viewer będzie zsynchronizowana z programem Microsoft Excel. Takie dane konfiguracyjne zapisane są w pliku arkusza kalkulacyjnego Excel i określają, które informacje Decision List Viewer przekaże do programu Excel i z niego odbierze.

Opisana poniżej procedura ma zastosowanie tylko wtedy, gdy zainstalowany jest program MS Excel. Jeśli program Excel nie jest zainstalowany, opcje synchronizacji modeli z tym programem nie są wyświetlane.

Aby zsynchronizować modele z programem MS Excel:

1. Otwórz model, uruchom sesję interaktywną i z menu Narzędzia wybierz polecenie **Organizuj miary modelu**.
2. Zaznacz **Tak** przy opcji **Oblicz niestandardowe miary w programie Excel**. Pole **Skoroszyt** stanie się aktywne i będzie w nim można wybrać wstępnie skonfigurowany szablon skoroszytu programu Excel.

3. Kliknij przycisk **Połącz z programem Excel**. Zostanie otwarte okno dialogowe Otwórz, w którym można w lokalnym lub sieciowym systemie plików wybrać położenie wstępnie skonfigurowanego szablonu.
4. Wybierz odpowiedni szablon skoroszytu programu Excel i kliknij przycisk **Otwórz**. Zostanie uruchomiony wybrany szablon skoroszytu programu Excel; za pomocą paska zadań systemu Windows (lub kombinacji klawiszy Alt-Tab wróć do okna dialogowego wyboru predyktorów dla miar użytkownika.
5. Wybierz odpowiednie odwzorowania między nazwami metryk zdefiniowanymi w szablonie skoroszytu programu Excel a nazwami metryk modelu i kliknij przycisk **OK**.

Po nawiązaniu połączenia program Excel zostanie uruchomiony ze wstępnie skonfigurowanym szablonem i wyświetli reguły modelu w arkuszu kalkulacyjnym. Decision List Viewer wyświetla wyniki obliczone w programie Excel jako nowe kolumny.

Uwaga: Metryki z programu Excel nie są zachowywane razem z modelem; metryki są ważne tylko w trakcie aktywnej sesji. Można jednak tworzyć obrazy stanu zawierające metryki programu Excel. Metryki programu Excel zapisane w widokach obrazów stanu mają zastosowanie wyłącznie do porównań historycznych i nie są odświeżane po ponownym otwarciu. Więcej informacji można znaleźć w temacie “Karta Obrazy stanu” na stronie 157. Metryki z programu Excel nie będą wyświetlane w obrazach stanu do czasu ponownego nawiązania połączenia z szablonem w programie Excel.

Konfiguracja integracji z programem MS Excel: Decision List Viewer i Microsoft współdziała z programem Microsoft Excel za pośrednictwem wstępnie skonfigurowanego szablonu skoroszytu programu Excel. Szablon ten zawiera trzy arkusze:

Model Measures. Miary, które do programu wyeksportowała Decision List Viewer, miary użytkownika w programie Excel oraz sumy obliczeń (zdefiniowane na arkuszu Settings).

Settings. Zawiera zmienne służące do generowania obliczeń na podstawie miar, które wyeksportowała Decision List Viewer, oraz miar użytkownika w programie Excel.

Configuration. Udostępnia opcje określające, które miary Decision List Viewer będzie eksportować, oraz służące do definiowania własnych miar w programie Excel.

OSTRZEŻENIE: Struktura arkusza Konfiguracja jest ściśle zdefiniowana. **NIE** wolno edytować żadnych komórek w obszarze zaznaczonym na zielono.

- **Metrics From Model.** Określa metryki Decision List Viewer z modelu używane w obliczeniach.
- **Metrics To Model.** Określa, które metryki wygenerowane przez program Excel zaimportuje Decision List Viewer. Decision List Viewer wyświetla metryki wygenerowane przez program Excel jako kolumny miar.

Uwaga: Metryki z programu Excel nie są zachowywane po wygenerowaniu nowego modelu; metryki są ważne tylko w trakcie aktywnej sesji.

Modyfikowanie miar modelu: Poniższy przykład ilustruje kilka sposobów modyfikowania miar modelu:

- Zmiana istniejącej miary.
- Importowanie dodatkowej miary standardowej z modelu.
- Eksportowanie dodatkowej miary użytkownika do modelu

Zmiana istniejącej miary

1. Otwórz szablon i wybierz arkusz Konfiguracja.
2. Edytuj dowolne pola **Name** (Nazwa) lub **Description** (Opis), zaznaczając i nadpisując ich zawartość.

   Uwaga: aby zmienić miarę — np. aby model pytał użytkownika o Prawdopodobieństwo, a nie Częstość — wystarczy zmienić nazwę i opis na arkuszu **Metrics From Model**. Zostaną one wówczas wyświetlone w modelu i użytkownik będzie mógł wybrać odpowiednią miarę do odwzorowania.

Importowanie dodatkowej miary standardowej z modelu

1. Otwórz szablon i wybierz arkusz Konfiguracja.
2. Z menu wybierz:

   Narzędzia > Ochrona > Nie chroń arkusza
3. Wybierz żółtą komórkę A5, która zawiera słowo **End** (Koniec).
4. Z menu wybierz:

   Wstaw > Wiersze
5. Wpisz nazwę (**Name**) i opis (**Description**) nowej miary. Na przykład **Błąd** i **Błąd skojarzony z segmentem**.
6. W komórce C5 wprowadź formułę **=COLUMN('Model Measures'!N3)**.
7. W komórce D5 wprowadź formułę **=ROW('Model Measures'!N3)+1**.

   Formuły te spowodują, że nowa miara pojawi się w kolumnie N arkusza Model Measures, która obecnie jest pusta.
8. Z menu wybierz:

   Narzędzia > Ochrona > Chroń arkusz
9. Kliknij przycisk **OK**.
10. Upewnij się, że komórka N3 na arkuszu Model Measures zawiera słowo **Error** (Błąd) jako tytuł nowej kolumny.
11. Zaznacz całą kolumnę N.
12. Z menu wybierz:

    Format > Komórki
13. Domyślnie wszystkie komórki mają format liczbowy **Ogólne**. Kliknij format **Procentowe**, aby zmienić sposób wyświetlania liczb. Ułatwi to sprawdzanie danych w programie Excel; ponadto umożliwi wykorzystanie danych w inny sposób, np. do utworzenia wykresu.
14. Kliknij przycisk **OK**.
15. Zapisz arkusz jako szablon w formacie Excel 2003, z unikalną nazwą i rozszerzeniem *.xlt*. Zaleca się zapisanie nowego szablonu w specjalnie wyznaczonym do tego celu miejscu w lokalnym lub sieciowym systemie plików, co ułatwi późniejsze odszukanie pliku.

Eksportowanie dodatkowej miary użytkownika do modelu

1. Otwórz szablon, do którego w poprzednim przykładzie dodano kolumnę Błąd; wybierz arkusz Configuration.
2. Z menu wybierz:

   Narzędzia > Ochrona > Nie chroń arkusza
3. Wybierz żółtą komórkę A14, która zawiera słowo **End** (Koniec).
4. Z menu wybierz:

   Wstaw > Wiersze
5. Wpisz nazwę (**Name**) i opis (**Description**) nowej miary. Na przykład **Przeskalowany błąd** i **Przeskalowany błąd z Excela**.
6. W komórce C14 wprowadź formułę **=COLUMN('Model Measures'!O3)**.
7. W komórce D14 wprowadź formułę **=ROW('Model Measures'!O3)+1**.

   Te formuły określają, że kolumna O będzie źródłem nowej miary dla modelu.
8. Wybierz arkusz Settings.
9. W kolumnie A17 wprowadź opis **'- Przeskalowany błąd**.
10. W komórce B17 wprowadź współczynnik skalowania **10**.
11. Na arkuszu Model Measures wprowadź opis **Przeskalowany błąd** w komórce O3 jako tytuł nowej kolumny.
12. W komórce O4 wprowadź formułę **=N4*Settings!B17**.
13. Zaznacz narożnik komórki O4 i przeciągnij go w dół do komórki O22, aby skopiować formułę do wszystkich komórek.
14. Z menu wybierz:

    Narzędzia > Ochrona > Chroń arkusz

15. Kliknij przycisk **OK**.
16. Zapisz arkusz jako szablon w formacie Excel 2003, z unikalną nazwą i rozszerzeniem *.xlt*. Zaleca się zapisanie nowego szablonu w specjalnie wyznaczonym do tego celu miejscu w lokalnym lub sieciowym systemie plików, co ułatwi późniejsze odszukanie pliku.

Po nawiązaniu połączenia z programem Excel przy użyciu tego szablonu wartość Błąd będzie dostępna jako nowa miara użytkownika.

Wizualizacja modeli

Wpływ modelu najłatwiej będzie ocenić, gdy model ten przedstawimy w postaci wizualnej. Korzystając z wykresu korzyści, można na bieżąco analizować korzyści biznesowe i techniczne, jakie przynosi model, w czasie rzeczywistym obserwując efekty różnych scenariuszy alternatywnych. W sekcji "Wykres korzyści" omówiono przewagi modelowania nad losowym podejmowaniem decyzji oraz możliwość bezpośredniego porównywania wielu wykresów utworzonych na podstawie modeli alternatywnych.

Wykres korzyści: Wykres korzyści przedstawia wartości z kolumny *Korzyści (%)* w tabeli. Korzyści są definiowane jako proporcja trafień w każdym przyroście względem łącznej liczby trafień w drzewie, według równania:

(trafień w przyroście / łączna liczba trafień) x 100%

Wykres korzyści obrazuje w istocie, jak szeroki zakres danych należy objąć, aby uzyskać daną wartość procentową wszystkich trafień w drzewie. Przekątna przedstawia oczekiwaną odpowiedź dla całej próby w przypadku, gdyby model nie został użyty. W takim przypadku wskaźnik odpowiedzi byłby stały, ponieważ prawdopodobieństwo udzielenia odpowiedzi przez poszczególne osoby jest jednakowe. Podwojenie wyniku wymagałoby przepytania dwukrotnie większej liczby osób. Krzywa wskazuje, o ile można poprawić odpowiedź, uwzględniając tylko osoby, w przypadku których korzyści mieszczą się w górnych percentylach. Na przykład uwzględnienie pierwszych 50% może dać więcej niż 70% odpowiedzi pozytywnych. Im większe nachylenie krzywej, tym wyższa korzyść.

Aby wyświetlić wykres korzyści:

1. Otwórz strumień zawierający węzeł Lista decyzyjna i uruchom z tego węzła sesję interaktywną.
2. Kliknij kartę **Korzyści**. W zależności od wybranych podzbiorów może być widoczny jeden wykres lub dwa wykresy (dwa wykresy pojawią się np. gdy dla miar modelu zdefiniowano zarówno podzbiór uczący, jak i testowy).

Domyślnie wykresy są wyświetlane segmentami. Można wybrać wyświetlanie kwantylami, wybierając opcję **Kwantyle**, a następnie wybierając odpowiednią metodę definiowania kwantyli z menu rozwijanego.

Opcje wykresu: Funkcja Opcje wykresu umożliwia wybranie modeli i obrazów stanu do przedstawienia na wykresie, wyboru podzbiorów do wykreślenia oraz włączenie lub wyłączenie wyświetlania etykiet segmentów.

Modele do wykreślenia

Bieżące modele. Umożliwia wybranie modeli, które mają być przedstawione na wykresie. Można wybrać model roboczy lub dowolne utworzone obrazy stanu modeli.

Podzbiory do wykreślenia

Podziały dla wykresu po lewej stronie. Lista rozwijana zawiera opcje wyświetlania wszystkich zdefiniowanych podzbiorów lub wszystkich danych.

Podziały dla wykresu po prawej stronie. Lista rozwijana zawiera opcje wyświetlania wszystkich zdefiniowanych podzbiorów, wszystkich danych lub tylko wykresu po lewej stronie. Gdy wybrana jest opcja **Przedstaw tylko lewy panel**, wyświetlany jest tylko lewy wykres.

Wyświetl etykiety segmentów. Gdy ta opcja jest włączona, na wykresach wyświetlane są etykiety segmentów.

Rozdział 10. Modele statystyczne

W modelach statystycznych stosowane są wyrażenia matematyczne umożliwiające kodowanie informacji wyodrębnionych z danych. W niektórych przypadkach techniki modelowania statystycznego bardzo szybko dostarczają odpowiednich modeli. Nawet dla problemów, w przypadku których bardziej elastyczne techniki uczenia maszyn (takie jak sieci neuronowe) mogą dawać znacznie lepsze wyniki, można użyć niektórych modeli statystycznych jako podstawowych modeli predykcyjnych do oceny działania bardziej zaawansowanych technik.

Dostępne są następujące węzły modelowania statystycznego.

Modele regresji liniowej przewidują docelową wartość ilościową na podstawie liniowych relacji między docelową wartością ilościową a jednym lub większą liczbą predyktorów.

Regresja logistyczna to technika statystyczna umożliwiająca klasyfikację rekordów na podstawie wartości zmiennych wejściowych. Jest ona analogiczna do regresji liniowej, lecz bazuje na przewidywanej zmiennej jakościowej zamiast na przedziale liczbowym.

Węzeł analizy PCA/czynnikowej udostępnia wydajne techniki redukcji danych pozwalające obniżyć stopień ich złożoności. Analiza głównych składowych (ang. Principal Components Analysis, PCA) znajduje kombinacje liniowe zmiennych wejściowych, które umożliwiają określenie wariancji w całym zestawie zmiennych, pod warunkiem że składowe są zlokalizowane ortogonalnie (prostopadle) do siebie. Analiza czynnikowa próbuje zidentyfikować współczynniki objaśniające wzory korelacji występujące w ramach zbiorów obserwowanych zmiennych. W przypadku obu podejść celem jest znalezienie niewielkiej liczby zmiennych wyliczanych w efektywny sposób, która podsumowuje informacje w oryginalnym zestawie zmiennych.

Analiza dyskryminacyjna opiera się na ściślejszych założeniach niż regresja logistyczna, lecz może stanowić wartościową alternatywę lub uzupełnienie analizy metodą regresji logistycznej w przypadku spełnienia tych założeń.

Węzeł Modele uogólnione rozszerza ogólny model liniowy w taki sposób, że zmienna zależna jest liniowo powiązana z czynnikami i współzmiennymi za pośrednictwem określonej funkcji łączenia. Model pozwala ponadto, aby zmienna zależna nie miała rozkładu normalnego. Obejmuje ona funkcjonalność dużej liczby modeli statystycznych, m.in. regresji liniowej, regresji logistycznej, modeli logarytmiczno-liniowych dla danych o liczebności.

Uogólniony liniowy model mieszany (GLMM) stanowi wersję modelu liniowego rozszerzoną w taki sposób, że zmienna przewidywana może mieć rozkład inny niż normalny, jest liniowo powiązana z czynnikami i współzmiennymi za pośrednictwem określonej funkcji łączenia, a obserwacje mogą być skorelowane. Uogólnione liniowe modele mieszane obejmują szeroki wachlarz modeli, począwszy od prostych modeli regresji liniowej, aż po złożone wielopoziomowe modele dla danych z obserwacji długofalowych nieposiadających rozkładu normalnego.

Węzeł regresji Coxa umożliwia utworzenie modelu przeżycia dla danych określających czasy do wystąpienia zdarzeń i zawierających ocenzurowane rekordy. Model generuje funkcję przeżycia przewidującą prawdopodobieństwo, że zdarzenie będące przedmiotem zainteresowania wystąpiło w określonym czasie (*t*) dla danych wartości zmiennych wejściowych.

Węzeł Liniowy

Regresja liniowa to często stosowana technika statystyczna przeznaczona do klasyfikowania rekordów na podstawie wartości numerycznych zmiennych wejściowych. Regresja liniowa dopasowuje prostą linię lub powierzchnię, która minimalizuje rozbieżności między wartościami przewidywanymi a rzeczywistymi wynikami.

Wymagania. W przypadku modelu regresji liniowej można stosować tylko zmienne numeryczne. Wymagana jest dokładnie jedna zmienna przewidywana (o roli ustawionej na **Zmienna przewidywana**) oraz jeden lub większa liczba predyktorów (o roli ustawionej na **Dane wejściowe**). Zmienne posiadające rolę **Łącznie** lub **Brak** są ignorowane — tak samo, jak zmienne nienumeryczne. (W razie potrzeby zmienne nienumeryczne mogą być rejestrowane przy użyciu węzła wyliczeń).

Mocne strony. Modele regresji liniowej są stosunkowo proste i zapewniają łatwy w interpretacji wzór matematyczny do generowania predykcji. Regresja liniowa jest od dawna stosowana jako procedura statystyczna, dlatego właściwości tych modeli są dobrze zrozumiałe. Uczenie modeli liniowych zwykle przebiega bardzo szybko. Węzeł Liniowy udostępnia metody automatycznego wyboru zmiennych w celu wyeliminowania z równania nieistotnych zmiennych wejściowych.

Uwaga: W sytuacjach, gdy zmienna przewidywana jest jakościowa, a nie jest zakresem ilościowym, np. **tak/nie** lub **odejście/brak odejścia**, jako alternatywę można stosować regresję logistyczną. Regresja logistyczna zapewnia także pokrycie nienumerycznych zmiennych wejściowych, co eliminuje potrzebę ponownego kodowania tych zmiennych. Więcej informacji można znaleźć w temacie “Węzeł logistyczny” na stronie 180.

Modele liniowe

Modele liniowe przewidują przewidywaną zmienną ilościową na podstawie liniowych relacji między przewidywaną a jednym lub większą liczbą predyktorów.

Modele liniowe są stosunkowo proste i zapewniają łatwy w interpretacji wzór matematyczny do oceny. W przeciwieństwie do innych typów modeli dla tego samego zbioru danych (takich jak sieci neuronowe czy drzewa decyzyjne), właściwości tych modeli łatwo zrozumieć i zwykle można się ich szybko nauczyć.

Przykład. Firma ubezpieczeniowa o ograniczonych środkach na sprawdzenie roszczeń ubezpieczeniowych właścicieli domów chce stworzyć model przybliżający koszty roszczeń. Po wdrożeniu tego modelu w centrach usługowych przedstawiciele będą mogli wprowadzać informacje na temat roszczeń podczas rozmów telefonicznych z klientem i natychmiast otrzymać „przybliżony” koszt roszczenia bazujący na wcześniejszych danych.

Wymagania dotyczące zmiennych. Musi istnieć Zmienna przewidywana i co najmniej jedna Zmienna wejściowa. Domyślnie pola ze wstępnie zdefiniowanymi rolami Obie lub Żadna nie są używane. Docelowa musi być zmienną ciągłą (ilościową). Nie ma żadnych ograniczeń poziomu pomiaru dla predyktorów (wejścia); zmienne jakościowe (flagi, nominalne oraz porządkowe) są używane jako czynniki w modelu, a zmienne ciągłe używane są jako współzmienne.

Cele

Co zamierzasz zrobić?

- **Utworzyć nowy model.** Stworzyć całkowicie nowy model. Jest to zwykle działanie węzła.
- **Kontynuować uczenie istniejącego modelu.** Szkolenie jest kontynuowane z wykorzystaniem ostatniego modelu, utworzonego z powodzeniem przez węzeł. Dzięki temu możliwa jest aktualizacja lub odświeżenie istniejącego

modelu bez konieczności wejścia do oryginalnych danych i może skutkować znacznie wydajniejszym działaniem, ponieważ tylko nowe lub zaktualizowane rekordy są podawane do strumienia. Szczegóły dotyczące poprzedniego modelu są zapisywane z węzłem modelowania, umożliwiając używanie tej opcji nawet, jeśli poprzednie wartościowe informacje z modelu są już niedostępne w strumieniu lub w palecie Modeli.

Uwaga: Gdy opcja ta jest włączona, wszystkie inne elementy sterowania w zakładkach Zmienne i Opcje budowania są wyłączone.

Jaki chcesz osiągnąć cel? Zaznacz odpowiedni cel.

- **Zbudować model standardowy.** Ta metoda tworzy pojedynczy model do przewidywania przy pomocy predyktorów. Ogólnie rzecz biorąc standardowe modele są łatwiejsze w interpretacji i można je szybciej ocenić w porównaniu ze wzmocnionymi, spakowanymi lub dużymi zestawami zbiorów danych.

  Uwaga: Aby użyć tej opcji w modelach rozdzielonych razem z opcją **Kontynuuj uczenie istniejącego modelu** należy mieć połączenie z produktem Analytic Server.
- **Zwiększyć dokładność modelu (boosting).** Metoda ta tworzy model zespolony przy pomocy wzmocnienia, który generuje sekwencję modeli w celu uzyskania bardziej precyzyjnych predykcji. Tworzenie i ocena zestawów mogą trwać dłużej niż w przypadku standardowego modelu.

  Wzmocnienie tworzy kolejność „modeli składników", z których każdy został skompilowany na podstawie całego zbioru danych. Przed skompilowaniem każdego kolejnego modelu składników, rekordy są ważone na podstawie reszt po poprzednich modelach składników. Obserwacje o dużej wartości reszt dają stosunkowo wyższe wagi analizy tak, że kolejny model składników będzie się skupiał na dobrym przewidywaniu tych rekordów. Te modele składników tworzą razem model zespolony. Model zespolony ocenia nowe rekordy przy pomocy reguły łączenia; dostępne reguły zależą od poziomu pomiaru celu.
- **Wzmocnić stabilność modelu (agregacja bootstrapowa).** Metoda ta tworzy model zespolony przy pomocy spakowania (agregacja metodą bootstrap), które generuje wiele modeli w celu uzyskania bardziej wiarygodnych predykcji. Tworzenie i ocena zestawów mogą trwać dłużej niż w przypadku standardowego modelu.

  Agregacja metodą bootstrap (bagging) powiela zespół danych z przyuczenia, tworząc próbkowanie poprzez zastąpienie oryginalnego zbioru danych. W wyniku tego powstają próby bootstrap, które mają taki sam rozmiar, jak oryginalny zbiór danych. Następnie na podstawie każdego powielania kompilowany jest „model składników". Te modele składników tworzą razem model zespolony. Model zespolony ocenia nowe rekordy przy pomocy reguły łączenia; dostępne reguły zależą od poziomu pomiaru celu.
- **Utworzyć model dla dużych zbiorów danych.** Metoda ta tworzy model zespolony przez podział zbioru danych na oddzielne bloki danych. Wybierz tę opcję, jeśli Twój zbiór danych jest zbyt duży do utworzenia któregokolwiek z powyższych modeli, lub aby utworzyć model przyrostowy. Tworzenie tej opcji może być szybsze, ale ocena może potrwać dłużej niż w przypadku standardowego modelu.

Patrz temat "Zestawy" na stronie 173 w celu zapoznania się z ustawieniami związanymi z wspomaganiem, agregacją metodą bootstrap lub bardzo dużymi zbiorami danych.

Podstawy

Automatycznie przygotuj dane. Opcja ta umożliwia przekształcenie docelowej i predyktorów przez tą procedurę w celu maksymalizacji siły predykcji modelu. Wszystkie transformacje są zapisywane razem z modelem i zastosowane dla nowych danych do oceny. Oryginalne wersje przekształconych zmiennych są wyłączone z modelu. Domyślnie odbywa się następujące, automatyczne przygotowanie danych.

- **Obsługa daty i czasu.** Każdy predyktor daty jest przekształcany na nowy predyktor ciągły zawierający czas, który upłynął od daty odniesienia (1970-01-01). Każdy predyktor czasu jest przekształcany na nowy predyktor ciągły, zawierający czas, który upłynął od godziny odniesienia (00:00:00).
- **Korekta poziomu pomiaru.** Predyktory ciągłe zawierające mniej niż 5 odrębnych wartości są uznane za predyktory porządkowe. Predyktory porządkowe zawierające więcej niż 10 odrębnych wartości są uznane za predyktory ciągłe.
- **Obsługa wartości odstających.** Wartości predyktorów ciągłych znajdujące się poza wartością odcięcia (3 standardowe odchylenia od średniej) są ustawione na wartości odcięcia.

- **Traktowanie braków danych.** Brakujące wartości nominalnych predyktorów są zastępowane trybem podziału szkoleniowego. Brakujące wartości porządkowych predyktorów są zastępowane medianą podziału szkoleniowego. Brakujące wartości predyktorów ciągłych są zastępowane średnią podziału szkoleniowego.
- **Nadzorowane scalanie.** Model staje się skromniejszy poprzez zmniejszenie liczby zmiennych do przetworzenia w powiązaniu z docelową. Podobne kategorie identyfikuje się na podstawie relacji między wejściem a zmienną przewidywaną. Scalane są kategorie, które znacząco się nie różnią (to znaczy takie, których wartość p jest większa niż 0,1). Jeśli wszystkie kategorie są scalone w jedną, oryginalne i wyliczone wersje zmiennej są wyłączone z modelu, ponieważ nie mają żadnej wartości jako predyktora.

Poziom ufności. Jest to poziom ufności używany do wyliczania oszacowań przedziałów współczynników modelu w widoku Współczynniki. Należy podać wartość większą od 0 i mniejszą od 100. Domyślna wartość to 95.

Wybór modelu

Metoda selekcji modelu. Wybierz jedną z metod wyboru modelu (szczegóły znajdują się poniżej) lub **Uwzględnij wszystkie predyktory**, co powoduje wprowadzenie wszystkich dostępnych predyktorów jako składniki modelu efektów głównych. Domyślnie używana jest opcja **Krokowa postępująca**.

Selekcja krokowa postępująca. Działanie to rozpoczyna się bez efektów w modelu i dodaje oraz usuwa każdorazowo po jednym efekcie do momentu, aż nie można już nic dodać ani usunąć żadnego efektu zgodnie z kryterium krokowej.

- **Kryterium dla wprowadzenia/usunięcia.** Jest to statystyka używana do określenia tego, czy należy dodać lub usunąć efekt z modelu. **Kryterium informacyjne (AICC)** bazuje na prawdopodobieństwie zestawu uczenia przy założeniu, że model jest przystosowany do personalizacji nadmiernie złożonych modeli. **Statystyka F** bazuje na teście statystycznym poprawy błędu modelu. **Poprawione R-kwadrat** bazuje na dopasowaniu zestawu uczenia i jest przystosowany do personalizacji nadmiernie złożonych modeli. **Kryterium zabezpieczenia przed przeuczeniem (ASE)** bazuje na dopasowaniu (średniego kwadratu błędu lub ASE) zbioru zabezpieczającego przed przeuczeniem. Zbiór zabezpieczający przed przeuczeniem jest losową podpróbą około 30% oryginalnego zbioru danych, który nie został użyty do uczenia modelu.

  W przypadku wybrania kryterium innego niż **Statystyka F**, na każdym kroku do modelu dodawany jest efekt, który odpowiada największemu, dodatniemu wzrostowi kryterium. Jakiekolwiek efekty w tym modelu, które odpowiadają spadkowi kryterium są usuwane.

  W przypadku wybrania **Statystyka F** jako kryterium, na każdym kroku do modelu dodawany jest efekt, który ma najniższą wartość *p*, mniejszą niż określony próg, **z uwzględnieniem efektów z wartościami p, mniejszymi niż**. Domyślną wartością jest 0,05. Jakiekolwiek efekty w modelu o wartości *p* większej niż określony próg, **Usuń efekty z wartościami p większymi od**, są usuwane. Domyślną wartością jest 0,10.
- **Maksimum efektów w modelu ostatecznym.** Domyślnie wszystkie efekty można wprowadzić do modelu. Alternatywnie, jeśli algorytm krokowy zakończy krok z określoną, maksymalną liczbą efektów, algorytm zatrzymuje się z bieżącym zestawem efektów.
- **Określ maksymalną liczbę kroków.** Algorytm krokowy zatrzymuje się po wykonaniu określonej liczby kroków. Domyślnie jest to 3-krotność liczby dostępnych efektów. Alternatywnie podaj dodatnią liczbę całkowitą w maksymalnej liczbie kroków.

Selekcja Najlepsze podzbiory. Opcja ta zaznacza „wszystkie możliwe” modele lub co najmniej większy podzbiór możliwych modeli, niż krokowa postępująca, w celu wybrania najlepszego, zgodnie z kryterium najlepszego podzbioru. **Kryterium informacyjne (AICC)** bazuje na prawdopodobieństwie zestawu uczenia przy założeniu, że model jest przystosowany do personalizacji nadmiernie złożonych modeli. **Poprawione R-kwadrat** bazuje na dopasowaniu zestawu uczenia i jest przystosowany do personalizacji nadmiernie złożonych modeli. **Kryterium zabezpieczenia przed przeuczeniem (ASE)** bazuje na dopasowaniu (średniego kwadratu błędu lub ASE) zbioru zabezpieczającego przed przeuczeniem. Zbiór zabezpieczający przed przeuczeniem jest losową podpróbą około 30% oryginalnego zbioru danych, który nie został użyty do uczenia modelu.

Model o największej wartości kryterium jest wybierany jako najlepszy model.

Uwaga: Wybór najlepszych podzbiorów może wymagać większej liczby obliczeń niż wybór krokowy, postępujący. Jeśli zadanie najlepszych podzbiorów jest wykonywane w połączeniu z wzmocnieniem, agregacją metodą bootstrap lub bardzo dużymi zbiorami danych, tworzenie modelu z wykorzystaniem wyboru krokowego, postępującego może zająć dużo więcej czasu, niż tworzenie standardowego modelu.

Zestawy

Ustawienia te determinują zachowanie tworzenia zespołów, które występuje, gdy w Celach pożądane jest wspomaganie, agregacja metodą bootstrap lub bardzo duże zbiory danych. Opcje, które nie mają zastosowania do wybranego celu są ignorowane.

Agregacja metodą bootstrap i bardzo duże zbiory danych. Podczas wybierania zestawu jest to reguła służąca do łączenia przewidywanych wartości z modeli podstawowych w celu wyliczenia wartości oceny zestawu.

- **Domyślna reguła zespolenia dla docelowych wartości ilościowych.** Przewidywane wartości zestawu dla jakościowych zmiennych docelowych można połączyć przy pomocy średniej lub mediany przewidywanych wartości z modeli podstawowych.

Należy zwrócić uwagę, że gdy celem jest zwiększenie dokładności modelu, wybory reguły łączenia są ignorowane. Wzmocnienie zawsze wykorzystuje głos ważonej większości do oceny jakościowych zmiennych docelowych i ważonej mediany do oceny jakościowych zmiennych docelowych.

Boosting i agregacja bootstrapowa. Podaj liczbę modeli podstawowych do utworzenia, gdy celem jest zwiększenie dokładności lub stabilności modelu; dla agregacji metodą bootstrap jest to liczba prób agregacji metodą bootstrap. Powinna to być dodatnia liczba całkowita.

Zaawansowane

Replikacja wyników. Ustawienie wartości początkowej generatora liczb losowych umożliwia powielenie analizy. Generator liczb pseudolosowych służy do wyboru rekordów, które znajdują się w zbiorze zabezpieczającym przed przeuczeniem. Podaj liczbę całkowitą lub kliknij przycisk **Generuj**, co spowoduje utworzenie pseudolosowej liczby całkowitej między 1 a 2147483647, włącznie. Domyślną wartością jest 54752075.

Opcje modelu

Nazwa modelu. Można automatycznie generować nazwę modelu na podstawie zmiennych docelowych lub podać nazwę użytkownika. Automatycznie wygenerowana nazwa jest nazwą zmiennej docelowej.

Należy zwrócić uwagę, że przewidywana wartość jest zawsze obliczana podczas oceny modelu. Nazwa nowej zmiennej jest nazwą zmiennej docelowej, poprzedzonej znakami *$L-*. Na przykład dla zmiennej docelowej o nazwie *sprzedaż*, nowa zmienna miałaby nazwę *$L-sprzedaż*.

Podsumowanie modelu

Widok Podsumowanie modelu to szybkie podsumowanie modelu i jego dopasowania.

Tabela Tabela określa między innymi następujące ustawienia modelu wysokiego poziomu:

- nazwa elementu docelowego określona na zakładce Zmienne,
- czy zostało wykonane automatyczne przygotowanie danych zgodnie z ustawieniem w obszarze Podstawowe,
- metoda selekcji modelu oraz kryteria określone w ustawieniach Wybór modelu. Wyświetlana jest także wartość wyboru kryterium dla modelu finalnego, przedstawiona w mniejszym i lepszym formacie.

Wykres

Na wykresie przedstawiono dokładność modelu finalnego, który jest przedstawiony w większym i lepszym formacie. Wartość wynosi 100 x skorygowana *R*2 dla modelu finalnego.

Automatyczne przygotowanie danych

Widok ten przedstawia informacje o tym, które zmienne zostały wyłączone i w jaki sposób przekształcone zmienne zostały uwzględnione w kroku automatycznego przygotowania danych (ADP). Dla każdej zmiennej, która została

przekształcona lub wyłączona, tabela zawiera nazwę zmiennej, jej rolę w analizie i działanie podjęte przez krok ADP. Zmienne są posortowane alfabetycznie, w kolejności rosnącej, według nazw zmiennych. Działania, które można podjąć dla każdego z pól to:

- **Wyliczanie czasu trwania: miesiące** wylicza czas (w miesiącach), który upłynął od wartości pola zawierającego daty do bieżącej daty systemowej.
- **Wyliczanie czasu trwania: godziny** wylicza czas (w godzinach), który upłynął od wartości pola zawierającego daty do bieżącego czasu systemu.
- **Zmień poziom pomiaru z ilościowego na porządkowy** konwertuje zmienne ciągłe zawierające mniej niż 5 różnych wartości na zmienne porządkowe.
- **Zmień poziom pomiaru z porządkowego na ilościowy** konwertuje zmienne porządkowe zawierające więcej niż 10 różnych wartości na zmienne ciągłe.
- **Obcięcie wartości odstających** ustawia wartości predyktorów ciągłych znajdujące się poza wartością odcięcia (3 standardowe odchylenia od średniej) na wartość odcięcia.
- **Zastąp braki danych** zastępuje brakujące wartości zmiennych nominalnych trybem, zmiennych porządkowych — medianą, a zmiennych ilościowych — średnią.
- **Łączenie kategorii w celu maksymalizacji związku ze zmienną przewidywaną** identyfikuje „podobne" kategorie predyktorów bazujące na relacji między wejściem a docelową. Scalane są kategorie, które znacząco się nie różnią (to znaczy takie, których wartość *p* jest większa niż 0,05).
- **Wyklucz predyktor o stałych wartościach / po obsłudze wartości odstających / po scaleniu kategorii** usuwa predyktory o pojedynczej wartości po (możliwym) wykonaniu innych działań ADP.

Ważność predyktorów

Zazwyczaj działania modelujące mają koncentrować się na zmiennych predyktorów, które są najważniejsze, a opuszczane lub ignorowane mają być te zmienne, które są najmniej ważne. Wykres ważności predyktorów pomaga osiągnąć ten cel przez wskazanie względnej ważności każdego predyktora przy szacowaniu modelu. Ponieważ wartości są względne, suma wartości wszystkich wyświetlanych predyktorów wynosi 1,0. Ważność predyktora nie jest powiązana z dokładnością modelu. Jest powiązana z ważnością każdego predyktora przy prognozach, a nie z tym, czy taka prognoza jest dokładna.

Przewidywane przez Obserwowane

Przedstawia on wykres rozrzutu z kategoryzacją przewidywanych wartości na osi pionowej przez obserwowane wartości na osi poziomej. W idealnym przypadku punkty te powinny leżeć na prostej nachylonej pod kątem 45 stopni; widok ten może stwierdzić, czy którekolwiek z wyników zostały przewidziane przez model w sposób oczywisty.

Reszty

Widok ten przedstawia wykres diagnostyczny reszt modelu.

Style wykresu. Dostępne są różne style wyświetlania, które są dostępne z poziomu listy rozwijanej **Styl**.

- **Histogram.** Jest to podzielony histogram studentyzowanych reszt z nakładaniem normalnego rozkładu. Modele liniowe zakładają, że reszty mają normalny rozkład tak, że histogram w idealnych warunkach powinien znajdować się maksymalnie blisko gładkiej linii.
- **Wykres P-P.** Jest to podzielony wykres prawdopodobieństwo-prawdopodobieństwo porównujący studentyzowane reszty z rozkładem normalnym. Jeśli nachylenie naniesionych punktów jest mniej strome niż normalna linia, reszta będzie bardziej różnorodna niż normalny rozkład; jeśli nachylenie będzie bardziej strome, reszta będzie mniej różnorodna niż normalny rozkład. Jeśli naniesione punkty mają krzywą w kształcie litery S, wówczas rozkład reszt jest skośny.

Odstające

Tabela ta zestawia rekordy, które wywierają nadmierny wpływ na model i przedstawia identyfikator rekordu (jeśli został podany w zakładce Zmienne), wartość docelową i odległość Cooka. Odległość Cooka jest miarą stopnia, w jakim zmieniłyby się reszty dla wszystkich rekordów, przy wykluczaniu poszczególnych rekordów z obliczeń współczynników modelu. Duża odległość Cooka wskazuje, że wykluczenie z rekordu znacząco zmienia współczynnik i dlatego powinno być uznawane za wpływowe.

Wpływowe rekordy należy uważnie zbadać w celu określenia, czy można im nadać mniejszą wagę podczas oceny modelu lub obciąć wartości odstające do jakiegoś dopuszczalnego progu, lub całkowicie usunąć wpływowe rekordy.

Efekty

Widok ten przedstawia rozmiar każdego efektu w modelu.

Style. Dostępne są różne style wyświetlania, które są dostępne z poziomu listy rozwijanej **Styl**.

- **Diagram.** Jest to wykres z efektami posortowanymi od góry do dołu wg malejącej ważności predyktora. Linie łączące w diagramie są ważone na podstawie istotności efektu, gdzie większa szerokość linii odpowiada bardziej istotnym efektom (niższe wartości *p*). Umieszczenie kursora nad linią łączącą powoduje wyświetlenie podpowiedzi wskazującej wartość *p* oraz ważność efektu. Jest to ustawienie domyślne.
- **Tabela.** Jest to tabela ANOVA dla ogólnego modelu całkowitych i pojedynczych efektów modelu. Pojedyncze efekty są posortowane od góry do dołu ze zmniejszającą się ważnością predyktora. Weź pod uwagę, że domyślnie tabela jest zwinięta i ukazuje tylko wyniki modelu ogólnego. Aby zobaczyć wyniki dla pojedynczego efektu modelu, kliknij w tabeli komórkę **Model skorygowany**.

Ważność predyktora. Dostępny jest suwak ważności predyktora, który steruje widocznością predyktorów w widoku. Nie zmienia to modelu, ale pozwala po prostu na skoncentrowanie się na najistotniejszych predyktorach. Domyślnie wyświetlanych jest 10 najistotniejszych efektów.

Istotność. Dostępny jest suwak istotności, który dalej steruje widocznością efektów w widoku, poza efektami pokazanymi na podstawie ważności predyktora. Efekty o wartościach istotności większych niż wartości suwaka, pozostają ukryte. Nie zmienia to modelu, ale pozwala po prostu na skoncentrowanie się na najistotniejszych efektach. Domyślną wartością jest 1,00 tak, że na podstawie istotności żadne efekty nie są filtrowane.

Współczynniki

Widok ten przedstawia wartość każdego współczynnika w modelu. Należy zwrócić uwagę, że czynniki (predyktory jakościowe) są kodowane wskaźnikami w ramach modelu tak, że **efekty** zawierające czynniki będą miały generalnie wiele powiązanych **współczynników**; po jednym dla każdej kategorii z wyjątkiem kategorii odpowiadającej parametrowi nadmiarowemu (odniesienia).

Style. Dostępne są różne style wyświetlania, które są dostępne z poziomu listy rozwijanej **Styl**.

- **Diagram.** Jest to wykres przedstawiający najpierw wyraz wolny, następnie sortujący efekty z góry do dołu wg malejącej ważności predyktora. W efektach zawierających czynniki, współczynniki są posortowane rosnąco według wartości danych. Linie łączące w diagramie są pokolorowane na podstawie znaku współczynnika (patrz klucz diagramu) i ważone na podstawie istotności współczynnika, gdzie większa szerokość linii odpowiada bardziej istotnym współczynnikom (niższe wartości *p*). Umieszczenie kursora nad linią łączącą powoduje wyświetlenie podpowiedzi wskazującej wartość współczynnika, jego wartości *p* oraz ważność efektu, z którym parametr jest powiązany. Jest to domyślny styl.
- **Tabela.** Pokazuje ona wartości, testy istotności i przedziały ufności dla poszczególnych współczynników modelu. Po wolnym wyrazie, efekty są posortowane od góry do dołu ze zmniejszającą się ważnością predyktora. W efektach zawierających czynniki, współczynniki są posortowane rosnąco według wartości danych. Zwróć uwagę, że domyślnie tabela jest zwinięta, aby pokazywać tylko współczynnik, istotność i ważność każdego z parametrów modelu. Aby zobaczyć błąd standardowy, statyczne *t* i przedział ufności, kliknij w tabeli komórkę **Współczynnik**. Umieszczenie kursora nad nazwą parametru modelu znajdującego się w tabeli powoduje wyświetlenie podpowiedzi wskazującej nazwę parametru, efektu powiązanego z parametrem oraz, dla predyktorów jakościowych, wartości etykiet związanych z parametrem modelu. Może to być szczególnie pomocne przy sprawdzaniu nowych kategorii stworzonych w czasie, gdy automatyczne przygotowanie danych scala podobne kategorie predyktora jakościowego.

Ważność predyktora. Dostępny jest suwak ważności predyktora, który steruje widocznością predyktorów w widoku. Nie zmienia to modelu, ale pozwala po prostu na skoncentrowanie się na najistotniejszych predyktorach. Domyślnie wyświetlanych jest 10 najistotniejszych efektów.

Istotność. Dostępny jest suwak istotności, który dalej steruje widocznością współczynników w widoku, poza współczynnikami pokazanymi na podstawie ważności predyktora. Współczynniki o wartościach istotności większych

niż wartości suwaka, pozostają ukryte. Nie zmienia to modelu, ale pozwala po prostu na skoncentrowanie się na najistotniejszych współczynnikach. Domyślną wartością jest 1,00 tak, że na podstawie istotności żadne współczynniki nie są filtrowane.

Oszacowanie średnie

Są to wykresy wyświetlane dla predyktorów istotności. Wykres ten przedstawia na osi pionowej wartości docelowe, szacowane z modelu, dla każdej wartości predyktora na osi poziomej, utrzymując wszystkie inne predyktory na stałym poziomie. Zapewnia on przydatną wizualizację efektów współczynników docelowej każdego predyktora.

Uwaga: Jeśli żaden predyktor nie jest istotny, nie powstaje żadna oszacowana średnia.

Podsumowanie tworzenia modelu

Gdy w Ustawieniach wyboru modelu wybierze się algorytm wyboru modelu inny niż **Brak**, spowoduje to dostarczenie niektórych szczegółów procesu tworzenia modelu.

Krokowa postępująca. Gdy algorytmem wyboru jest krokowa postępująca, tabela przedstawia 10 ostatnich kroków w algorytmie krokowym. Pokazana jest wartość kryterium wyboru i efekty w modelu na tym kroku. Dzięki temu użytkownik ma świadomość w jakim stopniu każdy krok wpływa na model. Każda kolumna pozwala posortowanie wierszy tak, aby można było łatwo zobaczyć, które efekty znajdują się w modelu na danym kroku.

Najlepsze podzbiory. Gdy algorytmem wyboru są najlepsze podzbiory, tabela przedstawia 10 najlepszych modeli. Dla każdego modelu pokazana jest wartość kryterium wyboru i efekty w modelu. Obrazuje to stabilność najlepszych modeli; czy mają tendencję do posiadania wielu podobnych efektów o niewielkich różnicach, wówczas użytkownik może mieć względną pewność w „najlepszym" modelu; jeśli mają one tendencję do posiadania bardzo różnych efektów, wówczas niektóre efekty mogą być zbyt proste i powinny być połączone (lub jeden usunięty). Każda kolumna pozwala posortowanie wierszy tak, aby można było łatwo zobaczyć, które efekty znajdują się w modelu na danym kroku.

Ustawienia

Należy zwrócić uwagę, że przewidywana wartość jest zawsze obliczana podczas oceny modelu. Nazwa nowej zmiennej jest nazwą zmiennej docelowej, poprzedzonej znakami *$L-*. Na przykład dla zmiennej docelowej o nazwie *sprzedaż*, nowa zmienna miałaby nazwę *$L-sprzedaż*.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę, wykorzystując natywny kod SQL** Jeśli ta opcja jest wybrana, generowany jest natywny kod SQL w celu oceny modelu w bazie danych.

  Uwaga: Ta opcja może szybciej zwracać wyniki, ale rozmiar i złożoność natywnego kodu SQL wzrastają wraz ze wzrostem złożoności modelu.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Węzeł Liniowy-AS

W produkcie IBM SPSS Modeler dostępne są dwie różne wersje węzła Liniowy:

- **Liniowy** to tradycyjny węzeł działający na serwerze IBM SPSS Modeler Server.
- **Liniowy-AS** działa po nawiązaniu połączenia z serwerem IBM SPSS Analytic Server.

Regresja liniowa to często stosowana technika statystyczna przeznaczona do klasyfikowania rekordów na podstawie wartości numerycznych zmiennych wejściowych. Regresja liniowa dopasowuje prostą linię lub powierzchnię, która minimalizuje rozbieżności między wartościami przewidywanymi a rzeczywistymi wynikami.

Wymagania. W przypadku modelu regresji liniowej można stosować tylko zmienne numeryczne i predyktory jakościowe. Wymagana jest dokładnie jedna zmienna przewidywana (o roli ustawionej na **Zmienna przewidywana**) oraz jeden lub większa liczba predyktorów (o roli ustawionej na **Dane wejściowe**). Zmienne posiadające rolę **Łącznie** lub **Brak** są ignorowane — tak samo, jak zmienne nienumeryczne. (W razie potrzeby zmienne nienumeryczne mogą być rejestrowane przy użyciu węzła wyliczeń).

Mocne strony. Modele regresji liniowej są stosunkowo proste i zapewniają łatwy w interpretacji wzór matematyczny do generowania predykcji. Regresja liniowa jest od dawna stosowana jako procedura statystyczna, dlatego właściwości tych modeli są dobrze zrozumiałe. Uczenie modeli liniowych zwykle przebiega bardzo szybko. Węzeł Liniowy udostępnia metody automatycznego wyboru zmiennych w celu wyeliminowania z równania nieistotnych zmiennych wejściowych.

Uwaga: W sytuacjach, gdy zmienna przewidywana jest jakościowa, a nie jest zakresem ilościowym, np. **tak/nie** lub **odejście/brak odejścia**, jako alternatywę można stosować regresję logistyczną. Regresja logistyczna zapewnia także pokrycie nienumerycznych zmiennych wejściowych, co eliminuje potrzebę ponownego kodowania tych zmiennych. Więcej informacji można znaleźć w temacie “Węzeł logistyczny” na stronie 180.

Modele Liniowy-AS

Modele liniowe przewidują przewidywaną zmienną ilościową na podstawie liniowych relacji między przewidywaną a jednym lub większą liczbą predyktorów.

Modele liniowe są stosunkowo proste i zapewniają łatwy w interpretacji wzór matematyczny do oceny. W przeciwieństwie do innych typów modeli dla tego samego zbioru danych (takich jak sieci neuronowe czy drzewa decyzyjne), właściwości tych modeli łatwo zrozumieć i zwykle można się ich szybko nauczyć.

Przykład. Firma ubezpieczeniowa o ograniczonych środkach na sprawdzenie roszczeń ubezpieczeniowych właścicieli domów chce stworzyć model przybliżający koszty roszczeń. Po wdrożeniu tego modelu w centrach usługowych przedstawiciele będą mogli wprowadzać informacje na temat roszczeń podczas rozmów telefonicznych z klientem i natychmiast otrzymać „przybliżony” koszt roszczenia bazujący na wcześniejszych danych.

Wymagania dotyczące zmiennych. Musi istnieć Zmienna przewidywana i co najmniej jedna Zmienna wejściowa. Domyślnie pola ze wstępnie zdefiniowanymi rolami Obie lub Żadna nie są używane. Docelowa musi być zmienną ciągłą (ilościową). Nie ma żadnych ograniczeń poziomu pomiaru dla predyktorów (wejścia); zmienne jakościowe (flagi, nominalne oraz porządkowe) są używane jako czynniki w modelu, a zmienne ciągłe używane są jako współzmienne.

Podstawowe

Uwzględnienie wyrazu wolnego. Ta opcja uwzględnia przesunięcie na osi y, gdy wartość na osi x wynosi 0. Wyraz wolny jest zwykle uwzględniany w modelu. Jeśli można założyć, że dane przechodzą przez początek układu współrzędnych, wyraz wolny można wyłączyć z modelu.

Uwzględniaj interakcje dwukierunkowe. Ta opcja stanowi dla modelu instrukcję, która powoduje porównanie poszczególnych możliwych par zmiennych wejściowych w celu sprawdzenia, czy trend jednej zmiennej wpływa na pozostałe. Jeśli wpływa, wówczas te zmienne wejściowe z większym prawdopodobieństwem uwzględniane będą w macierzy planu.

Przedział ufności dla oszacowań współczynników (%). Jest to przedział ufności używany do wyliczania oszacowań przedziałów współczynników modelu w widoku Współczynniki. Należy podać wartość większą od 0 i mniejszą od 100. Domyślna wartość to 95.

Porządek sortowania predyktorów jakościowych. Te elementy sterujące określają porządek kategorii dla czynników (jakościowych zmiennych wejściowych) na potrzeby określenia „ostatniej" kategorii. Ustawienie kolejności sortowania jest ignorowane, jeśli zmienne wejściowe nie są jakościowe lub jeśli określona jest niestandardowa kategoria odniesienia.

Wybór modelu

Metoda selekcji modelu. Wybierz jedną z metod wyboru modelu (szczegóły znajdują się poniżej) lub **Uwzględnij wszystkie predyktory**, co powoduje wprowadzenie wszystkich dostępnych predyktorów jako składniki modelu efektów głównych. Domyślnie używana jest opcja **Krokowa postępująca**.

Selekcja krokowa postępująca. Działanie to rozpoczyna się bez efektów w modelu i dodaje oraz usuwa każdorazowo po jednym efekcie do momentu, aż nie można już nic dodać ani usunąć żadnego efektu zgodnie z kryterium krokowej.

- **Kryterium dla wprowadzenia/usunięcia.** Jest to statystyka używana do określenia tego, czy należy dodać lub usunąć efekt z modelu. **Kryterium informacyjne (AICC)** bazuje na prawdopodobieństwie zestawu uczenia przy założeniu, że model jest przystosowany do personalizacji nadmiernie złożonych modeli. **Statystyka F** bazuje na teście statystycznym poprawy błędu modelu. **Poprawione R-kwadrat** bazuje na dopasowaniu zestawu uczenia i jest przystosowany do personalizacji nadmiernie złożonych modeli. **Kryterium zabezpieczenia przed przeuczeniem (ASE)** bazuje na dopasowaniu (średniego kwadratu błędu lub ASE) zbioru zabezpieczającego przed przeuczeniem. Zbiór zabezpieczający przed przeuczeniem jest losową podpróbą około 30% oryginalnego zbioru danych, który nie został użyty do uczenia modelu.

  W przypadku wybrania kryterium innego niż **Statystyka F**, na każdym kroku do modelu dodawany jest efekt, który odpowiada największemu, dodatniemu wzrostowi kryterium. Jakiekolwiek efekty w tym modelu, które odpowiadają spadkowi kryterium są usuwane.

  W przypadku wybrania **Statystyka F** jako kryterium, na każdym kroku do modelu dodawany jest efekt, który ma najniższą wartość *p*, mniejszą niż określony próg, **z uwzględnieniem efektów z wartościami p, mniejszymi niż**. Domyślną wartością jest 0,05. Jakiekolwiek efekty w modelu o wartości *p* większej niż określony próg, **Usuń efekty z wartościami p większymi od**, są usuwane. Domyślną wartością jest 0,10.
- **Maksimum efektów w modelu ostatecznym.** Domyślnie wszystkie efekty można wprowadzić do modelu. Alternatywnie, jeśli algorytm krokowy zakończy krok z określoną, maksymalną liczbą efektów, algorytm zatrzymuje się z bieżącym zestawem efektów.
- **Określ maksymalną liczbę kroków.** Algorytm krokowy zatrzymuje się po wykonaniu określonej liczby kroków. Domyślnie jest to 3-krotność liczby dostępnych efektów. Alternatywnie podaj dodatnią liczbę całkowitą w maksymalnej liczbie kroków.

Selekcja Najlepsze podzbiory. Opcja ta zaznacza „wszystkie możliwe" modele lub co najmniej większy podzbiór możliwych modeli, niż krokowa postępująca, w celu wybrania najlepszego, zgodnie z kryterium najlepszego podzbioru. **Kryterium informacyjne (AICC)** bazuje na prawdopodobieństwie zestawu uczenia przy założeniu, że model jest przystosowany do personalizacji nadmiernie złożonych modeli. **Poprawione R-kwadrat** bazuje na dopasowaniu zestawu uczenia i jest przystosowany do personalizacji nadmiernie złożonych modeli. **Kryterium zabezpieczenia przed przeuczeniem (ASE)** bazuje na dopasowaniu (średniego kwadratu błędu lub ASE) zbioru zabezpieczającego przed przeuczeniem. Zbiór zabezpieczający przed przeuczeniem jest losową podpróbą około 30% oryginalnego zbioru danych, który nie został użyty do uczenia modelu.

Model o największej wartości kryterium jest wybierany jako najlepszy model.

Uwaga: Wybór najlepszych podzbiorów może wymagać większej liczby obliczeń niż wybór krokowy, postępujący. Jeśli zadanie najlepszych podzbiorów jest wykonywane w połączeniu z wzmocnieniem, agregacją metodą bootstrap lub bardzo dużymi zbiorami danych, tworzenie modelu z wykorzystaniem wyboru krokowego, postępującego może zająć dużo więcej czasu, niż tworzenie standardowego modelu.

Opcje modelu

Nazwa modelu. Można automatycznie generować nazwę modelu na podstawie zmiennych docelowych lub podać nazwę użytkownika. Automatycznie wygenerowana nazwa jest nazwą zmiennej docelowej.

Należy zwrócić uwagę, że przewidywana wartość jest zawsze obliczana podczas oceny modelu. Nazwa nowej zmiennej jest nazwą zmiennej docelowej, poprzedzonej znakami *$L-*. Na przykład dla zmiennej docelowej o nazwie *sprzedaż*, nowa zmienna miałaby nazwę *$L-sprzedaż*.

Wyniki interaktywne

Po uruchomieniu modelu Liniowy-AS dostępne są następujące wyniki.

Informacje o modelu

Widok Informacje o modelu zawiera kluczowe informacje o modelu. Tabela określa niektóre ustawienia modelu wysokiego poziomu, takie jak:

- Nazwa elementu zmiennej przewidywanej określona na karcie Zmienne
- Zmienna ważąca Regresja
- Metoda budowania modelu określona w ustawieniach Wybór modelu
- Liczba predyktorów w danych wejściowych
- Liczba predyktorów użytych w modelu końcowym
- Skorygowane kryterium informacyjne Akaike (AICC). AICC jest miarą wybierania i porównywania modeli mieszanych tworzonych na podstawie -2 logarytmu wiarygodności (ograniczonego). Mniejsze wartości oznaczają lepszy model. Wartość AICC „poprawia” wartość AIC w przypadku małych prób. Przy wzroście wielkości próby wartość AICC zbiega do wartości AIC.
- R-kwadrat. Jest to miara dobroci dopasowania modelu liniowego, czasami nazywana współczynnikiem determinacji. Jest to część zmienności w zmiennej zależnej wyjaśniona przez model regresji. Przyjmuje wartości z przedziału od 0 do 1. Małe wartości statystyki wskazują na słabe dopasowanie modelu do danych.
- Skorygowany R kwadrat

Podsumowanie rekordów

Widok Podsumowanie rekordów przedstawia informacje o liczbie i wartości procentowej rekordów (obserwacji) uwzględnionych w modelu i wykluczonych z modelu.

Ważność predyktorów

Zazwyczaj działania modelujące mają koncentrować się na zmiennych predyktorów, które są najważniejsze, a opuszczane lub ignorowane mają być te zmienne, które są najmniej ważne. Wykres ważności predyktorów pomaga osiągnąć ten cel przez wskazanie względnej ważności każdego predyktora przy szacowaniu modelu. Ponieważ wartości są względne, suma wartości wszystkich wyświetlanych predyktorów wynosi 1,0. Ważność predyktora nie jest powiązana z dokładnością modelu. Jest powiązana z ważnością każdego predyktora przy prognozach, a nie z tym, czy taka prognoza jest dokładna.

Przewidywane według obserwowanych

Przedstawia on wykres rozrzutu z kategoryzacją przewidywanych wartości na osi pionowej przez obserwowane wartości na osi poziomej. W idealnym przypadku punkty te powinny leżeć na prostej nachylonej pod kątem 45 stopni; widok ten może stwierdzić, czy którekolwiek z wyników zostały przewidziane przez model w sposób oczywisty.

Ustawienia

Należy zwrócić uwagę, że przewidywana wartość jest zawsze obliczana podczas oceny modelu. Nazwa nowej zmiennej jest nazwą zmiennej docelowej, poprzedzonej znakami *$L-*. Na przykład dla zmiennej docelowej o nazwie *sprzedaż*, nowa zmienna miałaby nazwę *$L-sprzedaż*.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślne: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania.** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych.** Jeśli ta opcja jest wybrana, dane użytkownika pobierane są z bazy danych i oceniane w produkcie SPSS Modeler.

Węzeł logistyczny

Regresja logistyczna, znana również pod nazwą **regresji nominalnej**, to technika statystyczna umożliwiająca klasyfikację rekordów na podstawie wartości zmiennych wejściowych. Jest ona analogiczna do regresji liniowej, lecz bazuje na przewidywanej zmiennej jakościowej zamiast na liczbowej. Obsługiwane są zarówno modele dwumianowe (w przypadku zmiennych przewidywanych z dwiema kategoriami dyskretnymi), jak i wielomianowe (w przypadku zmiennych przewidywanych z więcej niż dwiema kategoriami).

Regresja logistyczna działa w oparciu o tworzony zestaw wyrażeń odnoszących wartości zmiennych wejściowych do prawdopodobieństw powiązanych z każdą z kategorii zmiennych wyjściowych. Po wygenerowaniu modelu może być on używany do oceny prawdopodobieństw dla nowych danych. Dla każdego rekordu dla każdej możliwej kategorii wyjściowej obliczane jest prawdopodobieństwo członkostwa. Jako predykowana wartość wyjściowa dla tego rekordu przypisywana jest kategoria zmiennej przewidywanej o najwyższym prawdopodobieństwie.

Przykład modelu dwumianowego. Operator telekomunikacyjny jest zaniepokojony liczbą klientów odchodzących do konkurencji. Korzystając z danych wykorzystania usług, można tworzyć modele dwumianowe umożliwiające predykcję list klientów, którzy z największym prawdopodobieństwem mogą przenieść się do innego operatora, a następnie przedstawiać tym klientom bardziej zindywidualizowaną ofertę w celu zatrzymania możliwie największej ich liczby. Model dwumianowy jest używany, ponieważ zmienna przewidywana ma dwie odrębne kategorie (klienci najprawdopodobniej zamierzający odejść oraz najprawdopodobniej niezamierzający odejść).

Uwaga: W przypadku modeli dwumianowych zmienne łańcuchowe mają maksymalnie osiem znaków. W razie potrzeby dłuższe łańcuchy mogą zostać zrekodowane za pomocą węzła Rekodowanie lub węzła anonimizacji.

Przykład wielomianowy. Operator telekomunikacyjny pogrupował bazę klientów wg wzorców korzystania z usług, tworząc cztery kategorie. Korzystając z danych demograficznych do predykcji członkostwa grupy, można utworzyć model wielomianowy umożliwiający klasyfikację przyszłych klientów na grupy, a następnie indywidualizować oferty dla poszczególnych klientów.

Wymagania. Jedna lub więcej zmiennych wejściowych oraz dokładnie jedna przewidywana zmienna jakościowa z dwiema lub większą liczbą kategorii. W przypadku modelu dwumianowego zmienna przewidywana musi mieć poziom pomiaru *Flaga*. W przypadku modelu wielomianowego zmienna przewidywana może mieć poziom pomiaru *Flaga* lub *Nominalny*, z dwiema lub większą liczbą kategorii. Zmienne o roli *Łącznie* lub *Żadna* są ignorowane. Typy zmiennych używanych w modelu muszą być w pełni zrealizowane jako instancje.

Mocne strony. Modele regresji logistycznej są często w miarę dokładne. Pozwalają one obsługiwać symboliczne i numeryczne zmienne wejściowe. Oferują one przewidywane prawdopodobieństwa dla wszystkich kategorii zmiennych

przewidywanych, tak że z łatwością można zidentyfikować drugą w kolejności prawdopodobną pozycję. Modele logistyczne są najefektywniejsze, kiedy przynależność do grupy jest zmienną prawdziwie jakościową; jeśli przynależność do grupy opiera się na wartościach przedziału ilościowego (na przykład wysoki a niski iloraz inteligencji), należy rozważyć możliwość wykorzystania regresji liniowej, tak aby skorzystać z bogatszych informacji oferowanych przez pełny zakres wartości. Modele logistyczne umożliwiają także automatyczny wybór zmiennych. Inne metody, takie jak modele drzew decyzyjnych czy Wybór predyktora, są jednak znacznie szybsze w przypadku dużych zbiorów danych. W końcu, ponieważ modele logistyczne są zrozumiałe dla wielu analityków i specjalistów eksploracji danych, mogą być one przez nich używane jako baza porównawcza dla innych technik modelowania.

Przetwarzając duże zbiory danych, można w zauważalny sposób poprawić wydajność, wyłączając test ilorazu wiarygodności i zaawansowaną opcję generowania wyników. Więcej informacji można znaleźć w temacie “Zaawansowane wyniki regresji logistycznej” na stronie 186.

Ważne: Jeśli na dysku brakuje tymczasowej przestrzeni, budowa modelu przy użyciu dwumianowej regresji logistycznej może się nie powieść i może zostać wyświetlony błąd. W przypadku dużych zestawów danych (10 GB lub więcej) wymagana jest taka sama ilość wolnej przestrzeni na dysku. Do ustawienia lokalizacji katalogu tymczasowego można wykorzystać zmienną środowiskową SPSSTMPDIR.

Opcje modelu z węzłem logistycznym

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Utwórz modele rozdzielone. Tworzy osobny model dla każdej możliwej wartości zmiennych wejściowych, jakie są określone jako zmienne podziału. Więcej informacji można znaleźć w “Budowanie modeli rozdzielonych” na stronie 28.

Procedura. Określa, czy tworzony jest węzeł dwumianowy, czy wielomianowy. Opcje dostępne w oknie dialogowym różnią się w zależności od typu wybranej procedury modelowania.

- **Dwumianowy.** Stosowany, gdy zmienna przewidywana jest typu Flaga lub nominalna z dwiema wartościami dyskretnymi (dychotomiczna), np. *yes*/*no*, *on*/*off*, *male*/*female*.
- **Wielomianowy.** Stosowany, gdy zmienna przewidywana jest zmienną nominalną z więcej niż dwiema wartościami. Można określić **Efekty główne**, **Pełny czynnikowy** lub **Użytkownika**.

Uwzględnij stałą w równaniu. Ta opcja określa, czy równania wynikowe będą zawierały składnik stały. W większości sytuacji należy pozostawić tę opcję zaznaczoną.

Modele dwumianowe

W przypadku modeli dwumianowych dostępne są następujące metody i opcje:

Metoda. Określ metodę, która będzie używana podczas budowania modelu regresji logistycznej.

- **Wprowadzanie.** Jest to metoda domyślna, która wprowadza wszystkie składniki bezpośrednio do równania. Podczas budowania modelu nie jest wykonywany wybór zmiennych.
- **Krokowa postępująca.** Metoda Krokowa postępująca wyboru zmiennych buduje równanie w sposób krokowy, jak wskazuje jej nazwa. Początkowy model jest najprostszym możliwym modelem bez składników (z wyjątkiem stałej) w równaniu. Na każdym kroku składniki, które nie zostały jeszcze dodane do modelu, są oceniane, a jeśli najlepszy składnik znacząco zwiększa jakość predykcji modelu, jest dodawany. Ponadto składniki, które aktualnie znajdują się w modelu, są ponownie oceniane w celu ustalenia, czy dowolny z nich może zostać usunięty bez znaczącego pogorszenia jakości modelu. Jeśli tak jest, składniki są usuwane. Proces jest powtarzany, a inne składniki są

dodawane i/lub usuwane. Jeśli nie można już dodać składników w celu poprawy modelu ani nie można usunąć żadnych składników bez pogorszenia jakości modelu, tworzony jest ostateczny model.

- **Krokowa wsteczna.** Metoda Krokowa wsteczna jest przeciwieństwem metody Krokowa postępująca. W przypadku tej metody model początkowy zawiera wszystkie składniki jako predyktory. Na każdym kroku składniki w modelu są oceniane, a dowolne składniki, które mogą być usuwane bez znaczącego umniejszania wartości modelu, są usuwane. Ponadto usunięte wcześniej składniki są oceniane ponownie, aby określić, czy najlepszy z nich powoduje znaczące zwiększenie jakości predykcji modelu. Jeśli tak jest, są dodawane ponownie do modelu. Jeśli nie można usunąć więcej składników bez znaczącego zmniejszenia wartości modelu i nie można dodać żadnych składników w celu poprawy modelu, tworzony jest ostateczny model.

Wejścia jakościowe. Lista zmiennych zidentyfikowanych jako jakościowe, czyli takie, w których poziom pomiaru jest typu flaga, nominalny lub porządkowy. Dla każdej zmiennej jakościowej można określić kontrast i kategorię odniesienia.

- **Nazwa zmiennej.** Ta kolumna zawiera nazwy wejściowych zmiennych jakościowych. W celu dodania do tej kolumny wejść ilościowych lub numerycznych kliknij ikonę Dodaj zmienne, która znajduje się po prawej stronie listy. Następnie wybierz żądane wejścia.
- **Kontrast.** Interpretacja współczynników regresji dla zmiennej jakościowej jest zależna od używanych kontrastów. Kontrast określa konfigurację testów hipotezy w celu porównania oszacowanych średnich. Na przykład, jeśli wiadomo, że zmienna jakościowa ma porządek domniemany, np. wzorzec lub grupowanie, wówczas można użyć kontrastu do zamodelowania tego porządku. Dostępne są następujące kontrasty:

  Wskaźnik. Kontrasty wskazują na obecność lub nieobecność kategorii. Jest to metoda domyślna.

  Prosty. Każda kategoria zmiennej predykcyjnej, z wyjątkiem kategorii odniesienia, jest porównywana do kategorii odniesienia.

  Różnica. Każda kategoria zmiennej predykcyjnej, z wyjątkiem pierwszej kategorii, jest porównywana do efektu średniego poprzednich kategorii. Kontrasty takie są określane również mianem odwrotnego kontrastu Helmerta.

  Helmerta. Każda kategoria zmiennej predykcyjnej, z wyjątkiem ostatniej kategorii, jest porównywana do efektu średniego kolejnych kategorii.

  Powtórzony. Każda kategoria zmiennej predykcyjnej, z wyjątkiem pierwszej kategorii, jest porównywana z kategorią, która ją poprzedza.

  Wielomianowy. Kontrasty wielomianowe ortogonalne. Zakłada się, że kategorie są równo rozłożone. Kontrasty wielomianowe są dostępne tylko w przypadku zmiennych numerycznych.

  Odchylenie. Każda kategoria zmiennej predykcyjnej, z wyjątkiem kategorii odniesienia, jest porównywana do efektu ogólnego.
- **Kategoria odniesienia.** Określa sposób ustalania kategorii odniesienia dla wybranego typu kontrastu. Wybierz opcję **Pierwsze**, aby użyć pierwszej kategorii dla zmiennej wejściowej — posortowanej alfabetycznie — albo wybierz opcję **Ostatnie**, aby użyć ostatniej kategorii. Domyślna kategoria bazowa ma zastosowanie do zmiennych wymienionych w obszarze **Wejścia jakościowe**.

  Uwaga: To pole jest niedostępne, jeśli kontrast jest ustawiony na wartość Różnica, Helmerta, Powtórzony lub Wielomianowy.

Oszacowanie efektu każdej zmiennej na odpowiedź ogólną jest obliczane jako wzrost lub spadek wiarygodności każdej innej kategorii w porównaniu do kategorii odniesienia. Może to ułatwić identyfikowanie zmiennych i wartości, które wykazują większą tendencję do konkretnej odpowiedzi.

Kategoria odniesienia jest przedstawiona w wynikach jako 0,0. Jest to spowodowane tym, że porównanie jej z nią samą zwraca wynik pusty. Wszystkie inne kategorie są przedstawione jako równania istotne dla kategorii odniesienia. Więcej informacji można znaleźć w temacie “Szczegóły modelu użytkowego Logistyczny” na stronie 188.

Modele wielomianowe

W przypadku modeli wielomianowych dostępne są następujące metody i opcje:

Metoda. Określ metodę, która będzie używana podczas budowania modelu regresji logistycznej.

- **Wprowadzanie.** Jest to metoda domyślna, która wprowadza wszystkie składniki bezpośrednio do równania. Podczas budowania modelu nie jest wykonywany wybór zmiennych.
- **Krokowa.** Metoda Krokowa wyboru zmiennych buduje równanie w sposób krokowy, jak wskazuje jej nazwa. Początkowy model jest najprostszym możliwym modelem bez składników (z wyjątkiem stałej) w równaniu. Na każdym kroku składniki, które nie zostały jeszcze dodane do modelu, są oceniane, a jeśli najlepszy składnik znacząco zwiększa jakość predykcji modelu, jest dodawany. Ponadto składniki, które aktualnie znajdują się w modelu, są ponownie oceniane w celu ustalenia, czy dowolny z nich może zostać usunięty bez znaczącego pogorszenia jakości modelu. Jeśli tak jest, składniki są usuwane. Proces jest powtarzany, a inne składniki są dodawane i/lub usuwane. Jeśli nie można już dodać składników w celu poprawy modelu ani nie można usunąć żadnych składników bez pogorszenia jakości modelu, tworzony jest ostateczny model.
- **Postępująca.** Metoda Postępująca wyboru zmiennych działa podobnie do metody krokowej, ponieważ model jest budowany krokowo. Jednak w przypadku tej metody początkowy model jest najprostszy i do niego może być dodawana tylko stała oraz składniki. W każdym kroku składniki, które nie znajdują się jeszcze w modelu, są testowane pod kątem tego, w jakim stopniu mogłyby poprawić model, a najlepszy z tych składników jest dodawany do modelu. Gdy nie można dodać już żadnych składników albo najlepszy składnik kandydacki nie powoduje znaczącej poprawy modelu, tworzony jest ostateczny model.
- **Eliminacja wsteczna.** Metoda Eliminacja wsteczna jest przeciwieństwem metody Krokowa. W tej metodzie model początkowy zawiera wszystkie składniki jako predyktory, a składniki mogą być tylko usuwane z modelu. Składniki modelu, które w niewielkim stopniu przyczyniają się do zmiany jakości modelu, są usuwane pojedynczo, aż do momentu, gdy nie można usunąć składników bez znaczącego pogorszenia modelu. Model w takim stanie jest ostateczny.
- **Krokowa wsteczna.** Metoda Krokowa wsteczna jest przeciwieństwem metody Krokowa. W przypadku tej metody model początkowy zawiera wszystkie składniki jako predyktory. Na każdym kroku składniki w modelu są oceniane, a dowolne składniki, które mogą być usuwane bez znaczącego umniejszania wartości modelu, są usuwane. Ponadto usunięte wcześniej składniki są oceniane ponownie, aby określić, czy najlepszy z nich powoduje znaczące zwiększenie jakości predykcji modelu. Jeśli tak jest, są dodawane ponownie do modelu. Jeśli nie można usunąć więcej składników bez znaczącego zmniejszenia wartości modelu i nie można dodać żadnych składników w celu poprawy modelu, tworzony jest ostateczny model.

Uwaga: Metody automatyczne, takie jak Krokowa, Postępująca i Eliminacja wsteczna, to wysoce adaptacyjne metody uczenia, które wykazują silną tendencję do przeuczania na podstawie danych uczących. Gdy te metody są stosowane, szczególnie ważne jest sprawdzenie poprawności modelu wynikowego z użyciem nowych danych albo próby testowej utworzonej z użyciem węzła podziału na podzbiory.

Kategoria odniesienia dla przewidywanej. Określa sposób ustalania kategorii odniesienia. Ta kategoria jest stosowana jako odniesienie, względem którego szacowane są równania regresji dla wszystkich innych kategorii. Wybierz opcję **Pierwsze**, aby użyć pierwszej kategorii dla bieżącej zmiennej przewidywanej — posortowanej alfabetycznie — albo wybierz opcję **Ostatnie**, aby użyć ostatniej kategorii. Można również wybrać opcję **Określ**, aby wybrać konkretną kategorię i wybrać żądaną wartość z listy. Dostępne wartości mogą być definiowane dla każdej zmiennej w węźle Typ.

Często użytkownicy wybierają kategorię, która najmniej interesuje ich jako kategoria odniesienia, np. produkt, który stracił pozycję lidera rynku. Następnie pozostałe kategorie są prezentowane w odniesieniu do tej kategorii odniesienia w sposób względny w celu zidentyfikowania tego, co może zwiększyć prawdopodobieństwo tego, że będą należeć do własnych kategorii. Może to ułatwić identyfikowanie zmiennych i wartości, które wykazują większą tendencję do konkretnej odpowiedzi.

Kategoria odniesienia jest przedstawiona w wynikach jako 0,0. Jest to spowodowane tym, że porównanie jej z nią samą zwraca wynik pusty. Wszystkie inne kategorie są przedstawione jako równania istotne dla kategorii odniesienia. Więcej informacji można znaleźć w temacie “Szczegóły modelu użytkowego Logistyczny” na stronie 188.

Typ modelu. Istnieją trzy opcje przeznaczone do zdefiniowania składników w modelu. Modele **efektów głównych** obejmują tylko pojedyncze zmienne wejściowe i nie testują interakcji (efektów multiplikatywnych) między zmiennymi wejściowymi. Modele **Pełny czynnikowy** obejmują wszystkie interakcje, a także efekty główne zmiennej wejściowej. Modele Pełny czynnikowy lepiej przechwytują złożone relacje, ale ich interpretowanie jest dużo trudniejsze i w ich przypadku częściej dochodzi do przeuczenia. Z powodu potencjalnie dużej liczby możliwych kombinacji metody automatycznego wyboru zmiennej (metody inne niż Wprowadzanie) są wyłączone dla modeli Pełny czynnikowy. Modele **użytkownika** uwzględniają tylko składniki (efekty główne i interakcje) określone przez użytkownika. W przypadku wyboru tej opcji należy użyć listy Składniki modelu, aby dodać lub usunąć składniki w modelu.

Składniki modelu. W przypadku budowania modelu Użytkownika konieczne będzie jawne określenie składników w modelu. Ta lista przedstawia bieżący zestaw składników dla modelu. Przyciski po prawej stronie listy Składniki modelu umożliwiają dodawanie i usuwanie składników modelu.

- Aby dodać składniki do modelu, kliknij przycisk *Dodaje nowe składniki modelu*.
- W celu usuwania składników należy wybrać żądane składniki i kliknąć przycisk *Usuwa wybrane składniki modelu*.

Dodawanie składników do modelu regresji logistycznej

W przypadku żądania niestandardowego modelu regresji logistycznej można dodawać składniki do modelu, klikając przycisk *Dodaje nowe składniki modelu* na karcie Model regresji logistycznej. Zostanie otwarte okno dialogowe Nowe składniki, w którym można określać składniki.

Typ dodawanego składnika. Istnieje kilka sposobów dodawania składników do modelu w zależności od wybranych zmiennych wejściowych na liście Dostępne zmienne.

- **Interakcja jednostkowa.** Umożliwia wstawienie składnika reprezentującego interakcję wszystkich wybranych zmiennych.
- **Efekty główne.** Wstawia jeden składnik efektu głównego (samą zmienną) dla każdej wybranej zmiennej wejściowej.
- **Wszystkie interakcje 2. rzędu.** Umożliwia wstawienie składnika interakcji 2. rzędu (jest to iloczyn zmiennych wejściowych) dla każdej możliwej pary wybranych zmiennych wejściowych. Na przykład, jeśli wybierzesz zmienne wejściowe *A*, *B*, oraz *C* na liście Dostępne zmienne, ta metoda spowoduje wstawienie składników *A * B*, *A * C* oraz *B * C*.
- **Wszystkie interakcje 3. rzędu.** Umożliwia wstawienie składnika interakcji 3. rzędu (jest to iloczyn zmiennych wejściowych) dla każdej możliwej kombinacji wybranych zmiennych wejściowych, pobieranych po trzy jednocześnie. Na przykład, jeśli wybierzesz zmienne wejściowe *A*, *B*, *C* oraz *D* na liście Dostępne zmienne, ta metoda spowoduje wstawienie składników *A * B * C*, *A * B * D*, *A * C * D* oraz *B * C * D*.
- **Wszystkie interakcje 4. rzędu.** Umożliwia wstawienie składnika interakcji 4. rzędu (jest to iloczyn zmiennych wejściowych) dla każdej możliwej kombinacji wybranych zmiennych wejściowych, pobieranych po cztery jednocześnie. Na przykład, jeśli wybierzesz zmienne wejściowe *A*, *B*, *C*, *D* oraz *E* na liście Dostępne zmienne, ta metoda spowoduje wstawienie składników *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* oraz *B * C * D * E*.

Dostępne zmienne. Ta opcja przedstawia listę dostępnych zmiennych wejściowych, która może być używana w celu konstruowania składników modelu.

Podgląd. Przedstawia składniki, które zostaną dodane do modelu w przypadku kliknięcia opcji **Wstaw** — odpowiednio do wybranych zmiennych i typów składników.

Wstaw. Umożliwia wstawienie składników do modelu (na podstawie aktualnie wybranych zmiennych oraz typu składnika) i powoduje zamknięcie okna dialogowego.

Opcje zaawansowane węzła logistycznego

Użytkownikom posiadającym szczegółową wiedzę na temat regresji logistycznej opcje zaawansowane umożliwiają precyzyjne dostosowywanie procesu uczenia. W celu uzyskania dostępu do opcji zaawansowanych należy ustawić opcję Tryb na wartość **Zaawansowany** na karcie Zaawansowany.

Skala (tylko modele wielomianowe). Można określić wartość skalowania dyspersji, która będzie używana w celu skorygowania oszacowania macierzy kowariancji parametrów. Wartość **Pearsona** umożliwia oszacowanie wartości skali przy użyciu statystyki chi-kwadrat Pearsona. Wartość **Odchylenie** umożliwia oszacowanie wartości skali przy użyciu statystyki funkcji odchylenia (wskaźnik chi-kwadrat wiarygodności). Można także określić własną wartość skalowania. Musi to być dodatnia wartość numeryczna.

Dołącz wszystkie prawdopodobieństwa. W przypadku wyboru tej opcji prawdopodobieństwa dla poszczególnych kategorii zmiennych wyjściowych będą dodawane do poszczególnych rekordów przetwarzanych przez węzeł. Jeśli ta opcja nie zostanie wybrana, wówczas zostanie dodane tylko prawdopodobieństwo przewidywanej kategorii.

Na przykład tabela zawierająca wyniki dla modelu wielomianowego z trzema kategoriami będzie zawierać pięć nowych kolumn. Jedna kolumna będzie zawierać prawdopodobieństwo poprawnego przewidzenia wyniku, następna prawdopodobieństwo tego, że predykcja jest trafiona lub chybiona, a dalsze kolumny będą przedstawiać prawdopodobieństwo tego, że predykcja każdej kategorii jest chybiona lub trafiona. Więcej informacji można znaleźć w temacie “Model użytkowy modelu logistycznego” na stronie 187.

Uwaga: Ta opcja jest zawsze wybierana w przypadku modeli dwumianowych.

Tolerancja osobliwości. Należy określić tolerancję używaną kontroli osobliwości.

Zbieżność. Te opcje umożliwiają sterowanie parametrami zbieżności modelu. Podczas wykonywania modelu ustawienia zbieżności kontrolują liczbę powtórzonych uruchomień różnych parametrów w celu sprawdzenia ich dopasowania. Im częściej parametry są wypróbowywane, tym bliższe będą wyniki (co oznacza, że wyniki uzyskają zbieżność). Więcej informacji można znaleźć w temacie “Opcje zbieżności regresji logistycznej”.

Wynik. Te opcje umożliwiają zażądanie dodatkowych statystyk, które będą wyświetlane w zaawansowanych wynikach modelu użytkowego budowanego przez węzeł. Więcej informacji można znaleźć w temacie “Zaawansowane wyniki regresji logistycznej” na stronie 186.

Krokowa. Te opcje umożliwiają określanie kryteriów dodawania i usuwania zmiennych w przypadku metod estymacji Krokowa, Postępująca, Eliminacja wsteczna oraz Krokowa wsteczna. (Przycisk jest wyłączony, jeśli wybrana jest metoda Wprowadzanie). Więcej informacji można znaleźć w temacie “Opcje metody krokowej regresji logistycznej” na stronie 186.

Opcje zbieżności regresji logistycznej

Parametry zbieżności można ustawić dla estymacji modelu regresji logistycznej.

Maksymalna liczba iteracji. Umożliwia określenie maksymalnej liczby iteracji na potrzeby estymacji modelu.

Maksimum kroków połowienia. Kroki połowienia to technika stosowana przez regresję logistyczną w przypadkach złożoności procesu estymacji. W normalnych okolicznościach wystarczy używać ustawienia domyślnego.

Zbieżność logarytmu wiarygodności. Iteracje zatrzymują się, gdy logarytm wiarygodności jest niższy niż ta wartość. Kryterium nie jest stosowane, jeśli wartość wynosi 0.

Zbieżność parametru. Iteracje zatrzymują się, jeśli zmiana bezwzględna lub względna w estymacjach parametru jest mniejsza niż ta wartość. Kryterium nie jest stosowane, jeśli wartość wynosi 0.

Delta (tylko modele wielomianowe). Można określić, że do każdej pustej komórki (kombinacja wartości zmiennej wejściowej i zmiennej wyjściowej) dodawana będzie wartość z zakresu od 0 do 1. Dzięki temu algorytm estymacji może przetwarzać dane, w których istnieje wiele możliwych kombinacji wartości zmiennych w porównaniu do liczby rekordów w danych. Domyślną wartością jest 0.

Zaawansowane wyniki regresji logistycznej

Należy wybrać opcjonalne wyniki, które będą wyświetlane w obszarze wyników zaawansowanych dla modelu użytkowego Regresja. W celu wyświetlenia zaawansowanych wyników przejdź do modelu użytkowego i kliknij kartę **Zaawansowane**. Więcej informacji można znaleźć w temacie “Zaawansowane wyniki modelu użytkowego Logistyczny” na stronie 190.

Opcje dwumianowe

Należy wybrać typ wyników generowanych dla modelu. Więcej informacji można znaleźć w temacie “Zaawansowane wyniki modelu użytkowego Logistyczny” na stronie 190.

Pokaż. Określ, czy wyniki będą wyświetlane przy każdym kroku, czy dopiero po przejściu wszystkich kroków.

CI dla exp(B) (%). Wybierz przedziały ufności dla każdego współczynnika (przedstawionego jako Beta) w wyrażeniu. Określ poziom przedziału ufności (domyślnie jest to 95%).

Diagnoza reszt. Umożliwia zażądanie tabeli diagnostyki obserwacji reszt.

- **Wartości odstające (odch. std.).** Wyświetlana jest tylko lista obserwacji resztowych, dla których bezwzględna wartość standaryzowana zmiennej z listy jest co najmniej tak duża, jak wartość określona przez użytkownika. Domyślna wartość to 2.
- **Wszystkie obserwacje.** W tabeli diagnostyki obserwacji reszt uwzględniane są wszystkie obserwacje.

  Uwaga: ta opcja powoduje wyświetlenie każdego z rekordów wejściowych, dlatego może spowodować zwrócenie bardzo dużej tabeli w raporcie, przy czym jedna linia odpowiada jednemu rekordowi.

Punkt odcięcia klasyfikacji. Ta opcja umożliwia określenie punktu odcięcia dla klasyfikowania obserwacji. Obserwacje z wartościami przewidywanymi, które przekraczają punkt odcięcia klasyfikacji, są klasyfikowane jako dodatnie, a wartości poniżej punktu odcięcia są klasyfikowane jako ujemne. W celu zmiany ustawienia domyślnego należy wprowadzić wartość zakresu od 0,01 do 0,99.

Opcje wielomianowe

Należy wybrać typ wyników generowanych dla modelu. Więcej informacji można znaleźć w temacie “Zaawansowane wyniki modelu użytkowego Logistyczny” na stronie 190.

Uwaga: wybranie opcji **Testy ilorazu wiarygodności** spowoduje znaczące wydłużenie czasu przetwarzania wymaganego do zbudowania modelu regresji logistycznej. Jeśli budowanie modelu trwa zbyt długo, należy rozważyć wyłączenie tej opcji albo użycie zamiast niej statystyki Walda i statystyki ocen. Więcej informacji można znaleźć w temacie “Opcje metody krokowej regresji logistycznej”.

Przebieg iteracji dla każdego. Wybierz interwał kroków, po którym przedstawiany będzie status iteracji w wynikach zaawansowanych.

Oszacowanie przedziału ufności. Przedziały ufności dla współczynników w równaniach. Określ poziom przedziału ufności (domyślnie jest to 95%).

Opcje metody krokowej regresji logistycznej

Te opcje umożliwiają określanie kryteriów dodawania i usuwania zmiennych w przypadku metod estymacji Krokowa, Postępująca, Eliminacja wsteczna oraz Krokowa wsteczna.

Liczba składników w modelu (tylko modele wielomianowe). Można określić maksymalną liczbę składników dla modeli Eliminacja wsteczna i Krokowa wsteczna, a także maksymalną liczbę składników dla modeli Postępująca i Krokowa. Jeśli zostanie określona wartość minimalna większa od 0, wówczas model będzie zawierał wiele składników, nawet jeśli na podstawie kryteriów statystycznych niektóre z tych składników zostałyby usunięte. Ustawienie minimum jest ignorowane w przypadku modeli Postępująca, Krokowa i Wprowadzanie. Jeśli zostanie określone maksimum,

niektóre składniki mogą być pomijane, nawet jeśli na podstawie kryteriów statystycznych zostałyby wybrane. Ustawienie **Określ maksimum** jest ignorowane w przypadku modeli Eliminacja wsteczna, Krokowa wsteczna oraz Wprowadzanie.

Kryterium wprowadzenia (tylko modele wielomianowe). Aby maksymalnie przyspieszyć przetwarzanie, wybierz opcję **Ocena**. Opcja **Iloraz wiarygodności** może udostępniać trochę bardziej odporne oszacowania, ale jej obliczenie trwa dłużej. Ustawieniem domyślnym jest użycie statystyki Ocena.

Kryterium usuwania. W celu uzyskania bardziej odpornego modelu należy wybrać opcję **Iloraz wiarygodności**. Aby skrócić czas wymagany do zbudowania modelu, można spróbować użyć opcję **Walda**. Jeśli jednak w danych występuje pełny lub quasi-pełny podział (co można określić poprzez użycie karty Zaawansowane w modelu użytkowym), wówczas statystyka Walda stanie się szczególnie nierzetelna i nie powinna być używana. Ustawieniem domyślnym jest użycie statystyki ilorazu wiarygodności. W przypadku modeli dwumianowych istnieje dodatkowa opcja **Warunkowe**. Ta opcja zapewnia testowanie usuwania na podstawie prawdopodobieństwa ilorazu wiarygodności wyliczonego na podstawie ocen parametrów warunkowych.

Wartości graniczne istotności dla kryteriów. Ta opcja umożliwia określenie kryteriów wyboru na podstawie prawdopodobieństwa statystycznego (wartości *p*) skojarzonego z poszczególnymi zmiennymi. Zmienne będą dodawane do modelu, pod warunkiem że powiązana wartość *p* jest mniejsza niż wartość **Wprowadzanie** i będą usuwane, jeśli wartość *p* jest większa niż wartość **Usunięcie**. Wartość **Wprowadzanie** musi być mniejsza niż wartość **Usunięcie**.

Wymagania dla wprowadzenia lub usunięcia (tylko modele wielomianowe). W przypadku niektórych zastosowań dodawanie składników interakcji do modelu nie ma sensu matematycznego, chyba że model zawiera również składniki niższego rzędu dla zmiennych stanowiących część składnika interakcji. Na przykład dodanie *A * B* do modelu może nie mieć sensu, chyba że *A* i *B* są również zawarte w modelu. Te opcje umożliwiają określenie sposobu postępowania w przypadku takich zależności podczas wyboru składników krokowych.

- **Hierarchia efektów czynnikowych.** Efekty wyższego rzędu (interakcje obejmujące większą liczbę zmiennych) zostaną wprowadzone do modelu, pod warunkiem że wszystkie efekty niższego rzędu (efekty główne lub interakcje obejmujące mniej zmiennych) dla istotnych zmiennych znajdują się już w modelu, a efekty niższego rzędu nie zostaną usunięte, jeśli efekty wyższego rzędu obejmujące te same zmienne znajdują się w modelu. Ta opcja ma zastosowanie tylko do zmiennych jakościowych.
- **Hierarchia wszystkich efektów.** Ta opcja działa w taki sam sposób, jak poprzednia opcja, ale obowiązuje w przypadku wszystkich zmiennych wejściowych.
- **Zawieranie się dla wszystkich efektów.** Efekty mogą być zawarte w modelu tylko wówczas, gdy wszystkie efekty zawarte w efekcie są również zawarte w modelu. Ta opcja jest podobna do opcji **Hierarchia wszystkich efektów**, ale zmienne ilościowe są traktowane trochę inaczej. Efekt może zawierać inny efekt, pod warunkiem że zawierany efekt (niższego rzędu) zawiera *wszystkie* zmienne ilościowe zawarte w efekcie zawierającym (wyższego rzędu), a zmienne ilościowe efektu zawartego muszą stanowić podzbiór tych efektów w efekcie zawierającym. Na przykład, jeśli *A* i *B* są zmiennymi jakościowymi, a *X* jest zmienną ilościową, wówczas składnik *A * B * X* zawiera składniki *A * X* oraz *B * X*.
- **Brak.** Żadne powiązania nie są wymuszane; składniki są dodawane i usuwane z modelu niezależnie.

Model użytkowy modelu logistycznego

Model użytkowy modelu logistycznego reprezentuje równanie estymowane przez węzeł logistyczny. Zawiera wszystkie informacje przechwytywane przez model regresji logistycznej, a także informacje dotyczące struktury i wydajności modelu. Ten typ równania może być również generowany przez inne modele, na przykład Oracle SVM.

W przypadku uruchomienia strumienia zawierającego model użytkowy modelu logistycznego węzeł dodaje dwie nowe zmienne zawierające predykcję modelu i powiązane prawdopodobieństwo. Nazwy nowych zmiennych pochodzą od nazwy przewidywanej zmiennej wyjściowej. Nazwy przewidywanych kategorii są poprzedzone znakami *$L-*, a nazwy powiązanych prawdopodobieństw są poprzedzone znakami *$LP-*. Na przykład w przypadku zmiennej wyjściowej o nazwie *colorpref* nowe zmienne będą miały nazwy *$L-colorpref* oraz *$LP-colorpref*. Jeśli dodatkowo wybrano opcję **Dołącz wszystkie prawdopodobieństwa** w węźle Logistyczne, wówczas dodatkowa zmienna zostanie dodana do

każdej kategorii zmiennej wyjściowej i będzie zawierać prawdopodobieństwo należące do odpowiedniej kategorii każdego rekordu. Te dodatkowe zmienne otrzymają nazwy na podstawie wartości zmiennej wynikowej z przedrostkiem *$LP-*. Na przykład, jeśli prawidłowe wartości *colorpref* to *Red*, *Green* i *Blue*, wówczas zostaną dodane trzy nowe zmienne: *$LP-Red*, *$LP-Green* oraz *$LP-Blue*.

Generowanie węzła filtrowania. Menu Utwórz umożliwia utworzenie nowego węzła filtrowania, który będzie przekazywany do zmiennych wejściowych na podstawie wyników modelu. Zmienne usuwane z modelu z powodu wielowspółliniowości będą filtrowane przez wygenerowany węzeł — tak samo, jak zmienne nieużywane w modelu.

Szczegóły modelu użytkowego Logistyczny

W przypadku modeli wielomianowych karta Model w modelu użytkowym Logistyczne jest podzielona. W panelu po lewej stronie przedstawione są równania modelu, a po prawej stronie ważność predyktora. W przypadku modeli dwumianowych ta karta przedstawia tylko ważność predyktora. Więcej informacji można znaleźć w temacie “Ważność predyktorów” na stronie 43.

Równania modelu

W przypadku modeli wielomianowych w lewym panelu wyświetlane są rzeczywiste równania podlegające ocenie dla modelu regresji logistycznej. Istnieje jedno równanie dla każdej kategorii w zmiennej przewidywanej z wyjątkiem kategorii bazowej. Równania są wyświetlane w formacie drzewa. Ten typ równania może być również generowany przez niektóre inne modele, na przykład Oracle SVM.

Równanie dla. Powoduje przedstawienie równań regresji używanych do uzyskania prawdopodobieństw kategorii zmiennej przewidywanej, na podstawie zestawu wartości predykcyjnych. Ostatnia kategoria zmiennej przewidywanej jest traktowana jako **kategoria bazowa**; przedstawione równania określają logarytm szans dla innych kategorii zmiennej przewidywanej w porównaniu do kategorii bazowej dla konkretnego zestawu wartości predykcyjnych. Przewidywane prawdopodobieństwo dla każdej kategorii danego wzorca predyktora jest uzyskiwane z tych wartości logarytmu szans.

Sposób obliczania prawdopodobieństw

Każde równanie oblicza logarytm szans dla konkretnej kategorii zmiennej przewidywanej względem kategorii bazowej. **Logarytm szans** określany również nazwą **logit** to współczynnik prawdopodobieństwa podanej kategorii zmiennej przewidywanej do kategorii bazowej, przy czym względem wyniku stosowana jest funkcja logarytmu naturalnego. W przypadku kategorii bazowej szanse kategorii w odniesieniu do niej samej wynoszą 1,0 i z tego względu logarytm szans wynosi 0. Można to traktować jako domniemane równanie dla kategorii bazowej, w której wszystkie współczynniki wynoszą 0.

W celu uzyskania prawdopodobieństwa z logarytmu szans dla konkretnej kategorii zmiennej przewidywanej należy przyjąć wartość logit obliczoną przez równanie dla tej kategorii i zastosować poniższą formułę:

$$P(group_i) = \exp(g_i) / \Sigma_k \exp(g_k)$$

gdzie *g* jest obliczonym logarytmem szans, *i* jest indeksem kategorii, a *k* przyjmuje wartości od 1 do liczby kategorii zmiennej przewidywanej.

Ważność predyktorów

Opcjonalnie na karcie Model może być również wyświetlany wykres przedstawiający względną ważność poszczególnych predyktorów w oszacowaniu modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Należy zwrócić uwagę, że ten wykres jest dostępny tylko po wybraniu opcji **Oblicz ważność predyktora** na karcie Analiza przed wygenerowaniem modelu. Więcej informacji można znaleźć w temacie “Ważność predyktorów” na stronie 43.

Uwaga: obliczenie ważności predyktora może trwać dłużej w przypadku regresji logistycznej niż w przypadku modelów innych typów i nie jest domyślnie wybrane na karcie Analiza. Wybranie tej opcji może spowodować spowolnienie działania, szczególnie w przypadku dużych zbiorów danych.

Podsumowanie modelu użytkowego Logistyczny

W podsumowaniu modelu regresji logistycznej wyświetlane są zmienne i ustawienia służące do wygenerowania modelu. Ponadto jeśli wykonano węzeł analizy dołączony do tego węzła modelowania, informacje z tej analizy również będą wyświetlane w tej sekcji. Ogólne informacje na temat korzystania z przeglądarki modelu zawiera sekcja “Przeglądanie modeli użytkowych” na stronie 42.

Ustawienia modelu użytkowego Logistyczny

Karta Ustawienia w modelu użytkowym Logistyczny określa opcje dotyczące ufności, prawdopodobieństw, ocen skłonności oraz generowania kodu SQL podczas oceniania modelu. Ta karta jest dostępna tylko po dodaniu modelu użytkowego do strumienia i zawiera różne opcje w zależności od typu modelu i typu zmiennej przewidywanej.

Modele wielomianowe

W przypadku modeli wielomianowych dostępne są następujące opcje.

Wylicz ufności Określa, czy ufności są obliczane podczas oceniania.

Wylicz surowe oceny skłonności (tylko dla przewidywanych zmiennych typu flaga) W przypadku modeli zawierających tylko zmienne przewidywane typu flaga można zażądać surowych ocen skłonności, które wskazują wiarygodność wyniku prawda określonego dla zmiennej przewidywanej. Te oceny są stosowane dodatkowo obok standardowych wartości predykcji i ufności. Skorygowane oceny skłonności są niedostępne. Więcej informacji można znaleźć w temacie “Opcje analizowania węzła modelowania” na stronie 34.

Dołącz wszystkie prawdopodobieństwa Określa, czy prawdopodobieństwa dla poszczególnych kategorii zmiennych wyjściowych są dodawane do poszczególnych rekordów przetwarzanych przez węzeł. Jeśli ta opcja nie zostanie wybrana, wówczas zostanie dodane tylko prawdopodobieństwo przewidywanej kategorii. Na przykład w przypadku nominalnej zmiennej przewidywanej z trzema kategoriami wynik oceny będzie zawierał kolumnę dla każdej z trzech kategorii plus czwarta kolumna wskazująca prawdopodobieństwo dowolnej przewidywanej kategorii. Na przykład, jeśli prawdopodobieństwa kategorii *Czerwone*, *Zielone* i *Niebieskie* wynoszą odpowiednio 0,6; 0,3 oraz 0,1, wówczas przewidywaną kategorią będzie *Czerwone* z prawdopodobieństwem 0,6.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę, wykorzystując natywny kod SQL** Jeśli ta opcja jest wybrana, generowany jest natywny kod SQL w celu oceny modelu w bazie danych.

  Uwaga: Ta opcja może szybciej zwracać wyniki, ale rozmiar i złożoność natywnego kodu SQL wzrastają wraz ze wzrostem złożoności modelu.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Uwaga: W przypadku modeli wielomianowych generowanie kodu SQL jest niedostępne, jeśli wybrano opcję **Dołącz wszystkie prawdopodobieństwa**, lub — w przypadku modeli z nominalnymi zmiennymi przewidywanymi — jeśli wybrano opcję **Wylicz ufności**. Generowanie kodu SQL z obliczeniami ufności jest obsługiwane w przypadku modeli wielomianowych zawierających tylko zmienne przewidywane typu flaga. Generowanie kodu SQL jest niedostępne dla modeli dwumianowych.

Modele dwumianowe

W przypadku modeli dwumianowych ufności i prawdopodobieństwa są zawsze włączone, a ustawienia umożliwiające wyłączenie tych opcji są niedostępne. Generowanie kodu SQL jest niedostępne dla modeli dwumianowych. Jedynym ustawieniem, które można zmienić dla modeli dwumianowych, jest możliwość obliczania surowych ocen skłonności. Zgodnie z tym, co zostało wspomniane wcześniej w odniesieniu do modeli wielomianowych — taki sposób działania obowiązuje tylko w przypadku modeli ze zmiennymi przewidywanymi typu flaga. Więcej informacji można znaleźć w temacie “Opcje analizowania węzła modelowania” na stronie 34.

Zaawansowane wyniki modelu użytkowego Logistyczny

Zaawansowane wyniki dla regresji logistycznej (znane również jako **wielomianowa regresja logistyczna (NOMREG)**) przedstawiają szczegółowe informacje o szacowanym modelu i jego wydajności. Większość informacji zawartych w zaawansowanych wynikach ma charakter techniczny i w celu poprawnej interpretacji takich wyników wymagana jest rozległa wiedza na temat analiz regresji logistycznej.

Ostrzeżenia. Wskazuje wszelkie ostrzeżenia i potencjalne problemy z wynikami.

Podsumowanie przetwarzania przypadku. Zawiera listę przetworzonych rekordów podzieloną na poszczególne zmienne jakościowe w modelu.

Podsumowanie kroku (opcjonalnie). Zawiera listę efektów dodanych lub usuniętych na każdym kroku tworzenia modelu, gdy używany jest automatyczny wybór zmiennej.

Uwaga: Stosowane tylko w przypadku metod Krokowa, Postępująca, Wsteczna i Krokowa wsteczna.

Przebieg iteracji (opcjonalnie). Przedstawia przebieg iteracji oszacowań parametru dla każdych *n* iteracji począwszy od oszacowań początkowych, gdzie *n* jest wartością interwału przedstawiania. Domyślnie przedstawianie zachodzi co każdą iterację (n=1).

Dopasowanie modelu (modele wielomianowe). Przedstawia test ilorazu wiarygodności modelu (końcowy) w porównaniu z tym, w których wszystkie współczynniki parametru mają wartość 0 (tylko wyraz wolny).

Klasyfikacja (opcjonalnie). Przedstawia macierz przewidywanych i rzeczywistych wartości zmiennych przewidywanych z procentami.

Statystyki dobroci dopasowania chi-kwadrat (opcjonalnie). Przedstawia statystyki chi-kwadrat Pearsona oraz ilorazu wiarygodności chi-kwadrat. Te statystyki testują ogólne dopasowanie modelu do danych uczących.

Statystyka dobroci dopasowania Hosmera-Lemeshowa (opcjonalnie). Przedstawia wyniki grupowania obserwacji w decyle ryzyka i porównania w ramach każdego decylu prawdopodobieństwa obserwowanego z prawdopodobieństwem oczekiwanym. Statystyka dobroci dopasowania jest bardziej odporna niż tradycyjne statystyki dobroci dopasowania stosowane w modelach wielomianowych, szczególnie dla modeli z ilościowymi współzmiennymi oraz dla analiz małych prób.

Pseudo R-kwadrat (opcjonalnie). Przedstawia następujące miary dopasowania modelu: Cox i Snell, Nagelkerke oraz *R*-kwadrat McFaddena. Te statystyki są pod pewnymi względami analogiczne do statystyki *R*-kwadrat w regresji liniowej.

Miary monotoniczności (opcjonalnie). Przedstawia liczbę par zgodnych, par niezgodnych i par związanych w danych, a także procent łącznej liczby par, jakie te pary reprezentują. W tabeli widoczne są także Współczynnik Somers' D, Lambda Goodmana i Kruskala, Współczynnik Tau-a Kendalla oraz Indeks zgodności C.

Kryteria informacyjne (opcjonalnie). Przedstawia Kryterium informacyjne Akaike (AIC) oraz kryterium informacyjne Bayesowskie Schwarza (BIC).

Testy ilorazu wiarygodności (opcjonalnie). Przedstawia statystyki testujące, czy współczynniki efektów modelu są statystycznie różne od 0. Znaczące zmienne wejściowe to zmienne o bardzo niskich poziomach istotności w wynikach (oznakowane jako *Istotność*).

Oszacowania parametrów (opcjonalnie). Przedstawia oszacowania współczynników równania, testy tych współczynników, ilorazy szans bazujące na współczynnikach o etykietach *Exp(B)*, a także przedziały ufności dla ilorazów szans.

Asymptotyczna macierz kowariancji/korelacji (opcjonalnie). Przedstawia asymptotyczne kowariancje i/lub korelacje oszacowań współczynników.

Częstości obserwowane i przewidywane (opcjonalnie). Dla każdego układu współzmiennych przedstawia częstości obserwowane i przewidywane dla każdej wartości zmiennej wyjściowej. Ta tabela może być dosyć duża, szczególnie w przypadku modeli z liczbowymi zmiennymi wejściowymi. Jeśli tabela wynikowa jest zbyt duża i z tego powodu jest niepraktyczna, jest pomijana i wyświetlane jest ostrzeżenie.

Węzeł analizy PCA/czynnikowej

Węzeł analizy PCA/czynnikowej udostępnia wydajne techniki redukcji danych pozwalające obniżyć stopień ich złożoności. Udostępniono dwa podobne, jednak różniące się rozwiązania.

- **Analiza głównych składowych** (ang. Principal Components Analysis, PCA) znajduje kombinacje liniowe zmiennych wejściowych, które umożliwiają przechwytywanie wariancji w całym zestawie zmiennych, pod warunkiem że składowe są zlokalizowane ortogonalnie (prostopadle) do siebie. PCA skupia się na wszystkich wariancjach, z uwzględnieniem wariancji współużytkowanych i unikalnych.
- **Analiza czynnikowa** służy identyfikacji koncepcji lub **czynników**, które wyjaśniają wzory korelacji występujące w ramach zbiorów obserwowanych zmiennych. Analiza czynnikowa skupia się wyłącznie na wariancji współużytkowanej. Wariancja, która jest unikalna dla konkretnych zmiennych, nie jest uwzględniana w oszacowaniu modelu. Węzeł analizy PCA/czynnikowej udostępnia kilka metod przeprowadzania analizy czynnikowej.

W przypadku obu rozwiązań celem jest znalezienie niewielkiej liczby zmiennych pochodnych w efektywny sposób podsumowujących informacje w oryginalnym zestawie zmiennych.

Wymagania. W modelu PCA/czynnikowym mogą być używane tylko zmienne liczbowe. Aby oszacować analizę czynnikową lub PCA, wymagana jest co najmniej jedna zmienna z rolą ustawioną na zmienne *Input*. Zmienne z rolą ustawioną na *Target*, *Both* lub *None* są ignorowane, ponieważ nie są zmiennymi numerycznymi.

Mocne strony. Analiza czynnikowa i PCA może skutecznie zmniejszyć złożoność danych bez utraty znacznej treści informacji. Techniki te ułatwiają tworzenie bardziej solidnych modeli, które będą działały szybciej niż byłoby to możliwe w przypadku surowych zmiennych wejściowych.

Opcje modelu węzła analizy PCA/czynnikowej

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Metoda wyodrębniania. Określa metodę, jaka będzie użyta do redukcji danych.

- **Głównych składników.** Jest to metoda domyślna, używająca PCA do wyszukania składników, które podsumowują zmienne wejściowe.
- **Nieważonych najmniejszych kwadratów.** Ta metoda analizy czynnikowej działa na zasadzie wyszukiwania zestawu czynników, które najlepiej odtworzą wzorzec relacji (korelacji) dla zmiennych wejściowych.
- **Uogólnionych najmniejszych kwadratów.** Ta metoda analizy czynnikowej jest podobna do metody nieważonych najmniejszych kwadratów z takim wyjątkiem, że używa ważenia do zmniejszenia znaczenia zmiennych z wieloma unikalnymi (niewspólnymi) wariancjami.
- **Największej wiarygodności.** Ta metoda analizy czynnikowej tworzy równania czynnikowe, które z największym prawdopodobieństwem utworzyły obserwowany wzorzec relacji (korelacji) w zmiennych wejściowych w oparciu o założenia dotyczące formy tych relacji. W szczególności metoda ta zakłada, że dane uczące podlegają normalnemu rozkładowi wielu zmiennych.
- **Osi głównych.** Ta metoda analizy czynnikowej jest bardzo podobna do metody głównych składników z takim wyjątkiem, że skupia się tylko na wariancji współużytkowanej.
- **Czynnikowa Alfa.** Ta metoda analizy czynnikowej zakłada, że zmienne wykorzystane do analizy stanowią próbę z uniwersum potencjalnych zmiennych wejściowych. Maksymalizuje wartość rzetelności statystycznej czynników.
- **Czynnikowa obrazu.** Ta metoda analizy czynnikowej korzysta z oszacowania danych w celu wyizolowania wariancji wspólnej i wyszukania czynników, które ją opisują.

Zaawansowane opcje węzła analizy PCA/czynnikowej

W przypadku szczegółowej wiedzy na temat analizy czynnikowej i PCA opcje zaawansowane pozwolą na dostosowanie procesu uczenia. W celu uzyskania dostępu do opcji zaawansowanych należy ustawić opcję Tryb na wartość **Zaawansowany** na karcie Zaawansowany.

Braki danych. Domyślnie program IBM SPSS Modeler korzysta tylko z rekordów zawierających poprawne wartości dla wszystkich zmiennych użytych w modelu. (Niekiedy jest to zwane **usuwaniem obserwacjami** brakujących wartości). Jeśli istnieje dużo braków danych, może okazać się, że to rozwiązanie eliminuje zbyt wiele rekordów, a ilość pozostałych danych jest zbyt mała, aby wygenerować dobry model. W takich przypadkach można usunąć zaznaczenie opcji **Używaj tylko kompletnych rekordów**. Wówczas program IBM SPSS Modeler podejmie próbę użycia jak największej ilości informacji do oszacowania modelu, łącznie z rekordami, w których dla niektórych zmiennych istnieją braki danych. (Niekiedy jest to zwane **usuwaniem parami** brakujących wartości). Jednak w niektórych sytuacjach użycie niekompletnych rekordów w taki sposób może prowadzić do problemów obliczeniowych podczas szacowania modelu.

Pola. Należy określić, czy w oszacowaniu modelu ma być użyta macierz korelacji (domyślnie), czy też macierz kowariancji zmiennych wejściowych.

Maksimum iteracji dla uzyskania zbieżności. Umożliwia określenie maksymalnej liczby iteracji na potrzeby estymacji modelu.

Wyodrębniaj czynniki. Istnieją dwa sposoby wyboru liczby czynników do wyodrębnienia ze zmiennych wejściowych.

- **Wartości własne powyżej.** Ta opcja zachowa wszystkie czynniki lub składniki z wartościami własnymi większymi niż określone kryterium. Opcja **Wartości własne** mierzy zdolność każdego czynnika lub składnika do podsumowania wariancji w zbiorze zmiennych wejściowych. W przypadku użycia macierzy korelacji model zachowa wszystkie czynniki lub składniki z wartościami własnymi większymi niż określona wartość. Podczas użycia macierzy korelacji kryterium stanowi określona wartość razy średnia wartość własna. Takie skalowanie sprawia, że opcja ta ma podobne znaczenie dla obu typów macierzy.
- **Maksymalna liczba.** Ta opcja zachowa określoną liczbę czynników lub składników w kolejności malejącej wartości własnych. Inaczej mówiąc, zachowywane są czynniki lub składniki odpowiadające liczbie *n* najwyższych wartości własnych, gdzie *n* to określone kryterium. Domyślnie jako kryterium wyodrębniania ustawionych jest pięć czynników/składników.

Format macierzy składowych/czynników. Opcje te kontrolują format macierzy czynników (lub macierzy składowych w modelach PCA).

- **Sortuj wartości.** Jeśli ta opcja jest zaznaczona, ładunki czynnikowe w wynikach modelu będą posortowane numerycznie.
- **Ukryj wartości poniżej.** Jeśli ta opcja jest zaznaczona, wyniki poniżej określonej wartości progowej będą w macierzy ukryte, aby ułatwić dostrzeżenie wzorca w macierzy.

Rotacja. Te opcje umożliwiają kontrolowanie metody rotacji dla modelu. Więcej informacji można znaleźć w temacie "Opcje rotacji węzła analizy PCA/czynnikowej".

Opcje rotacji węzła analizy PCA/czynnikowej

W wielu przypadkach matematyczna rotacja zestawu zachowanych czynników może zwiększyć ich użyteczność, a w szczególności możliwość ich interpretacji. Można wybrać następujące metody rotacji:

- **Bez rotacji.** Opcja domyślna. Nie jest stosowana żadna rotacja.
- **Varimax.** Metoda rotacji ortogonalnej, która minimalizuje liczbę zmiennych z wysokimi ładunkami czynnikowymi. Upraszcza interpretację czynników.
- **Prosta Oblimin.** Metoda rotacji ukośnej (nieortogonalnej). Kiedy wartość **Delta** jest równa 0 (ustawienie domyślne) osie czynników są najbardziej ukośne. Im większą wartość ujemną przyjmie wskaźnik delta, tym mniej ukośne będą osie czynników. Aby zmienić domyślną wartość delty (równą 0) należy wprowadzić liczbę mniejszą od lub równą 0,8.
- **Quartimax.** Metoda ortogonalna, która minimalizuje liczbę czynników potrzebnych do wyjaśnienia każdej zmiennej. Upraszcza interpretację obserwowanych zmiennych.
- **Equamax.** Metoda rotacji, która jest kombinacją metody Varimax upraszczającej interpretację czynników i metody Quartimax upraszczającej interpretację zmiennych. Technika ta minimalizuje liczbę zmiennych, które mają wysokie ładunki na poszczególnych czynnikach oraz liczbę czynników potrzebnych do wyjaśnienia poszczególnych zmiennych.
- **Promax.** Rotacja ukośna, która umożliwia skorelowanie czynników. Można ją wyliczyć szybciej niż rotację prostą Oblimin, dlatego jest ona użyteczna w przypadku dużych zbiorów danych. **Kappa** kontroluje ukośność rozwiązania (zakres, w jakim czynniki mogą zostać skorelowane).

Model użytkowy analizy PCA/czynnikowej

Model użytkowy analizy PCA/czynnikowej reprezentuje model analizy czynnikowej i model analizy głównych składowych (PCA) utworzony za pomocą węzła analizy PCA/czynnikowej. Zawierają one wszystkie informacje przechwycone przez wyuczony model, jak również informacje na temat wydajności i cech modelu.

Po uruchomieniu strumienia zawierającego model równania czynnikowego węzeł dodaje nową zmienną dla każdego czynnika lub składnika w modelu. Nazwy nowych zmiennych są pochodnymi nazwy modelu z przedrostkiem *$F-* i przyrostkiem *-n*, gdzie *n* oznacza numer czynnika lub składnika. Na przykład, jeśli model ma nazwę *Factor* i zawiera trzy czynniki nowe zmienne będą miały następujące nazwy: *$F-Factor-1*, *$F-Factor-2* i *$F-Factor-3*.

Aby lepiej zrozumieć, co model czynnikowy zakodował, można przeprowadzić analizę w dalszych węzłach strumienia. Przydatnym sposobem wyświetlania wyniku dla modelu czynnikowego jest wyświetlanie wszystkich korelacji pomiędzy czynnikami i zmiennymi wejściowymi za pośrednictwem węzła statystyk. Pozwoli to zobaczyć, które zmienne wejściowe bardziej wpływają na określone czynniki oraz pomoże wykryć, czy istnieje dla czynników ukryte znaczenie lub interpretacja.

Model czynnikowy można również ocenić na podstawie informacji dostępnych w wynikach zaawansowanych. Aby wyświetlić wyniki zaawansowane, należy kliknąć zakładkę **Zaawansowane** w przeglądarce modeli użytkowych. Wyniki zaawansowane zawierają wiele szczegółowych informacji i są przeznaczone dla użytkowników z dużą wiedzą na temat analizy czynnikowej lub PCA. Więcej informacji można znaleźć w temacie "Wyniki zaawansowane modelu użytkowego analizy PCA/czynnikowej" na stronie 194.

Wyrażenia modelu użytkowego analizy PCA/czynnikowej

Karta Model dla modelu użytkowego Czynnik prezentuje wyrażenie oceny czynnikowej dla każdego czynnika. Oceny czynnikowe lub komponentowe są obliczane w wyniku pomnożenia każdej wartości zmiennej wejściowej przez współczynnik i zsumowanie wyników.

Podsumowanie modelu użytkowego analizy PCA/czynnikowej

Karta Podsumowanie dla modelu czynnikowego prezentuje liczbę czynników zachowanych w modelu czynnikowym/PCA, wraz z dodatkowymi informacjami na temat zmiennych i ustawień służących do generowania modelu. Więcej informacji można znaleźć w temacie “Przeglądanie modeli użytkowych” na stronie 42.

Wyniki zaawansowane modelu użytkowego analizy PCA/czynnikowej

Wyniki zaawansowane dla analizy czynnikowej udostępniają szczegółowe informacje na temat oszacowanego modelu i wydajności. Większość informacji zawartych w wynikach zaawansowanych to informacje ściśle techniczne; dlatego właściwa interpretacja tych wyników wymaga dużej wiedzy w zakresie analizy czynnikowej.

Ostrzeżenia. Wskazuje wszelkie ostrzeżenia i potencjalne problemy z wynikami.

Zasoby zmienności wspólnej. Przedstawia proporcje wariancji każdej zmiennej, jaka jest wyliczana na podstawie czynników lub składników. *Początkowe* zwraca początkowe zasoby zmienności wspólnej z całym zestawem czynników (model jest uruchamiany z liczbą czynników odpowiadającą liczbie zmiennych wejściowych); *Po wyodrębnieniu* zwraca zasoby zmienności wspólnej w oparciu o zachowany zestaw czynników.

Łączna wariancja wyjaśniona. Przedstawia łączną wariancję wyjaśnioną przez czynniki w modelu. Opcja *Początkowe wartości własne* przedstawia wariancję wyjaśnioną przez pełny zestaw czynników początkowych. Opcja *Sumy kwadratów ładunków po wyodrębnieniu* przedstawia wariancję wyjaśnioną przez czynniki zachowane w modelu. Opcja *Sumy kwadratów ładunków po rotacji* przedstawia wariancję wyjaśnioną przez czynniki poddane rotacji. Należy zwrócić uwagę, że dla rotacji ukośnych opcja *Sumy kwadratów ładunków po rotacji* będzie przedstawiała tylko sumy kwadratów ładunków i nie będzie przedstawiała wartości procentowych wariancji.

Macierz czynników (lub składników). Przedstawia korelacje pomiędzy zmiennymi wejściowymi a czynnikami nierotowanymi.

Macierz rotowanych czynników (lub składników). Przedstawia korelacje pomiędzy zmiennymi wejściowymi a zrotowanymi czynnikami dla rotacji ortogonalnych.

Macierz modelowa. Przedstawia częściowe korelacje pomiędzy zmiennymi wejściowymi a zrotowanymi czynnikami dla rotacji ukośnych.

Macierz struktury. Przedstawia proste korelacje pomiędzy zmiennymi wejściowymi a zrotowanymi czynnikami dla rotacji ukośnych.

Macierz korelacji czynników. Przedstawia korelacje pomiędzy czynnikami dla rotacji ukośnych.

---

węzeł Analiza dyskryminacyjna

Analiza dyskryminacyjna umożliwia budowanie modelu prognozowego przynależności do grup. Model jest budowany na podstawie funkcji dyskryminacyjnej (lub, dla więcej niż dwóch grup, zestawu funkcji dyskryminacyjnych) na podstawie liniowych kombinacji predyktorów, zapewniających najlepsze rozróżnienie między grupami. Funkcje są generowane z próbki obserwacji, których przynależność do grupy jest znana. Funkcje mogą następnie zostać zastosowane do nowych obserwacji, gdzie znane są miary dla predyktorów, ale nie przynależność do grupy.

Przykład. Operator telekomunikacyjny może zastosować analizę dyskryminacyjną w celu podzielenia klientów na grupy na podstawie danych o wykorzystaniu usług. Dzięki temu możliwe będzie dokonanie oceny potencjalnych klientów i skierowanie oferty do tych, którzy z największym prawdopodobieństwem należą do najbardziej wartościowych grup.

Wymagania. Potrzebna jest najmniej jedna zmienna wejściowa i dokładnie jedna zmienna przewidywana. Zmienna przewidywana musi być jakościowa (z poziomem pomiaru *Flaga* lub *Nominalna*) i składowana jako łańcuch lub liczba całkowita. (W razie potrzeby sposób składowania można przekształcić za pomocą węzła wypełniania lub węzła wyliczeń). Zmienne o roli *Łącznie* lub *Brak* są ignorowane. Typy zmiennych używanych w modelu muszą być w pełni zrealizowane jako instancje.

Mocne strony. Zarówno analiza dyskryminacyjna, jak i regresja logistyczna są wartościowymi modelami klasyfikacji. Jednak w analizie dyskryminacyjnej czyni się więcej założeń co do zmiennych wejściowych — np. że mają rozkład normalny i powinny być ilościowe. Analiza daje lepsze wyniki, gdy te założenia są spełnione, zwłaszcza przy małych próbach.

Opcje modelu węzła Analiza dyskryminacyjna

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Utwórz modele rozdzielone. Tworzy osobny model dla każdej możliwej wartości zmiennych wejściowych, jakie są określone jako zmienne podziału. Więcej informacji można znaleźć w “Budowanie modeli rozdzielonych” na stronie 28.

Metoda. Dostępne są następujące opcje związane z wprowadzaniem predyktorów do modelu:

- **Wprowadzanie.** Jest to metoda domyślna, która wprowadza wszystkie składniki bezpośrednio do równania. Składniki, które nie zwiększają istotnie jakości predykcji modelu, nie są dodawane.
- **Krokowa.** Początkowy model jest najprostszym możliwym modelem bez składników (z wyjątkiem stałej) w równaniu. Na każdym kroku składniki, które nie zostały jeszcze dodane do modelu, są oceniane, a jeśli najlepszy składnik znacząco zwiększa jakość predykcji modelu, jest dodawany.

Uwaga: Metoda Krokowa ma silną tendencję do przeuczania. Gdy te metody są stosowane, szczególnie ważne jest sprawdzenie poprawności modelu wynikowego z użyciem nowych danych albo próby testowej.

Opcje zaawansowanego węzła Analiza dyskryminacyjna

Użytkownikom dysponującym gruntowną wiedzą na temat algorytmów analizy dyskryminacyjnej opcje zaawansowane umożliwiają precyzyjne dostosowanie procesu uczenia. W celu uzyskania dostępu do opcji zaawansowanych należy ustawić opcję **Tryb** na wartość **Zaawansowany** na karcie Zaawansowany.

Prawdopodobieństwa a priori. Dzięki tej opcji można określić, czy współczynniki klasyfikacji są dostosowane do wiedzy a priori o przynależności do grup.

- **Dla wszystkich grup równe.** Dla wszystkich grup przyjmowane są równe prawdopodobieństwa wstępne. Nie ma to wpływu na współczynniki.
- **Oblicz na podstawie wielkości grup.** Wstępne prawdopodobieństwa przynależności do grup są określane na podstawie zaobserwowanych w próbce rozmiarów grup. Na przykład jeśli 50% obserwacji włączonych do analizy należy do pierwszej grupy, 25% do drugiej i 25% do trzeciej, współczynniki klasyfikacji są dopasowane do zwiększonego prawdopodobieństwa przynależności do pierwszej grupy w stosunku do pozostałych dwóch.

Użyj macierzy kowariancji. Możesz klasyfikować obserwacje z wykorzystaniem macierzy kowariancji wewnątrzgrupowych lub macierzy kowariancji odrębnych dla grup.

- *Wewnątrzgrupowe*. Do klasyfikacji obserwacji wykorzystywana jest połączona macierz kowariancji wewnątrzgrupowych.
- *Odrębne dla grup*. Wykorzystuje do klasyfikacji macierze kowariancji dla poszczególnych grup. Ponieważ klasyfikacja oparta jest na funkcjach dyskryminacyjnych a nie na pierwotnych zmiennych, opcja ta nie zawsze jest równoważna dyskryminacji kwadratowej.

Wynik. Te opcje umożliwiają zażądanie dodatkowych statystyk, które będą wyświetlane w zaawansowanych wynikach modelu użytkowego budowanego przez węzeł. Więcej informacji można znaleźć w temacie "Opcje wyników węzła Analiza dyskryminacyjna".

Krokowa. Te opcje umożliwiają określanie kryteriów dodawania i usuwania zmiennych w przypadku krokowej metody estymacji. (Przycisk jest wyłączony, jeśli wybrana jest metoda Wprowadzanie). Więcej informacji można znaleźć w temacie "Opcje metody krokowej węzła Analiza dyskryminacyjna" na stronie 197.

Opcje wyników węzła Analiza dyskryminacyjna

Należy wybrać opcjonalne wyniki, które będą wyświetlane w obszarze wyników zaawansowanych dla modelu użytkowego regresji logistycznej. W celu wyświetlenia zaawansowanych wyników przejdź do modelu użytkowego i kliknij kartę **Zaawansowane**. Więcej informacji można znaleźć w temacie "Zaawansowane wyniki modelu użytkowego Analiza dyskryminacyjna" na stronie 198.

Statystyki opisowe. Dostępne opcje to: średnia (w tym odchylenia standardowe), ANOVA dla każdej zmiennej i test *M* Boxa.

- *Średnie*. Wyświetla średnią ogólną i średnie w grupach oraz odchylenia standardowe dla zmiennych niezależnych.
- *ANOVA dla każdej zmiennej*. Dla każdej zmiennej niezależnej wykonuje test istotności różnic między średnimi grupowymi metodą jednoczynnikowej analizy wariancji.
- *M Boxa*. Test równości macierzy kowariancji grupowych. Przy odpowiednio dużych wielkościach prób nieistotna wartość p oznacza, że dowód nierówności macierzy jest niewystarczający. Test jest wrażliwy na odstępstwa od normalności rozkładu wielowymiarowego.

Współczynniki funkcji. Dostępne opcje to: współczynniki klasyfikacji Fishera i niestandaryzowane współczynniki.

- *Fishera*. Wyświetla współczynniki funkcji klasyfikacyjnej Fishera, które mogą być bezpośrednio używane do klasyfikowania. Dla każdej grupy otrzymywany jest oddzielny zestaw współczynników funkcji klasyfikacji, a przypadek klasyfikuje się do tej grupy, dla której ma najwyższą ocenę dyskryminacyjną (wartość funkcji klasyfikacji).
- *Niestandaryzowane*. Wyświetla niestandaryzowane współczynniki funkcji dyskryminacyjnej.

Macierze. Dostępne macierze współczynników dla zmiennych niezależnych to: macierz korelacji wewnątrzgrupowej, macierz kowariancji wewnątrzgrupowej, macierz kowariancji dla odrębnych grup i macierz kowariancji całkowitej.

- *Korelacja wewnątrzgrupowa*. Wyświetla macierz sumarycznych (połączonych) korelacji wewnątrzgrupowych, uzyskiwaną przez uśrednienie macierzy kowariancji dla wszystkich grup przed obliczeniem korelacji.
- *Kowariancja wewnątrzgrupowa*. Wyświetla macierz sumarycznych (połączonych) kowariancji wewnątrzgrupowych, która może być różna od całkowitej macierzy kowariancji. Macierz jest uzyskiwana przez uśrednienie poszczególnych macierzy kowariancji dla wszystkich grup.
- *Kowariancje dla odrębnych grup*. Wyświetla osobne macierze kowariancji dla każdej grupy.
- *Kowariancja całkowita*. Wyświetla macierz kowariancji obliczanych na podstawie wszystkich obserwacji, tak jakby pochodziły z jednej próby.

Klasyfikacja. Wyniki zaawansowane dotyczą wyników klasyfikacji.

- *Wyniki obserwacjami*. Dla każdej wyświetlane są kody rzeczywistej grupy, przewidywanej grupy, prawdopodobieństw a posteriori i ocen dyskryminacyjnych.
- *Tabela podsumowań*. Liczba obserwacji prawidłowo i nieprawidłowo przypisanych do każdej grupy na podstawie analizy dyskryminacyjnej. Czasem zwana „Macierzą nieporozumień".

- *Klasyfikacja typu pozostaw-jedną-poza*. Każda analizowana obserwacja jest klasyfikowana przez funkcję wyprowadzoną w oparciu o wszystkie pozostałe obserwacje z wyłączeniem tej jednej. Znana również jako „metoda U".
- *Mapa terytorialna*. Oparty o wartości funkcji dyskryminacyjnej wykres granic, wykorzystany do klasyfikowania obserwacji do grup. Liczby odpowiadają grupom, do których zostały zaklasyfikowane poszczególne obserwacje. Średnie dla kolejnych grup są na wykresie oznaczone gwiazdkami, które znajdują się wewnątrz granic określonych dla tych grup. Mapa nie zostaje wyświetlona wtedy, gdy istnieje tylko jedna funkcja dyskryminacyjna.
- *Połączone grupy*. Po zaznaczeniu tej opcji tworzony jest wykres rozrzutu wartości dwóch pierwszych funkcji dyskryminacyjnych, obejmujący wszystkie grupy. Jeśli istnieje tylko jedna funkcja, wyświetlany jest histogram.
- *Odrębne dla grup*. Tworzy wykresy rozrzutu oddzielne dla każdej z grup, z uwzględnieniem pierwszych dwu funkcji dyskryminacyjnych. Jeśli istnieje tylko jedna funkcja, wyświetlone zostaną histogramy.

Krokowa. Podsumowanie dla kolejnych kroków umożliwia wyświetlenie statystyk dla wszystkich zmiennych po każdym kroku; opcja **F dla odległości parami** umożliwia wyświetlenie macierzy połączonych w pary ilorazów *F* dla każdej pary grup. Ilorazy *F* mogą być używane do testowania istotności odległości Mahalanobisa między grupami.

Opcje metody krokowej węzła Analiza dyskryminacyjna

Metoda. Wybierz statystykę, która ma być wykorzystywana do wprowadzania lub usuwania nowych zmiennych. Dostępne opcje to: lambda Wilksa, Wariancja niewyjaśniona, Odległość Mahalanobisa, Najmniejszy iloraz *F* i *V* RAO. *V* Rao umożliwia określenie minimalnego przyrostu *V* dla wprowadzanej zmiennej.

- *Lambda Wilksa*. Metoda doboru zmiennych w krokowej analizie dyskryminacyjnej, przy której wybierane są takie zmienne, które po wprowadzeniu do równania najbardziej zmniejszą współczynnik lambda Wilksa. W każdym kolejnym kroku procedury wprowadzona zostaje ta zmienna, która minimalizuje wartość tego współczynnika.
- *Wariancja niewyjaśniona*. W każdym kolejnym kroku analizy do modelu wprowadzana jest zmienna, która minimalizuje sumę niewyjaśnionej zmienności między grupami.
- *Odległość Mahalanobisa*. Miara stopnia, w jakim wartości zmiennych niezależnych dla danej obserwacji różnią się od wartości przeciętnej dla wszystkich obserwacji. Duże wartości wskaźnika Mahalanobisa oznaczają, że obserwacja zawiera skrajne wartości jednej albo większej liczby zmiennych niezależnych.
- *Najmniejszy iloraz F*. Metoda doboru zmiennych przy analizie metodą krokową, oparta na maksymalizacji ilorazu F, obliczanego na podstawie odległości Mahalanobisa pomiędzy grupami.
- *V Rao*. Miara różnic między średnimi grupowymi. Znana jest także pod nazwą śladu Lawleya-Hotellinga. W każdym kolejnym kroku procedury wprowadzona zostaje ta zmienna, która powoduje największy wzrost wskaźnika V. Po wybraniu tej opcji wprowadź minimalną wartość, którą zmienna musi posiadać, aby została wprowadzona do analizy.

Kryteria. Dostępne opcje to: **Użyj wartości F** i **Zastosuj prawdopodobieństwo F.** Podaj wartości wykorzystywane do wprowadzania i usuwania zmiennych.

- *Użyj wartości F*. Zmienna zostaje wprowadzona do modelu, jeśli wartość F jest większa niż określona wartość kryterium wprowadzenia, a zostaje wyłączona, jeśli wartość F jest mniejsza niż wartość przyjęta jako kryterium usunięcia. Wartość wprowadzenia musi być większa od wartości usunięcia i obie muszą być dodatnie. Chcąc wprowadzić więcej zmiennych do modelu, należy obniżyć wartość wprowadzenia. Chcąc usunąć więcej zmiennych, należy zwiększyć wartość usunięcia.
- *Zastosuj prawdopodobieństwo F*. Zmienna zostaje wprowadzona do modelu, jeśli oszacowany dla niej poziom istotności dla wartości F jest mniejszy niż określona wartość kryterium wprowadzenia, a zostaje wyłączona, jeśli poziom istotności jest większy niż wartość przyjęta jako kryterium usunięcia. Wartość wprowadzenia musi być mniejsza od wartości usunięcia i obie muszą być dodatnie. Chcąc wprowadzić więcej zmiennych do modelu, należy zwiększyć wartość wprowadzenia. Aby usunąć więcej zmiennych, należy zmniejszyć wartość usunięcia.

Model użytkowy Analiza dyskryminacyjna

Model użytkowy Analiza dyskryminacyjna odzwierciedla równania oszacowane przez węzły Analiza dyskryminacyjna. Zawierają one wszystkie informacje przechwytywane przez model Analiza dyskryminacyjna, a także informacje dotyczące struktury i wydajności modelu.

W przypadku uruchomienia strumienia zawierającego model użytkowy modelu Analiza dyskryminacyjna węzeł dodaje dwie nowe zmienne zawierające predykcję modelu i powiązane prawdopodobieństwo. Nazwy nowych zmiennych pochodzą od nazwy przewidywanej zmiennej wyjściowej. Nazwy przewidywanych kategorii są poprzedzone znakami *$D-*, a nazwy powiązanych prawdopodobieństw są poprzedzone znakami *$DP-*. Na przykład w przypadku zmiennej wyjściowej o nazwie *colorpref* nowe zmienne będą miały nazwy *$D-colorpref* i *$DP-colorpref*.

Generowanie węzła filtrowania. Menu Utwórz umożliwia utworzenie nowego węzła filtrowania, który będzie przekazywał zmienne wejściowe na podstawie wyników modelu.

Ważność predyktorów

Opcjonalnie na karcie Model może być również wyświetlany wykres przedstawiający względną ważność poszczególnych predyktorów w oszacowaniu modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Należy zwrócić uwagę, że ten wykres jest dostępny tylko po wybraniu opcji **Oblicz ważność predyktora** na karcie Analiza przed wygenerowaniem modelu. Więcej informacji można znaleźć w temacie “Ważność predyktorów” na stronie 43.

Zaawansowane wyniki modelu użytkowego Analiza dyskryminacyjna

Wyniki zaawansowane dla analizy dyskryminacyjnej udostępniają szczegółowe informacje na temat oszacowanego modelu i wydajności. Większość informacji zawartych w wynikach zaawansowanych to informacje ściśle techniczne; dlatego właściwa interpretacja tych wyników wymaga dużej wiedzy w zakresie analizy dyskryminacyjnej. Więcej informacji można znaleźć w temacie “Opcje wyników węzła Analiza dyskryminacyjna” na stronie 196.

Ustawienia modelu użytkowego Analiza dyskryminacyjna

Podczas oceniania modelu karta Ustawienia dla modelu użytkowego Analiza dyskryminacyjna umożliwia uzyskanie ocen skłonności. Ta karta jest dostępna dla modeli zawierających tylko zmienne przewidywane typu flaga i tylko w przypadku gdy do strumienia dodano model użytkowy.

Wylicz surowe oceny skłonności. W przypadku modeli z przewidywaną zmienną typu flaga (zwracających predykcje tak lub nie) można wyliczyć oceny skłonności definiujące prawdopodobieństwo prawdziwego wyniku określonego dla zmiennej przewidywanej. Stanowią one uzupełnienie pozostałych współczynników ufności i wartości predykcyjnych, jakie mogą zostać wygenerowane podczas oceniania.

Wylicz skorygowane oceny skłonności. Surowe oceny skłonności bazują na danych uczących i mogą być zbyt optymistyczne z uwagi na tendencję wielu modeli do przeuczania tych danych. Opcja wyliczania skłonności skorygowanych próbuje skompensować tę tendencję na drodze oceny wydajności modelu w teście lub w podzbiorze walidacyjnym. Opcja ta wymaga, aby zmienna dzieląca na podzbiory była zdefiniowana w strumieniu oraz aby skorygowane oceny skłonności były włączone w węźle modelowania przed przystąpieniem do generowania modelu.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Podsumowanie modelu użytkowego Analiza dyskryminacyjna

Na karcie Podsumowanie modelu użytkowego Analiza dyskryminacyjne wyświetlane są zmienne i ustawienia służące do wygenerowania modelu. Ponadto jeśli wykonano węzeł analizy dołączony do tego węzła modelowania, informacje z

tej analizy również będą wyświetlane w tej sekcji. Ogólne informacje na temat korzystania z przeglądarki modelu zawiera sekcja “Przeglądanie modeli użytkowych” na stronie 42.

Węzeł Modele uogólnione

Uogólniony model liniowy rozszerza ogólny model liniowy w taki sposób, że zmienna zależna jest liniowo powiązana z czynnikami i współzmiennymi za pośrednictwem określonej funkcji łączenia. Model pozwala ponadto, aby zmienna zależna nie miała rozkładu normalnego. Dzięki bardzo ogólnej postaci wzoru modelu obejmuje on wiele modeli statystycznych, takich jak regresja liniowa dla odpowiedzi o rozkładzie normalnym, modele logistyczne dla danych binarnych, modele logarytmiczno-liniowe dla danych o liczebności i wiele innych modeli statystycznych.

Przykłady. Firma transportowa może używać uogólnionych modeli liniowych do dopasowania regresji Poissona w celu zliczenia uszkodzeń dla kilku typów statków zbudowanych w różnym okresie, a model wynikowy może ułatwić określenie, które typy statków są najbardziej podatne na uszkodzenia.

Firma zajmująca się ubezpieczeniami samochodów może używać uogólnionych modeli liniowych w celu dopasowania regresji gamma do roszczeń związanych z uszkodzeniami samochodów, a model wynikowy może pomóc w ustaleniu czynników, jakie wpłynęły na wysokość większości roszczeń.

Badacze w dziedzinie medycyny mogą używać uogólnionych modeli liniowych do dopasowania regresji komplementarnej log-log do danych przeżycia obciętych przedziałowych w celu ustalenia predykcji czasu potrzebnego do ponownego wystąpienia określonego stanu zdrowia.

Działanie uogólnionych modeli liniowych polega na budowaniu równania, które tworzy relację pomiędzy wartościami zmiennych wejściowych a wartościami zmiennych wyjściowych. Po wygenerowaniu modelu może być on używany do oszacowania wartości dla nowych danych. Dla każdego rekordu dla każdej możliwej kategorii wyjściowej obliczane jest prawdopodobieństwo członkostwa. Jako predykowana wartość wyjściowa dla tego rekordu przypisywana jest kategoria zmiennej przewidywanej o najwyższym prawdopodobieństwie.

Wymagania. Wymagana jest co najmniej jedna zmienna wejściowa i dokładnie jedna zmienna przewidywana (której poziom pomiaru może być określony jako *Ilościowa* lub *Flaga*) z co najmniej dwoma kategoriami. Typy zmiennych używanych w modelu muszą być w pełni zrealizowane jako instancje.

Mocne strony. Uogólniony model liniowy jest niezwykle elastyczny, ale proces wyboru struktury modelu nie jest zautomatyzowany; dlatego konieczna jest znajomość danych, co nie jest wymagane w przypadku algorytmów typu „czarna skrzynka”.

Opcje zmiennych węzła Modele uogólnione

Oprócz opcji zmiennych przewidywanych, wejściowych i dzielących na podzbiory zwykle dostępnych na kartach Zmienne węzła modelowania (patrz “Opcje zmiennych węzła modelowania” na stronie 31), węzeł Modele uogólnione zapewnia następujące dodatkowe funkcje.

Użyj zmiennej ważącej. Parametr skali to oszacowanie parametru modelu w odniesieniu do wariancji odpowiedzi. Wagi parametrów skali są wartościami „znanymi”, które mogą się różnić między obserwacjami. Jeśli określona jest zmienna wagi parametru skali, wówczas parametr skali, który jest powiązany z wariancją odpowiedzi, jest dzielony przez jej wartości dla każdej obserwacji. W analizie nie są używane rekordy z wartościami wagi parametru skali mniejszymi od zera lub równymi 0 ani obserwacje brakujące.

Zmienna przewidywana przedstawia liczbę zdarzeń w zbiorze prób. Gdy odpowiedź jest liczbą zdarzeń występujących w zbiorze prób, zmienna przewidywana zawiera liczbę zdarzeń. Można wybrać dodatkową zmienną zawierającą liczbę prób. Jeśli natomiast liczba prób jest taka sama we wszystkich obiektach, oznacza to, że próby mogły być określone za pomocą wartości stałej. Liczba prób powinna być większa niż lub równa liczbie zdarzeń w każdym rekordzie. Zdarzenia powinny być nieujemnymi liczbami całkowitymi, a próby powinny być dodatnimi liczbami całkowitymi.

Opcje modelu węzła Modele uogólnione

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Utwórz modele rozdzielone. Tworzy osobny model dla każdej możliwej wartości zmiennych wejściowych, jakie są określone jako zmienne podziału. Więcej informacji można znaleźć w “Budowanie modeli rozdzielonych” na stronie 28.

Typ modelu. Dostępne są dwie opcje typu modelu do zbudowania. Po wybraniu opcji **Tylko efekt główny** model uwzględnia tylko pojedyncze zmienne wejściowe i nie przeprowadza i nie testuje interakcji (efektów multiplikatywnych) między zmiennymi wejściowymi. Opcja **Efekty główne i wszystkie interakcje drugiego rzędu** uwzględnia wszystkie iteracje drugiego rzędu a także efekty główne zmiennej wejściowej.

Przesunięcie. Składnik przesunięcia jest predyktorem „strukturalnym”. Jego wskaźnik nie jest szacowany przez model, ale przyjmuje się, że ma wartość 1; dlatego wartości przesunięcia są po prostu dodawane do predyktora liniowego zmiennej przewidywanej. Jest to szczególnie przydatne w modelach regresji Poissona, w których każda obserwacja może mieć inny poziom ekspozycji na badane zdarzenie.

Na przykład przy modelowaniu częstości wypadków wśród poszczególnych kierowców należy pamiętać o istotnej różnicy między kierowcą, który miał jeden wypadek w ciągu trzech lat, a kierowcą, który miał jeden wypadek w ciągu 25 lat! Liczba wypadków może być modelowana jako odpowiedź o rozkładzie Poissona lub odpowiedź o rozkładzie ujemnym dwumianowym z logarytmiczną funkcją łączenia, jeśli logarytm naturalny doświadczenia kierowcy jest uwzględniony w składniku przesunięcia.

Inne kombinacje typów rozkładu i funkcji łączenia będą wymagały przekształcenia zmiennej przesunięcia.

Uwaga: jeśli używane są różne zmienne przesunięcia, określona zmienna również nie powinna być używana jako wejściowa. W razie konieczności w poprzedzającym węźle źródłowym lub węźle Typ rolę zmiennej przesunięcia należy ustawić na **Brak**.

Kategoria odniesienia dla przewidywanej flagi.

W przypadku zmiennych dychotomicznych można wybrać kategorię odniesienia dla zmiennej zależnej. Może to wpływać na określone wyniki, takie jak oszacowania parametrów i zapisane wartości, ale nie powinno zmienić dopasowania modelu. Na przykład, jeśli zmienna dychotomiczna przyjmuje wartości 0 i 1:

- Domyślnie procedura wybierze jako kategorię odniesienia ostatnią (o najwyższej wartości) kategorię lub 1. W tej sytuacji prawdopodobieństwa zapisane w modelu oszacują szansę, że dana obserwacja będzie miała wartość 0, a oszacowania parametrów powinny być interpretowane w odniesieniu do wiarygodności kategorii 0.
- Jeśli jako kategoria odniesienia wybrana zostanie pierwsza (o najniższej wartości) kategoria lub 0, wówczas prawdopodobieństwa zapisane w modelu oszacują szansę, że dana obserwacja będzie miała wartość 1.
- Jeśli wybrana zostanie kategoria użytkownika, a zmienna zawiera zdefiniowane etykiety, kategorię odniesienia można ustawić, wybierając wartość z listy. Może to być wygodne, kiedy w trakcie określania modelu, użytkownik nie pamięta dokładnie, jak była zakodowana konkretna zmienna.

Uwzględnij wyraz wolny w modelu. Wyraz wolny jest zwykle uwzględniany w modelu. Jeśli można założyć, że dane przechodzą przez początek układu współrzędnych, wyraz wolny można wyłączyć z modelu.

Opcje zaawansowane węzła Modele uogólnione

Użytkownikom dysponującym gruntowną wiedzą na temat uogólnionych modeli liniowych opcje zaawansowane umożliwiają precyzyjne dostosowanie procesu uczenia. W celu uzyskania dostępu do opcji zaawansowanych należy ustawić opcję **Tryb** na wartość **Zaawansowany** na karcie Zaawansowany.

Rozkład zmiennej przewidywanej i funkcja łączenia

Rozkład.

Ten wybór określa rozkład zmiennej zależnej. Możliwość określenia rozkładu innego niż normalny i nietożsamościowej funkcji łączenia jest istotnym ulepszeniem uogólnionego modelu liniowego w porównaniu do ogólnego modelu liniowego. Istnieje wiele możliwych kombinacji rozkład-funkcja łączenia, a kilka z nich może być odpowiednich dla dowolnego zbioru danych, dlatego wybór może być zależny od rozważań teoretycznych apriori lub tego, która kombinacja wydaje się zapewniać najlepsze dopasowanie.

- **Dwumianowy.** Ten rozkład jest odpowiedni tylko dla zmiennych, które reprezentują zmienne dychotomiczne lub liczbę zdarzeń.
- **Gamma.** Ten rozkład jest odpowiedni dla zmiennych z dodatnimi wartościami skali, które są skośne w kierunku większych wartości dodatnich. Jeśli wartość danych jest mniejsza niż lub równa 0 lub występuje brak wartości, wówczas dana obserwacja nie jest wykorzystywana w analizie.
- **Odwrócony Gaussa.** Ten rozkład jest odpowiedni dla zmiennych z dodatnimi wartościami skali, które są skośne w kierunku większych wartości dodatnich. Jeśli wartość danych jest mniejsza niż lub równa 0 lub występuje brak wartości, wówczas dana obserwacja nie jest wykorzystywana w analizie.
- **Ujemny dwumianowy.** Ten rozkład może być traktowany jako seria prób wymaganych do zaobserwowania *k* sukcesów i jest odpowiedni dla zmiennych z nieujemnymi liczbami całkowitymi. Jeśli wartość danych nie jest liczbą całkowitą, jest mniejsza od 0 lub występuje brak danych, wówczas dana obserwacja nie jest wykorzystywana w analizie. Wartość stała parametru dodatkowego ujemnego rozkładu dwumianowego może być dowolną liczbą większą niż lub równą 0. Jeśli parametr dodatkowy jest ustawiony na 0, użycie tego rozkładu da takie same efekty, jak użycie rozkładu Poissona.
- **Normalny.** Ten rozkład jest odpowiedni dla zmiennych ilościowych, których wartości rozkładają się symetrycznie, w kształcie dzwonu, wokół wartości centralnej (średniej). Zmienna zależna musi być typu liczbowego.
- **Poissona.** Ten rozkład można traktować jako liczbę wystąpień zdarzenia badanego w ustalonym okresie i jest odpowiedni dla zmiennych o nieujemnych wartościach całkowitych. Jeśli wartość danych nie jest liczbą całkowitą, jest mniejsza od 0 lub występuje brak danych, wówczas dana obserwacja nie jest wykorzystywana w analizie.
- **Tweedie'go.** Ten rozkład jest odpowiedni dla zmiennych, które mogą być reprezentowane przez poissonowsko mieszane rozkłady gamma; rozkład ten jest mieszany, to znaczy że łączy właściwości rozkładu ciągłego (nieujemne wartości rzeczywiste) i dyskretnego (prawdopodobieństwo dodatnie dla pojedynczej wartości, 0). Zmienna zależna musi być liczbowa, z wartościami danych większymi niż lub równymi zero. Jeśli wartość danych jest mniejsza niż zero lub występuje brak danych, wówczas dana obserwacja nie jest wykorzystywana w analizie. Wartość stała parametru rozkładu Tweedie'go może być dowolną liczbą większą niż jeden i mniejszą niż dwa.
- **Wielomianowy.** Ten rozkład jest odpowiedni dla zmiennych, które reprezentują odpowiedzi porządkowe. Zmienna zależna może być liniowa lub łańcuchowa i musi zawierać co najmniej dwie różniące się poprawne wartości danych.

Funkcje łączenia.

Funkcja łączenia to przekształcenie zmiennej zależnej, które umożliwia estymację modelu. Dostępne są następujące funkcje:

- **Tożsamość.** $f(x)=x$. Zmienna zależna nie jest przekształcana. To połączenie może być używane dla dowolnego rozkładu.
- **Komplementarny log-log.** $f(x)=\log(-\log(1-x))$. Ma zastosowanie tylko w przypadku rozkładu dwumianowego.
- **Skumulowany Cauchit.** $f(x) = \tan(\pi (x - 0{,}5))$; ma zastosowanie do skumulowanego prawdopodobieństwa dla każdej kategorii odpowiedzi. Ma zastosowanie tylko w przypadku rozkładu wielomianowego.

- **Skumulowany logarytmiczno-logarytmiczny dopełnienia.** $f(x)=\ln(-\ln(1-x))$; ma zastosowanie do skumulowanego prawdopodobieństwa dla każdej kategorii odpowiedzi. Ma zastosowanie tylko w przypadku rozkładu wielomianowego.
- **Skumulowany logit.** $f(x)=\ln(x / (1-x))$; ma zastosowanie do skumulowanego prawdopodobieństwa dla każdej kategorii odpowiedzi. Ma zastosowanie tylko w przypadku rozkładu wielomianowego.
- **Skumulowany ujemny logarytmiczno-logarytmiczny.** $f(x)=-\ln(-\ln(x))$; ma zastosowanie do skumulowanego prawdopodobieństwa dla każdej kategorii odpowiedzi. Ma zastosowanie tylko w przypadku rozkładu wielomianowego.
- **Skumulowany probit.** $f(x)=\Phi^{-1}(x)$; ma zastosowanie do skumulowanego prawdopodobieństwa dla każdej kategorii odpowiedzi, gdzie Φ^{-1} jest odwrotnością funkcji skumulowanego rozkładu standardowego normalnego. Ma zastosowanie tylko w przypadku rozkładu wielomianowego.
- **Logarytm.** $f(x)=\log(x)$. To połączenie może być używane dla dowolnego rozkładu.
- **Logarytmiczny dopełnienia.** $f(x)=\log(1-x)$. Ma zastosowanie tylko w przypadku rozkładu dwumianowego.
- **Logit.** $f(x)=\log(x / (1-x))$. Ma zastosowanie tylko w przypadku rozkładu dwumianowego.
- **Ujemny dwumianowy.** $f(x)=\log(x / (x+k^{-1}))$, gdzie k to parametr dodatkowy ujemnego rozkładu dwumianowego. Ma zastosowanie tylko w przypadku ujemnego rozkładu dwumianowego.
- **Ujemny log-log.** $f(x)=-\log(-\log(x))$. Ma zastosowanie tylko w przypadku rozkładu dwumianowego.
- **Nieparzysty potęgowy.** $f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$, jeśli $\alpha \neq 0$. $f(x)=\log(x)$, jeśli $\alpha=0$. α jest wymaganą specyfikacją liczbową i musi być liczbą rzeczywistą. Ma zastosowanie tylko w przypadku rozkładu dwumianowego.
- **Probit.** $f(x)=\Phi^{-1}(x)$, gdzie Φ^{-1} jest odwrotnością funkcji skumulowanego rozkładu standardowego normalnego. Ma zastosowanie tylko w przypadku rozkładu dwumianowego.
- **Potęgowy.** $f(x)=x^{\alpha}$, jeśli $\alpha \neq 0$. $f(x)=\log(x)$, jeśli $\alpha=0$. α jest wymaganą specyfikacją liczbową i musi być liczbą rzeczywistą. To połączenie może być używane dla dowolnego rozkładu.

Parametry. Elementy sterujące w tej grupie pozwalają na określenie wartości parametrów po wybraniu określonych opcji rozkładu.

- **Parametry dla ujemnego dwumianowego.** W przypadku ujemnego rozkładu dwumianowego można wybrać, czy wartość ma zostać określona, czy też system ma wprowadzić wartość oszacowaną.
- **Parametry dla Tweedie'go.** W przypadku rozkładu Tweedie'go należy wybrać wartość z zakresu od 1,0 do 2,0 dla wartości stałej.

  Estymacja parametru. Elementy sterujące w tej grupie pozwalają na określenie metod estymacji i udostępniają wartości początkowe dla oszacowania parametrów.

  – **Metoda.** Można wybrać metodę estymacji parametru. Można wybrać metodę oceny Newtona-Raphsona, Fishera lub metodę hybrydową, w której interakcje oceny Fishera są wykonywane przed przejściem do metody Newtona-Raphsona. Jeśli zbieżność zostanie osiągnięta w fazie oceny Fishera metodą hybrydową przed wykonaniem maksymalnej liczby iteracji Fishera, algorytm będzie kontynuował działanie, stosując metodę Newtona-Raphsona.

  – **Metoda parametru skali.** Można wybrać metodę parametru skali. Metoda maksymalnej wiarygodności umożliwia oszacowanie parametru skali wraz z efektami modelu; należy zauważyć, że ta opcja nie jest odpowiednia, jeśli odpowiedź jest w rozkładzie ujemnym dwumianowym, Poissona lub rozkładzie dwumianowym . Opcje odchylenia i chi-kwadratu Pearsona pozwalają oszacować parametr skali na podstawie wartości tych statystyk. Alternatywnie, można określić wartość stałą parametru skali.
- **Macierz kowariancji.** Dla uogólnionej odwrotności macierzy Hessego estymator oparty na modelu ma wartość ujemną. Estymator odporny (zwany również estymatorem Hubera/White'a/kanapkowym) to „skorygowany” estymator oparty na modelu, który zapewnia zgodne oszacowanie kowariancji, nawet jeśli specyfikacja wariancji i funkcji łączenia jest niepoprawna.

Iteracje. Te opcje umożliwiają sterowanie parametrami zbieżności modelu. Więcej informacji można znaleźć w temacie “Iteracje uogólnionych modeli liniowych” na stronie 203.

Wynik. Te opcje umożliwiają zażądanie dodatkowych statystyk, które będą wyświetlane w zaawansowanych wynikach modelu użytkowego budowanego przez węzeł. Więcej informacji można znaleźć w temacie "Zaawansowane wyniki uogólnionych modeli liniowych".

Tolerancja osobliwości. Macierze osobliwe (lub nieodwracalne) zawierają liniowo zależne kolumny, które mogą powodować problemy z algorytmem estymacji. Nawet macierze prawie osobliwe mogą powodować uzyskiwanie słabych wyników, dlatego procedura będzie traktować macierze, których wyznacznik jest mniejszy zakres tolerancji, jako osobliwe. Należy podać wartość dodatnią.

Iteracje uogólnionych modeli liniowych

Istnieje możliwość ustawienia parametrów zbieżności dla oszacowania uogólnionego modelu liniowego.

Iteracje. Dostępne są następujące opcje:

- **Maksymalna liczba iteracji.** Maksymalna liczba iteracji, jakie wykona algorytm. Podaj nieujemną liczbę całkowitą.
- **Maksimum kroków połowienia.** Przy każdej iteracji rozmiar kroku jest zmniejszany o 0,5, aż do momentu wzrostu logarytmu wiarygodności lub osiągnięcia maksymalnej wartości połowienia kroków. Określ dodatnią liczbę całkowitą.
- **Sprawdź separację punktów danych.** Po wybraniu tej opcji algorytm przeprowadza testy, aby sprawdzić, czy oszacowania parametru mają unikalne wartości. Rozdzielenie następuje, kiedy procedura może stworzyć model, który poprawnie klasyfikuje wszystkie obserwacje. Ta opcja jest dostępna dla odpowiedzi dwumianowych w układzie dychotomicznym .

Regresja logistyczna: Kryteria zbieżności. Dostępne są następujące opcje

- **Zbieżność parametru.** Po zaznaczeniu tej opcji algorytm zatrzymuje się, gdy bezwzględna lub względna zmiana oszacowania parametru jest mniejsza od określonej wartości, która musi być dodatnia.
- **Zbieżność logarytmu wiarygodności.** Po zaznaczeniu tej opcji algorytm zatrzymuje się, gdy bezwzględna lub względna zmiana funkcji logarytmu wiarygodności jest mniejsza od określonej wartości, która musi być dodatnia.
- **Zbieżność Hessego.** Dla specyfikacji bezwzględnej zakłada się zbieżność, jeśli statystyka w oparciu o zbieżność Hessego jest mniejsza od określonej wartości dodatniej. Dla specyfikacji względnej zakłada się zbieżność, jeśli statystyka jest mniejsza niż iloczyn określonej wartości dodatniej i wartości bezwzględnej logarytmu wiarygodności.

Zaawansowane wyniki uogólnionych modeli liniowych

Należy wybrać opcjonalne wyniki, które będą wyświetlane w obszarze wyników zaawansowanych dla modelu użytkowego uogólnionego modelu liniowego. W celu wyświetlenia zaawansowanych wyników przejdź do modelu użytkowego i kliknij kartę **Zaawansowane**. Więcej informacji można znaleźć w temacie "Zaawansowane wyniki modelu użytkowego Modele uogólnione" na stronie 205.

Dostępne są następujące wyniki:

- **Podsumowanie przetwarzania przypadku.** Wyświetla liczbę i wartość procentową obserwacji uwzględnionych w analizie i z niej wykluczonych oraz tabelę Podsumowanie danych skorelowanych.
- **Statystyki opisowe.** Wyświetla statystyki opisowe i informacje podsumowujące dotyczące zmiennej zależnej, współzmiennych i czynników.
- **Informacje o modelu.** Wyświetla nazwę zbioru danych, zmienną zależną lub zdarzenia i zmienne prób, zmienną przesunięcia, ilościową zmienną ważącą, rozkład prawdopodobieństwa i funkcję łączenia.
- **Statystyki dobroci dopasowania.** Wyświetla odchylenie i skalowane odchylenie, chi-kwadrat Pearsona i skalowany chi-kwadrat Pearsona, logarytm wiarygodności, AIC (kryterium informacyjne Akaike), AICC (skończone skorygowane AIC próby), BIC (Bayesowskie kryterium informacyjne) oraz CAIC (spójne AIC).
- **Statystyki podsumowujące modelu.** Wyświetla testy dopasowania modelu, z uwzględnieniem statystyki ilorazu wiarygodności dla testu typu omnibus dopasowania modelu oraz statystyki dla kontrastu typu I lub III dla każdego efektu.

- **Oceny parametrów.** Wyświetlane są oszacowania parametrów i odpowiednie statystyki testu oraz przedziały ufności. Opcjonalnie oprócz surowych oszacowań parametrów można wyświetlić wykładnicze oszacowania parametrów.
- **Macierz kowariancji szacowanych parametrów** Wyświetla macierz kowariancji szacowanych parametrów.
- **Macierz korelacji szacowanych parametrów.** Wyświetla macierz korelacji oszacowanych parametrów.
- **Macierz współczynników kontrastów (L).** Wyświetla współczynniki kontrastu dla domyślnych efektów i dla szacowanych średnich brzegowych, o ile zażądano ich na karcie Średnie EM.
- **Ogólne funkcje estymowalne.** Wyświetla macierze do generowania macierzy współczynnika kontrastu (L).
- **Przebieg iteracji.** Wyświetla przebieg iteracji dla oszacowań parametrów i logarytmu wiarygodności oraz drukuje ostatnią ewaluację wektora gradientu i macierzy Hessego. Tabela przebiegu iteracji wyświetla oszacowania parametrów dla co *n*-tej iteracji począwszy od iteracji 0 (oszacowanie początkowe), gdzie *n* jest wartością przedziału drukowania. Jeśli tworzone jest żądanie przebiegu iteracji, wówczas ostatnia iteracja jest zawsze wyświetlana, niezależnie od wartości *n*.
- **Test mnożnika Lagrange'a.** Wyświetla statystyki mnożnika Lagrange'a dla oceny ważności parametru skali, jaki jest wyliczany za pośrednictwem odchylenia lub rozkładu chi-kwadrat Pearsona lub jest ustawiany na wartość stałą dla rozkładu normalnego, gamma i odwróconego Gaussa. W przypadku ujemnego rozkładu dwumianowego testowany jest stały parametr dodatkowy.

Efekty modelu. Dostępne są następujące opcje:

- **Typ analizy.** Określa typ analizy, jaka ma zostać utworzona. Analiza typu I jest ogólnie odpowiednia w przypadku znanych przyczyn a priori dot. sortowania predyktorów w modelu; typ III ma szersze zastosowanie. Statystyki Walda lub ilorazu wiarygodności są obliczane na podstawie wyboru dokonanego w grupie statystyk chi-kwadrat.
- **Przedziały ufności.** Określa poziom ufności większy niż 50 i mniejszy niż 100. Przedziały Walda są tworzone przy założeniu, że parametry mają asymptotyczny rozkład normalny; przedziały wiarygodności profilu są bardziej dokładne, ale wymagające obliczeniowo. Poziom tolerancji dla przedziałów wiarygodności profilu stanowi kryterium zatrzymywania algorytmu iteracyjnego używanego do obliczenia interwałów.
- **Funkcja logarytmu wiarygodności.** Kontroluje format wyświetlania funkcji logarytmu wiarygodności. Pełna funkcja obejmuje dodatkowy składnik, który jest stały w odniesieniu do oszacowań parametrów; nie ma wpływu ma oszacowanie parametrów i w niektórych produktach oprogramowania nie jest wyświetlany.

Model użytkowy Modele uogólnione

Model użytkowy Modele uogólnione reprezentuje równania oszacowane przez model Modele uogólnione. Zawierają one wszystkie informacje przechwytywane przez model, a także informacje dotyczące struktury i wydajności modelu.

Po uruchomieniu strumienia zawierającego model użytkowy Modele uogólnione węzeł dodaje nowe zmienne, których zawartość zależy od charakteru zmiennej przewidywanej:

- **Przewidywana zmienna typu flaga.** Dodaje zmienne zawierające przewidywaną kategorię i powiązane prawdopodobieństwo oraz prawdopodobieństwa dla każdej kategorii. Nazwy dwóch pierwszych nowych zmiennych pochodzą od nazwy przewidywanej zmiennej wyjściowej. Nazwy przewidywanych kategorii są poprzedzone przedrostkiem *$G-*, a nazwy powiązanych prawdopodobieństw są poprzedzone przedrostkiem *$GP-*. Przykładowo dla zmiennej wyjściowej o nazwie *default* nowe zmienne będą miały nazwę *$G-default* i *$GP-default*. Kolejne dwie dodatkowe zmienne otrzymają nazwy na podstawie wartości zmiennej wynikowej z przedrostkiem *$GP-*. Na przykład, jeśli prawidłowe wartości *default* to *Yes* i *No*, nowe zmienne będą miały nazwy *$GP-Yes* i *$GP-No*.
- **Przewidywana zmienna ilościowa.** Dodaje zmienne zawierające przewidywaną średnią i błąd standardowy.
- **Przewidywana zmienna ilościowa, reprezentująca liczbę zdarzeń w serii prób.** Dodaje zmienne zawierające przewidywaną średnią i błąd standardowy.
- **Przewidywana zmienna porządkowa.** Dodaje zmienne zawierające przewidywaną kategorię i powiązane prawdopodobieństwo dla każdej wartości w uporządkowanym zestawie. Nazwy zmiennych pochodzą od wartości z przewidywanego uporządkowanego zestawu. Nazwy przewidywanych kategorii są poprzedzone przedrostkiem *$G-*, a nazwy powiązanych prawdopodobieństw są poprzedzone przedrostkiem *$GP-*.

Generowanie węzła filtrowania. Menu Utwórz umożliwia utworzenie nowego węzła filtrowania, który będzie przekazywał zmienne wejściowe na podstawie wyników modelu.

Ważność predyktorów

Opcjonalnie na karcie Model może być również wyświetlany wykres przedstawiający względną ważność poszczególnych predyktorów w oszacowaniu modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Należy zwrócić uwagę, że ten wykres jest dostępny tylko po wybraniu opcji **Oblicz ważność predyktora** na karcie Analiza przed wygenerowaniem modelu. Więcej informacji można znaleźć w temacie “Ważność predyktorów” na stronie 43.

Zaawansowane wyniki modelu użytkowego Modele uogólnione

Wyniki zaawansowane dla uogólnionych modeli liniowych udostępniają szczegółowe informacje na temat estymowanego modelu i jego wydajności. Większość informacji zawartych w zaawansowanych wynikach ma charakter techniczny i w celu poprawnej interpretacji takich wyników wymagana jest rozległa wiedza na temat analiz tego typu. Więcej informacji można znaleźć w temacie “Zaawansowane wyniki uogólnionych modeli liniowych” na stronie 203.

Ustawienia modelu użytkowego Modele uogólnione

Podczas oceniania modelu karta Ustawienia dla modelu użytkowego Modele uogólnione umożliwia uzyskanie ocen skłonności, a także wygenerowanie kodu SQL. Ta karta jest dostępna dla modeli zawierających tylko zmienne przewidywane typu flaga i tylko w przypadku gdy do strumienia dodano model użytkowy.

Wylicz surowe oceny skłonności. W przypadku modeli z przewidywaną zmienną typu flaga (zwracających predykcje tak lub nie) można wyliczyć oceny skłonności definiujące prawdopodobieństwo prawdziwego wyniku określonego dla zmiennej przewidywanej. Stanowią one uzupełnienie pozostałych współczynników ufności i wartości predykcyjnych, jakie mogą zostać wygenerowane podczas oceniania.

Wylicz skorygowane oceny skłonności. Surowe oceny skłonności bazują na danych uczących i mogą być zbyt optymistyczne z uwagi na tendencję wielu modeli do przeuczania tych danych. Opcja wyliczania skłonności skorygowanych próbuje skompensować tę tendencję na drodze oceny wydajności modelu w teście lub w podzbiorze walidacyjnym. Opcja ta wymaga, aby zmienna dzieląca na podzbiory była zdefiniowana w strumieniu oraz aby skorygowane oceny skłonności były włączone w węźle modelowania przed przystąpieniem do generowania modelu.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Podsumowanie modelu użytkowego Modele uogólnione

Na karcie Podsumowanie modelu użytkowego Modele uogólnione wyświetlane są zmienne i ustawienia służące do wygenerowania modelu. Ponadto jeśli wykonano węzeł analizy dołączony do tego węzła modelowania, informacje z tej analizy również będą wyświetlane w tej sekcji. Ogólne informacje na temat korzystania z przeglądarki modelu zawiera sekcja “Przeglądanie modeli użytkowych” na stronie 42.

Uogólnione liniowe modele mieszane

Węzeł GLMM

Ten węzeł umożliwia utworzenie uogólnionego liniowego modelu mieszanego (GLMM).

Uogólnione liniowe modele mieszane

Uogólnione liniowe modele mieszane rozszerzają model liniowy w następujący sposób:

- Zmienna przewidywana jest związana liniowo z czynnikami i współzmiennymi za pomocą określonej funkcji łączenia.
- Zmienna przewidywana może mieć rozkład inny niż normalny.
- Obserwacje mogą być skorelowane.

Uogólnione liniowe modele mieszane obejmują szeroki wachlarz modeli, począwszy od prostych modeli regresji liniowej, aż po złożone wielopoziomowe modele dla danych z obserwacji długofalowych nieposiadających rozkładu normalnego.

Przykłady. Rada szkoły dzielnicowej może używać uogólnionego liniowego modelu mieszanego w celu ustalenia, czy eksperymentalna metoda nauczania skutecznie poprawia wyniki z matematyki. Uczniowie z tej samej klasy powinni zostać skorelowani, ponieważ są nauczani przez tego samego nauczyciela, a klasy w tej samej szkole również mogą być skorelowane, co umożliwia nam uwzględnienie efektów losowych na poziomie szkoły i klasy w celu uwzględnienia różnych źródeł zmienności.

Badacze pracujący w dziedzinach medycznych mogą używać uogólnionego liniowego modelu mieszanego w celu ustalenia, czy nowy lek przeciwdrgawkowy może zmniejszać częstotliwość ataków padaczkowych u pacjentów. Powtarzalne pomiary dla tego samego pacjenta są zwykle skorelowane pozytywnie, dlatego odpowiedni powinien być model mieszany z pewnymi efektami losowymi. Zmienna przewidywana — liczba ataków — przyjmuje dodatnie wartości całkowite, dlatego odpowiedni może być uogólniony liniowy model mieszany o rozkładzie Poissona i logarytmiczna funkcja łączenia.

Dostawcy telewizji kablowej, telefonii przewodowej i internetu przewodowego mogą używać uogólnionego liniowego modelu mieszanego w celu uzyskania dodatkowych informacji na temat potencjalnych klientów. Możliwe odpowiedzi charakteryzują się nominalnymi poziomami pomiaru, dlatego analityk firmowy stosuje uogólniony liniowy mieszany model logit z losowym wyrazem wolnym w celu przechwytywania korelacji między odpowiedziami na pytania o korzystanie z usług różnych typów (telewizja, telefon, internet) w odpowiedziach konkretnego respondenta uczestniczącego w ankiecie.

Karta struktury danych umożliwia określenie zależności strukturalnych między rekordami w bazie danych po skorelowaniu obserwacji. Jeśli rekordy w zestawie danych reprezentują niezależne obserwacje, nie ma potrzeby określać niczego na tej karcie.

Obiekty. Połączenie wartości wybranych zmiennych jakościowych powinno w sposób jednoznaczny definiować obiekty w zbiorze danych. Na przykład jedna zmienna *ID pacjenta* powinna wystarczyć do zdefiniowania obiektów w jednym szpitalu. Jeśli jednak numery identyfikacyjne pacjenta nie identyfikują jednoznacznie pacjentów w różnych szpitalach, wówczas konieczna może być kombinacja *ID szpitala* i *ID pacjenta*. Przy wielokrotnych pomiarach dla każdego obiektu zapisywane są wielokrotne obserwacje. Z tego powodu jeden obiekt może być w zbiorze danych przedstawiany w wielu rekordach.

Obiekt jest jednostka obserwacyjną, która może być rozważana niezależnie od innych obiektów. Na przykład odczyty ciśnienia krwi pacjenta w badaniu medycznym mogą być traktowane jako niezależne od odczytów innych pacjentów. Definiowanie obiektów staje się szczególnie istotne, gdy istnieją wielokrotne pomiary na każdy obiekt i gdy wymagane jest modelowanie korelacji między tymi obserwacjami. Na przykład można oczekiwać, że pomiary ciśnienia krwi u jednego pacjenta podczas kolejnych wizyt u lekarza będą skorelowane.

Wszystkie zmienne określone jako **obiekty** na karcie struktury danych są używane do definiowania obiektów dla struktury kowariancji rezydualnej i udostępniają listę możliwych zmiennych do definiowania obiektów dla struktur kowariancji efektów losowych w bloku efektów losowych.

Powtarzane pomiary. Zmienne określone w tym miejscu służą do identyfikowania obserwacji powtórzonych. Na przykład pojedyncza zmienna *Tydzień* może identyfikować 10 tygodni obserwacji w badaniu medycznym, a w celu identyfikacji obserwacji codziennych w ciągu roku można używać razem zmiennych *Miesiąc* i *Dzień*.

Definiuj grupy kowariancji według. Zmienne jakościowe określone w tym miejscu definiują niezależne zestawy parametrów kowariancji efektów powtarzanych ; po jednej dla każdej kategorii zdefiniowanej przez klasyfikację krzyżową zmiennych grupujących. Wszystkie obiekty mają ten sam typ kowariancji; obiekty w tej samej grupie kowariancji będą miały te same wartości dla parametrów.

Współrzędne kowariancji przestrzennej. Gdy jeden z typów kowariancji przestrzennej wybrano jako typ struktury kowariancji, zmienne na tej liście określają współrzędne powtarzanych obserwacji.

Typ kowariancji powtórzonej. Określa strukturę kowariancji dla reszt. Dostępne są następujące struktury:

- Autoregresja pierwszego rzędu (AR1)
- Autoregresyjna średnia ruchoma (1,1) (ARMA11)
- Symetria złożona
- Przekątna
- Tożsamość skalowana
- Przestrzenna: potęgowa
- Przestrzenna: wykładnicza
- Przestrzenna: Gaussa
- Przestrzenna: liniowa
- Przestrzenna: liniowo-logarytmiczna
- Przestrzenna: sferyczna
- Toeplitz
- Nieustrukturalizowana
- Składowe wariancji

Zmienna przewidywana: Te ustawienia definiują zmienną przewidywaną, jej rozkład, a także jej relację z predyktorami przez funkcję łączenia.

Zmienna przewidywana. Zmienna przewidywana jest wymagana. Może mieć dowolny poziom pomiaru, a poziom pomiaru zmiennej przewidywanej ogranicza to, które rozkłady i funkcje łączenia są odpowiednie.

- **Użyj liczby prób jako mianownika.** Gdy zmienna przewidywana odpowiedzi jest liczbą zdarzeń występujących w zbiorze prób, zmienna przewidywana zawiera liczbę zdarzeń. Można wybrać dodatkową zmienną zawierającą liczbę prób. Na przykład podczas testowania nowego pestycydu można wystawiać mrówki na działanie pestycydu w różnych stężeniach, a następnie rejestrować liczbę mrówek zabitych i liczbę mrówek w każdej próbie. W tym przypadku zmienna rejestrująca liczbę mrówek zabitych powinna być określona jako zmienna przewidywana (zdarzenia), a zmienna rejestrująca liczbę mrówek w każdej próbie powinna być określona jako zmienna prób. Jeśli liczba mrówek jest taka sama dla każdej próby, wówczas liczba prób może być określona przy użyciu wartości stałej.

  Liczba prób powinna być większa niż lub równa liczbie zdarzeń w każdym rekordzie. Zdarzenia powinny być nieujemnymi liczbami całkowitymi, a próby powinny być dodatnimi liczbami całkowitymi.
- **Dostosuj kategorię odniesienia.** W przypadku przewidywanej zmiennej jakościowej można wybrać kategorię odniesienia. To może wpłynąć na niektóre wyniki, takie jak oszacowania parametrów, ale nie powinno zmienić dopasowania modelu. Na przykład, jeśli zmienna przewidywana przyjmuje domyślnie wartości 0, 1 i 2, wówczas procedura ustawia ostatnią kategorię (o najwyższej wartości) — czyli 2 — jako kategorię odniesienia. W tej sytuacji oszacowania parametrów powinny być interpretowane jako odnoszące się do wiarygodności kategorii 0 lub 1 *w*

odniesieniu do wiarygodności kategorii 2. Jeśli zostanie określona kategoria niestandardowa, a zmienna przewidywana zawiera zdefiniowane etykiety, można ustawić kategorię odniesienia, wybierając wartość z listy. To może być wygodne, jeśli podczas określania modelu użytkownik nie pamięta dokładnie sposobu zakodowania konkretnej zmiennej.

Rozkład zmiennej przewidywanej i związek (połączenie) z modelem liniowym. Na podstawie wartości predyktorów model oczekuje, że rozkład wartości zmiennej przewidywanej będzie zgodny z określonym kształtem, a w przypadku wartości zmiennej przewidywanej, że będą powiązane liniowo z predyktorami przez określoną funkcję łączenia. Dostępne są skróty dla kilku typowych modeli. Jeśli istnieje konkretna kombinacja rozkładu i funkcji łączenia, dla której użytkownik planuje znaleźć dopasowanie, a która nie jest dostępna na liście skrótów, wówczas można również wybrać ustawienie **Użytkownika**.

- **Model liniowy.** Określa rozkład normalny z łączem tożsamości, które jest użyteczne, gdy zmienna przewidywana może zostać przewidziana z użyciem modelu regresji liniowej lub modelu ANOVA.
- **Regresja gamma.** Określa rozkład gamma z logarytmiczną funkcją łączenia, który powinien być używany, gdy zmienna przewidywana zawiera wszystkie wartości dodatnie i wykazuje skośność w stronę wyższych wartości.
- **Analiza logliniowa (LOGLINEAR).** Określa rozkład Poissona z logarytmiczną funkcją łączenia, który powinien być używany, gdy zmienna przewidywana reprezentuje liczbę wystąpień w stałym okresie czasu.
- **Regresja ujemna dwumianowa.** Określa rozkład ujemny dwumianowy z logarytmiczną funkcją łączenia, który powinien być używany, gdy zmienna przewidywana i mianownik reprezentują liczbę prób wymaganych do zaobserwowania *k* sukcesów.
- **Wielomianowa regresja logistyczna.** Określa rozkład wielomianowy, który powinien być używany, gdy zmienna przewidywana jest odpowiedzią wielokategoryjną. Używa skumulowanej funkcji łączenia logit (wyniki porządkowe) lub uogólnionej funkcji łączenia logit (wielokategoryjne odpowiedzi nominalne).
- **Binarna regresja logistyczna.** Określa rozkład dwumianowy z funkcją łączenia logit, który powinien być używany, gdy zmienna przewidywana jest odpowiedzią binarną przewidywaną przez model regresji logistycznej.
- **Binarny probit.** Określa rozkład dwumianowy z funkcją łączenia probit, który powinien być używany, gdy zmienna przewidywana jest odpowiedzią binarną z bazowym rozkładem binarnym.
- **Przeżycia obcięte przedziałowe.** Określa rozkład dwumianowy z funkcją łączenia komplementarny log-log, która jest użyteczna w analizie przeżycia, gdy niektóre obserwacje nie mają zdarzenia kończącego.

Rozkład

Ta opcja określa rozkład zmiennej przewidywanej. Możliwość określenia rozkładu innego niż normalny i funkcji łączenia innej niż tożsamość stanowi kluczowe udoskonalenie uogólnionego liniowego modelu mieszanego w porównaniu do liniowego modelu mieszanego. Istnieje wiele możliwych kombinacji rozkład-funkcja łączenia, a kilka z nich może być odpowiednich dla dowolnego zbioru danych, dlatego wybór może być zależny od rozważań teoretycznych apriori lub tego, która kombinacja wydaje się zapewniać najlepsze dopasowanie.

Dwumianowy

Ten rozkład jest odpowiedni tylko dla zmiennej przewidywanej, która reprezentuje odpowiedź binarną lub liczbę zdarzeń.

Gamma

Ten rozkład jest odpowiedni dla zmiennych przewidywanych z wartościami w skali dodatniej, które wykazują skośność w stronę większych wartości dodatnich. Jeśli wartość danych jest mniejsza niż lub równa 0 lub występuje brak wartości, wówczas dana obserwacja nie jest wykorzystywana w analizie.

Odwrócony Gaussa

Ten rozkład jest odpowiedni dla zmiennych przewidywanych z wartościami w skali dodatniej, które wykazują skośność w stronę większych wartości dodatnich. Jeśli wartość danych jest mniejsza niż lub równa 0 lub występuje brak wartości, wówczas dana obserwacja nie jest wykorzystywana w analizie.

Wielomianowy

Ten rozkład jest dla zmiennej przewidywanej, która reprezentuje odpowiedź wielokategoryjną. Forma modelu będzie zależna od poziomu pomiaru zmiennej przewidywanej.

Wynikiem dla **nominalnej** zmiennej przewidywanej będzie nominalny model wielomianowy, w którym szacowany jest osobny zestaw parametrów modelu dla każdej kategorii zmiennej przewidywanej (z wyjątkiem kategorii odniesienia). Oszacowania parametrów dla konkretnego predyktora przedstawiają związek między predyktorem a wiarygodnością dla każdej kategorii zmiennej przewidywanej, względem kategorii odniesienia.

Wynikiem dla **porządkowej** zmiennej przewidywanej będzie porządkowy model wielomianowy, w którym tradycyjny składnik stałej jest zastępowany przez zestaw parametrów **progowych**, które odnoszą się do prawdopodobieństwo skumulowanego kategorii zmiennej przewidywanej.

Ujemny dwumianowy

W regresji ujemnej dwumianowej używany jest ujemny rozkład dwumianowy z logarytmiczną funkcją łączenia, który powinien być używany, gdy zmienna przewidywana reprezentuje liczbę wystąpień o wysokiej wariancji.

Normalny

Jest odpowiedni w przypadku ilościowej zmiennej przewidywanej, której wartości przyjmują symetryczny rozkład w kształcie dzwona, z centralną wartością średnią.

Poissona

Ten rozkład można traktować jako liczbę wystąpień zdarzenia badanego w ustalonym okresie i jest odpowiedni dla zmiennych o nieujemnych wartościach całkowitych. Jeśli wartość danych nie jest liczbą całkowitą, jest mniejsza od 0 lub występuje brak danych, wówczas dana obserwacja nie jest wykorzystywana w analizie.

Funkcja łączenia

Funkcja łączenia to transformacja zmiennej przewidywanej, która umożliwia estymację modelu. Dostępne są następujące funkcje:

Tożsamość

$f(x)=x$. Zmienna przewidywana nie jest transformowana. Ta funkcja łączenia może być używana z dowolnym rozkładem, z wyjątkiem rozkładu wielomianowego.

Komplementarny log-log

$f(x)=\log(-\log(1-x))$. Jest odpowiednia tylko w przypadku rozkładu dwumianowego i wielomianowego.

Cauchit

$f(x)=\tan(\pi(x-0.5))$. Jest odpowiednia tylko w przypadku rozkładu dwumianowego i wielomianowego.

Log $f(x)=\log(x)$. Ta funkcja łączenia może być używana z dowolnym rozkładem, z wyjątkiem rozkładu wielomianowego.

Komplementarny log

$f(x)=\log(1-x)$. Ma zastosowanie tylko w przypadku rozkładu dwumianowego.

Logit $f(x)=\log(x / (1-x))$. Jest odpowiednia tylko w przypadku rozkładu dwumianowego i wielomianowego.

Ujemny log-log

$f(x)=-\log(-\log(x))$. Jest odpowiednia tylko w przypadku rozkładu dwumianowego i wielomianowego.

Probit $f(x)=\Phi^{-1}(x)$, gdzie Φ^{-1} jest odwrotnością funkcji skumulowanego rozkładu standardowego normalnego. Jest odpowiednia tylko w przypadku rozkładu dwumianowego i wielomianowego.

Potęgowy

$f(x)=x^{\alpha}$, jeśli $\alpha \neq 0$. $f(x)=\log(x)$, jeśli $\alpha=0$. α jest wymaganą specyfikacją liczbową i musi być liczbą rzeczywistą. Ta funkcja łączenia może być używana z dowolnym rozkładem, z wyjątkiem rozkładu wielomianowego.

Efekty stałe: Współczynniki efektów stałych są zwykle traktowane jako zmienne, których wszystkie wartości badane są reprezentowane w zbiorze danych, i mogą być używane podczas oceniania. Domyślnie zmienne z predefiniowaną rolą w danych wejściowych, które nie są określone w innym miejscu w oknie dialogowym, są wprowadzane w części modelu dotyczącej efektów stałych. Zmienne jakościowe (flaga, nominalne i porządkowe) są używane jako współczynniki w modelu, a zmienne ilościowe są używane jako współzmienne.

Wprowadź efekty do modelu, zaznaczając jedną zmienną lub większą liczbę zmiennych na liście źródłowej i przeciągając do listy efektów. Typ tworzonego efektu jest zależny od tego, do którego obszaru aktywnego zostanie przeciągnięte zaznaczenie.

- **Główne.** Zmienne odrzucane są wyświetlane jako osobne efekty główne u dołu listy efektów.
- **2. rzędu.** Wszystkie możliwe pary zmiennych odrzucanych pojawiają się jako interakcje 2. rzędu u dołu listy efektów.
- **3. rzędu.** Wszystkie możliwe trójki zmiennych odrzucanych pojawiają się jako interakcje 3. rzędu u dołu listy efektów.
- ***.** Kombinacja wszystkich zmiennych odrzucanych pojawia się jako pojedyncza interakcja u dołu listy efektów.

Przyciski po prawej stronie kreatora efektów umożliwiają wykonywanie różnych działań.

Tabela 10. Opisy przycisków kreatora efektów

Ikona	Opis
	Umożliwia usuwanie składników z modelu efektów stałych poprzez wybranie składników przeznaczonych do usunięcia i kliknięcie przycisku usuwania.
	Umożliwia reorganizację składników w modelu efektów stałych poprzez wybranie składników przeznaczonych do reorganizacji i kliknięcie strzałki w górę lub w dół.
	Umożliwia dodawanie składników zagnieżdżonych przy użyciu okna dialogowego “Dodaj składnik zdefiniowany przez użytkownika” poprzez kliknięcie przycisku Dodaj składnik zdefiniowany przez użytkownika.

Uwzględnij wyraz wolny. Wyraz wolny jest zwykle uwzględniany w modelu. Jeśli można założyć, że dane przechodzą przez początek układu współrzędnych, wyraz wolny można wyłączyć z modelu.

Dodaj składnik zdefiniowany przez użytkownika: W tej procedurze można zbudować składniki zagnieżdżone dla modelu. Składniki zagnieżdżone są przydatne do modelowania efektu czynników lub współzmiennych, których wartości nie wchodzą w interakcje z poziomami innego czynnika. Na przykład sieć sklepów spożywczych może analizować zwyczaje zakupowe swoich klientów w kilku sklepach. Ponieważ każdy klient bywa regularnie tylko w jednym z tych sklepów, efekt *Klient* jest **zagnieżdżony w** efekcie *Lokalizacja sklepu*.

Ponadto można uwzględnić efekty interakcji, takie jak składniki wielomianowe z tą samą współzmienną, lub dodać wiele poziomów zagnieżdżenia do składnika zagnieżdżonego.

Ograniczenia: W odniesieniu do składników zagnieżdżonych obowiązują następujące ograniczenia:

- Wszystkie czynniki w interakcji muszą być unikalne. A zatem, jeśli *A* jest czynnikiem, to niedozwolone jest określenie *A*A*.
- Wszystkie czynniki w efekcie zagnieżdżonym muszą być unikalne. A zatem, jeśli *A* jest czynnikiem, to niedozwolone jest określenie *A(A)*.
- Efekt nie może być zagnieżdżony w obrębie współzmiennej. A zatem, jeśli *A* jest czynnikiem, a *X* jest współzmienną, to określenie *A(X)* jest niedozwolone.

Tworzenie zagnieżdżonego składnika

1. Wybierz czynnik lub współzmienną zagnieżdżony/-ą w innym czynniku, a następnie kliknij przycisk ze strzałką.
2. Kliknij opcję **(W)**.
3. Wybierz czynnik, w którym zagnieżdżony jest poprzedni czynnik lub współzmienna, a następnie kliknij przycisk ze strzałką.
4. Kliknij opcję **Dodaj składnik**.

Opcjonalnie można uwzględnić efekty interakcji lub dodać wiele poziomów zagnieżdżania do zagnieżdżonego składnika.

Efekty losowe: Współczynniki efektów losowych to zmienne, których wartości w pliku danych mogą być traktowane jako próba losowa z większej populacji wartości. Są użyteczne, gdy konieczne jest wyjaśnienie nadmiernej zmienności w zmiennej przewidywanej. Domyślnie w przypadku wyboru więcej niż jednego obiektu na karcie struktury danych dla każdego obiektu poza obiektem najbardziej wewnętrznym zostanie utworzony blok efektów losowych. Na przykład, jeśli wybrano Szkołę, Klasę i Ucznia jako obiekty na karcie struktury danych, wówczas następujące bloki efektów losowych są tworzone automatycznie:

- Efekt losowy 1: obiektem jest szkoła (brak efektów, tylko wyraz wolny)
- Efekt losowy 2: obiektem jest szkoła * klasa (brak efektów, tylko wyraz wolny)

Z bloków efektów losowych można korzystać na dwa sposoby:

1. W celu dodania nowego bloku kliknij opcję **Dodaj blok...** To spowoduje otwarcie okna dialogowego “Blok efektów losowych”.
2. W celu dodania istniejącego bloku wybierz blok, który chcesz edytować, a następnie kliknij opcję **Edytuj blok...** To spowoduje otwarcie okna dialogowego “Blok efektów losowych”.
3. W celu usunięcia jednego lub większej liczby bloków, wybierz bloki, które chcesz usunąć, a następnie kliknij przycisk usuwania.

Blok efektów losowych: Wprowadź efekty do modelu, zaznaczając jedną zmienną lub większą liczbę zmiennych na liście źródłowej i przeciągając do listy efektów. Typ tworzonego efektu jest zależny od tego, do którego obszaru aktywnego zostanie przeciągnięte zaznaczenie. Zmienne jakościowe (flaga, nominalne i porządkowe) są używane jako współczynniki w modelu, a zmienne ilościowe są używane jako współzmienne.

- **Główne.** Zmienne odrzucane są wyświetlane jako osobne efekty główne u dołu listy efektów.
- **2. rzędu.** Wszystkie możliwe pary zmiennych odrzucanych pojawiają się jako interakcje 2. rzędu u dołu listy efektów.
- **3. rzędu.** Wszystkie możliwe trójki zmiennych odrzucanych pojawiają się jako interakcje 3. rzędu u dołu listy efektów.
- ***.** Kombinacja wszystkich zmiennych odrzucanych pojawia się jako pojedyncza interakcja u dołu listy efektów.

Przyciski po prawej stronie kreatora efektów umożliwiają wykonywanie różnych działań.

Tabela 11. Opisy przycisków kreatora efektów

Ikona	Opis
	Umożliwia usuwanie składników z modelu poprzez wybranie składników przeznaczonych do usunięcia i kliknięcie przycisku usuwania.
	Umożliwia reorganizację składników w modelu poprzez wybranie składników przeznaczonych do reorganizacji i kliknięcie strzałki w górę lub w dół.

Tabela 11. Opisy przycisków kreatora efektów (kontynuacja)

Ikona	Opis
	Umożliwia dodawanie składników zagnieżdżonych przy użyciu okna dialogowego “Dodaj składnik zdefiniowany przez użytkownika” na stronie 210 poprzez kliknięcie przycisku Dodaj składnik zdefiniowany przez użytkownika.

Uwzględnij wyraz wolny. Wyraz wolny nie jest domyślnie uwzględniony w modelu efektów losowych. Jeśli można założyć, że dane przechodzą przez początek układu współrzędnych, wyraz wolny można wyłączyć z modelu.

Wyświetl parametry predykcji dla tego bloku. Pozwala wyświetlać szacowania parametrów według estymatora efektów losowych.

Definiuj grupy kowariancji według. Zmienne jakościowe określone w tym miejscu definiują niezależne zestawy parametrów kowariancji efektów losowych; po jednej dla każdej kategorii zdefiniowanej przez klasyfikację krzyżową zmiennych grupujących. Dla każdego bloku efektów losowych można określić inny zestaw zmiennych grupujących. Wszystkie obiekty mają ten sam typ kowariancji; obiekty w tej samej grupie kowariancji będą miały te same wartości dla parametrów.

Kombinacja obiektów. Dzięki tej opcji możliwe jest określanie obiektów efektów losowych z obecnej kombinacji obiektów z karty Struktura danych. Na przykład, jeśli *Szkoła*, *Klasa* i *Uczeń* są w tej kolejności zdefiniowane jako obiekty na karcie Struktura danych, wówczas lista rozwijana Kombinacja obiektów będzie zawierała opcje **Brak**, **Szkoła**, **Szkoła * Klasa** oraz **Szkoła * Klasa * Uczeń**.

Typ kowariancji efektu losowego. Określa strukturę kowariancji dla reszt. Dostępne są następujące struktury:

- Autoregresja pierwszego rzędu (AR1)
- Autoregresyjna średnia ruchoma (1,1) (ARMA11)
- Symetria złożona
- Przekątna
- Tożsamość skalowana
- Toeplitz
- Nieustrukturalizowana
- Składowe wariancji

Waga i przesunięcie: Waga analizy. Parametr skali to oszacowanie parametru modelu w odniesieniu do wariancji odpowiedzi. Wagi analiz są wartościami „znanymi”, które mogą się różnić między obserwacjami. Jeśli określona jest zmienna wagi analizy, wówczas parametr skali, który jest powiązany z wariancją odpowiedzi, jest dzielony przez wartości wagi analizy dla każdej obserwacji. W analizie nie są używane rekordy z wartościami wagi analizy mniejszymi od zera lub równymi zeru ani obserwacje brakujące.

Przesunięcie. Składnik przesunięcia jest predyktorem „strukturalnym”. Jego wskaźnik nie jest szacowany przez model, ale przyjmuje się, że ma wartość 1; dlatego wartości przesunięcia są po prostu dodawane do predyktora liniowego zmiennej przewidywanej. Jest to szczególnie przydatne w modelach regresji Poissona, w których każda obserwacja może mieć inny poziom ekspozycji na badane zdarzenie.

Na przykład przy modelowaniu częstości wypadków wśród poszczególnych kierowców należy pamiętać o istotnej różnicy między kierowcą, który miał jeden wypadek w ciągu trzech lat, a kierowcą, który miał jeden wypadek w ciągu 25 lat! Liczba wypadków może być modelowana jako odpowiedź o rozkładzie Poissona lub odpowiedź o rozkładzie ujemnym dwumianowym z logarytmiczną funkcją łączenia, jeśli logarytm naturalny doświadczenia kierowcy jest uwzględniony w składniku przesunięcia.

Inne kombinacje typów rozkładu i funkcji łączenia będą wymagały przekształcenia zmiennej przesunięcia.

Ogólne opcje budowania: Te opcje określają niektóre bardziej zaawansowane kryteria używane do budowania modelu.

Porządek sortowania

Te elementy sterujące określają porządek kategorii dla zmiennych przewidywanych i czynników (jakościowych zmiennych wejściowych) na potrzeby określenia „ostatniej" kategorii. Ustawienie porządku sortowania zmiennych przewidywanych jest ignorowane, jeśli zmienne przewidywane nie jest jakościowe lub jeśli w ustawieniach "Zmienna przewidywana" na stronie 207 określona jest niestandardowa kategoria odniesienia.

Reguły zatrzymujące

Możliwe jest określenie maksymalnej liczby iteracji, jakie wykona algorytm. W algorytmie wykorzystywany jest proces z podwójną iteracją, który obejmuje pętlę wewnętrzną i zewnętrzną. Wartość określona dla maksymalnej liczby iteracji ma zastosowanie do obu pętli. Podaj nieujemną liczbę całkowitą. Domyślną wartością jest 100.

Ustawienia po estymacji

Te ustawienia określają sposób obliczania niektórych wyników modelu na potrzeby wyświetlania.

Poziom ufności (%)

Jest to poziom ufności używany do wyliczania oszacowań przedziałów współczynników modelu. Należy podać wartość większą od 0 i mniejszą od 100. Domyślna wartość to 95.

Stopnie swobody

Określa sposób obliczania stopni swobody dla testów istotności. Wybierz opcję **Ustalone dla wszystkich testów (metoda resztowa)**, jeśli próba jest wystarczająco duża lub dane są zrównoważone, albo model używa kowariancji prostszego typu — na przykład przekątna lub tożsamość skalowana. Jest to ustawienie domyślne. Wybierz opcję **Różne pomiędzy testami (przybliżenie Satterthwaite'a)**, jeśli próba jest niewielka, dane są niezrównoważone lub w modelu używany jest skomplikowany typ kowariancji — na przykład nieustrukturalizowana.

Testy efektów stałych i współczynników

Jest to metoda obliczania macierzy kowariancji oszacowań parametrów. Wybierz mocne oszacowanie, jeśli martwi Cię możliwość naruszenia założeń modelu.

Oszacowanie: W algorytmie budowania modelu wykorzystywany jest proces z podwójną iteracją, który obejmuje pętlę wewnętrzną i zewnętrzną. Względem pętli wewnętrznej zastosowanie mają następujące ustawienia.

Zbieżność parametru.

Zbieżność jest zakładana, jeśli maksymalna zmiana bezwzględna lub względna w oszacowaniach parametru jest mniejsza niż podana wartość, która musi być nieujemna. Kryterium nie jest stosowane, jeśli określona wartość jest równa 0.

Zbieżność logarytmu wiarygodności.

Zbieżność jest zakładana, jeśli zmiana bezwzględna lub względna w funkcji logarytmu wiarygodności jest mniejsza niż podana wartość, która musi być nieujemna. Kryterium nie jest stosowane, jeśli określona wartość jest równa 0.

Zbieżność Hessego.

W przypadku specyfikacji **Wartości bezwzględne** zakładana jest zbieżność, jeśli statystyka oparta na macierzy Hessego jest mniejsza niż określona wartość. W przypadku specyfikacji **Względne** zbieżność jest zakładana, jeśli statystyka jest mniejsza niż iloczyn wartości określonej i wartości bezwzględnej logarytmu wiarygodności. Kryterium nie jest stosowane, jeśli określona wartość jest równa 0.

Maksymalna liczba kroków oceny Fishera.

Podaj nieujemną liczbę całkowitą. Wartość 0 określa metodę Newtona-Raphsona. Wartości większe od 0 określają użycie algorytmu oceny Fishera aż do iteracji o numerze *n*, gdzie *n* jest podaną liczbą całkowitą, a następnie metody Newtona-Raphsona.

Tolerancja osobliwości.

Ta wartość jest stosowana jako tolerancja podczas kontroli osobliwości. Podaj wartość dodatnią.

Uwaga: Domyślnie używana jest zbieżność parametru, gdzie sprawdzana jest maksymalna **bezwzględna** zmiana przy tolerancji 1E-6. To ustawienie może zwracać wyniki różniące się od wyników uzyskiwanych w wersjach wcześniejszych niż wersja 22. W celu odtworzenia wyników z wersji wcześniejszych niż wersja 22 należy użyć opcji **Względne** dla kryterium zbieżności parametru i zachować domyślną wartość tolerancji równą 1E-6.

Ogólne: Nazwa modelu. Można automatycznie generować nazwę modelu na podstawie zmiennych docelowych lub podać nazwę użytkownika. Automatycznie wygenerowana nazwa jest nazwą zmiennej docelowej. Jeśli istnieje więcej niż jedna zmienna przewidywana, nazwa modelu składa się z listy nazw zmiennych połączonych ampersandami. Na przykład, jeśli zmienne przewidywane to *field1 field2 field3*, nazwa modelu będzie miała postać: *field1 & field2 & field3*.

Udostępnij do oceniania. Podczas oceniania modelu powinny być wygenerowane zaznaczone elementy z tej grupy. Przy ocenie modelu zawsze obliczana jest wartość przewidywana (dla wszystkich zmiennych przewidywanych) i ufność (dla jakościowych zmiennych przewidywanych). Obliczona ufność może być oparta na prawdopodobieństwie przewidywanej wartości (najwyższe przewidywane prawdopodobieństwo) lub różnicy między najwyższym przewidywanym prawdopodobieństwem a drugim co do wysokości przewidywanym prawdopodobieństwem.

- **Przewidywane prawdopodobieństwo dla przewidywanych zmiennych jakościowych.** Generuje przewidywane prawdopodobieństwa dla jakościowych zmiennych przewidywanych. Dla każdej kategorii tworzona jest jedna zmienna.
- **Oceny skłonności (ważne tylko dla przewidywanych zmiennych typu flaga).** W przypadku modeli z przewidywaną zmienną typu flaga (zwracających predykcje tak lub nie) można wyliczyć oceny skłonności definiujące prawdopodobieństwo prawdziwego wyniku określonego dla zmiennej przewidywanej. Model generuje surowe oceny skłonności; jeśli stosowane są podzbiory, model generuje także skorygowane oceny skłonności na podstawie podzbioru testowego.

Oszacowane średnie: Ta karta umożliwia wyświetlenie szacowanych średnich brzegowych dla poziomów współczynników i interakcji czynników. Szacowane średnie brzegowe są niedostępne w przypadku modeli wielomianowych.

Składniki
Ta lista zawiera składniki modelu w efektach stałych, które są w całości zbudowane ze zmiennych jakościowych. Należy sprawdzić każdy składnik, dla którego model ma zwracać szacowane średnie brzegowe.

Typ kontrastu
Określa typ kontrastu używanego do poziomów zmiennej kontrastu.

Brak Nie zostaną wygenerowane żadne kontrasty.

Parami
Powoduje uzyskiwanie porównań parami dla kombinacji wszystkich poziomów podanych czynników. Jest to jedyny kontrast dostępny dla interakcji czynników.

Odchylenie
Kontrasty porównują każdy poziom czynnika do średniej głównej.

Proste Kontrasty porównują każdy poziom czynnika, z wyjątkiem ostatniego, do ostatniego poziomu. Poziom „ostatni” jest określony przez porządek sortowania czynników określonych w opcjach budowania. Należy zwrócić uwagę na to, że żaden z tych typów kontrastów nie jest ortogonalny.

Zmienna kontrastu
Określa czynnik, którego poziomy są porównywane z użyciem wybranego typu kontrastu. Jeśli jako typ kontrastu wybrano **Brak**, wówczas nie można (i nie potrzeba) wybrać żadnej zmiennej kontrastu.

Zmienna ciągła
Zmienne ilościowe z listy są wyodrębniane ze składników w efektach stałych, w których stosowane są zmienne ilościowe. Podczas obliczania szacowanych średnich brzegowych współzmienne są ustalone na podanych wartościach. Wybierz średnią lub podaj wartość użytkownika.

Dla porównań wielokrotnych skoryguj metodą

Podczas testowania hipotez z użyciem kontrastów wielokrotnych całościowy poziom istotności można skorygować na podstawie poziomów istotności dla uwzględnionych kontrastów. Dzięki temu możliwe jest wybranie metody korekty.

Najmniejsza istotna różnica

Ta metoda nie kontroluje ogólnego prawdopodobieństwa odrzucenia hipotez, które stwierdzają, że niektóre kontrasty liniowe różnią się od wartości hipotezy zerowej.

Sekwencyjna Bonferroniego

Jest to sekwencyjne zstępująca odrzucająca procedura Bonferroniego, która jest mniej konserwatywna w zakresie odrzucania indywidualnych hipotez, ale zachowuje identyczny całościowy poziom istotności.

Sekwencyjna Sidaka

Jest to sekwencyjnie zstępująca odrzucająca procedura Bonferroniego, która jest mniej konserwatywna w zakresie odrzucania indywidualnych hipotez, ale zachowuje identyczny całościowy poziom istotności.

Metoda najmniejszej istotnej różnicy jest mniej konserwatywna niż metoda liczby sekwencyjnej Sidaka, która z kolei jest mniej konserwatywna niż metoda sekwencyjna Bonferroniego; oznacza to, że najmniejsza istotna różnica odrzuci co najmniej taką liczbę pojedynczych hipotez, co metoda sekwencyjna Sidaka, która z kolei odrzuci co najmniej taką liczbę pojedynczych hipotez, co metoda Sekwencyjna Bonferroniego.

Wyświetlaj estymowane średnie w odniesieniu do

Ta opcja określa, czy szacowane średnie brzegowe są obliczane na podstawie pierwotnej skali zmiennej przewidywanej, czy na podstawie transformacji funkcji łączenia.

Oryginalna skala zmiennej zależnej

Oblicza szacowane średnie brzegowe dla zmiennej przewidywanej. Należy zwrócić uwagę na to, że jeśli zmienna przewidywana zostanie określona z użyciem opcji zdarzeń/prób, zwróci szacowane średnie brzegowe dla proporcji zdarzeń/prób, a nie dla liczby zdarzeń.

Transformacja funkcji łączenia

Oblicza szacowane średnie brzegowe dla predyktora liniowego.

Widok modelu: Domyślnie wyświetlany jest widok Podsumowanie modelu. Aby wyświetlić inny widok, należy wybrać go z miniatur widoków.

Podsumowanie modelu: Widok jest obrazem stanu, szybkim podsumowaniem modelu i jego dopasowania.

Tabela. Tabela identyfikuje zmienną przewidywaną, rozkład prawdopodobieństwa oraz funkcja łączenia, które są określone w obszarze Ustawienia zmiennej przewidywanej. Jeśli zmienna przewidywana jest zdefiniowana przez zdarzenia i próby, komórka zostanie podzielona w celu przedstawienia zmiennej zdarzeń i zmiennej prób albo stałą liczbę prób. Dodatkowo zostanie wyświetlone skorygowane kryterium informacyjne Akaike (AICC) i Bayesowskie kryterium informacyjne (BIC) dla próby skończonej.

- *Skorygowane Akaike*. Miara wybierania i porównywania modeli mieszanych tworzonych na podstawie -2 logarytmu wiarygodności (ograniczonego). Mniejsze wartości oznaczają lepszy model. Wartość AICC „poprawia” wartość AIC w przypadku małych prób. Przy wzroście wielkości próby wartość AICC zbiega do wartości AIC.
- *Bayesowskie*. Miara wybierania i porównywania modeli mieszanych tworzonych na podstawie -2 logarytmu wiarygodności. Mniejsze wartości oznaczają lepszy model. Wartość BIC „karze” także modele przeparametryzowane (na przykład złożone modele z dużą liczbą danych wejściowych), jednak silniej niż miara AIC.

Wykres. Jeśli zmienna przewidywana jest jakościowa, na wykresie zostanie wyświetlona dokładność modelu końcowego, która jest procentem poprawnych klasyfikacji.

Struktura danych: Ten widok udostępnia podsumowanie struktury danych określonej przez użytkownika i ułatwia sprawdzenie, czy obiekty i powtarzane pomiary zostały poprawnie określone. Informacje obserwowane dla pierwszego

obiektu są wyświetlane dla każdej zmiennej obiektowej i każdej zmiennej pomiarów powtarzanych, a także dla zmiennej przewidywanej. Dodatkowo wyświetlana jest liczba poziomów dla każdej zmiennej obiektowej i zmiennej pomiarów powtarzanych.

Przewidywane według obserwowanych: W przypadku ilościowych zmiennych przewidywanych określonych jako zdarzenia/próby ten wykres przedstawia wykres rozrzutu z kategoryzacją wartości przewidywanych (na osi pionowej) na wartości obserwowane (na osi poziomej). W idealnym przypadku punkty te powinny leżeć na prostej nachylonej pod kątem 45 stopni; widok ten może stwierdzić, czy którekolwiek z wyników zostały przewidziane przez model w sposób oczywisty.

Klasyfikacja: W przypadku jakościowych zmiennych przewidywanych wyświetlana jest klasyfikacja krzyżowa wartości obserwowanych względem przewidywanych w mapie natężeń, a dodatkowo przedstawiany jest ogólny procent poprawnych.

Style tabel. Istnieje kilka różnych stylów wyświetlania, które są dostępne z poziomu listy rozwijanej **Styl**.

- **Procent w wierszu.** Wyświetla procent w wierszu (liczby komórek wyrażone jako procent sum z wierszy) w komórkach. Jest to ustawienie domyślne.
- **Liczby komórek.** Wyświetla liczby komórek w komórkach. Cieniowanie mapy komórek jest w dalszym ciągu wyrażone jako procent w wierszu.
- **Mapa natężeń.** W komórkach nie są wyświetlane żadne wartości, tylko cieniowanie.
- **Skompresowane.** Nagłówki kolumn ani wierszy nie są wyświetlane. Nie są wyświetlane również wartości w komórkach. Taki styl może być użyteczny, gdy zmienna przewidywana zawiera wiele kategorii.

Braki danych. Jeśli w jakichkolwiek rekordach brakuje wartości w zmiennej przewidywanej, wówczas takie rekordy są wyświetlane w wierszu **(Braki danych)**, który jest wyświetlany pod poprawnymi wierszami. Rekordy z brakami danych nie zwiększają ogólnego procentu poprawnych.

Wiele zmiennych przewidywanych. Jeśli istnieje wiele przewidywanych zmiennych jakościowych, wówczas każda zmienna przewidywana jest wyświetlana w osobnej tabeli i istnieje lista rozwijana **Zmienna przewidywana**, która kontroluje wyświetlane zmienne przewidywane.

Duże tabele. Jeśli wyświetlana zmienna przewidywana zawiera ponad 100 kategorii, żadna tabela nie jest wyświetlana.

Efekty stałe: Ten widok przedstawia rozmiar każdego efektu stałego w modelu.

Style. Dostępne są różne style wyświetlania, które są dostępne z poziomu listy rozwijanej **Styl**.

- **Diagram.** Jest to wykres, w którym efekty są posortowane od góry do dołu w kolejności, w jakiej zostały określone w ustawieniach Efekty stałe. Linie łączące w diagramie są ważone na podstawie istotności efektu, gdzie większa szerokość linii odpowiada bardziej istotnym efektom (niższe wartości *p*). Jest to ustawienie domyślne.
- **Tabela.** Jest to tabela ANOVA dla ogólnego modelu całkowitych i pojedynczych efektów modelu. Poszczególne efekty są posortowane od góry do dołu w kolejności, w jakiej zostały określone w ustawieniach Efekty stałe.

Istotność. Dostępny jest suwak Istotność, który steruje widocznością efektów w widoku. Efekty o wartościach istotności większych niż wartości suwaka, pozostają ukryte. Nie zmienia to modelu, ale pozwala po prostu na skoncentrowanie się na najistotniejszych efektach. Domyślną wartością jest 1,00 tak, że na podstawie istotności żadne efekty nie są filtrowane.

Współczynniki stałe: Widok ten przedstawia wartość każdego współczynnika stałego w modelu. Należy zwrócić uwagę, że czynniki (predyktory jakościowe) są kodowane wskaźnikami w ramach modelu tak, że **efekty** zawierające czynniki będą miały generalnie wiele powiązanych **współczynników**; po jednym dla każdej kategorii z wyjątkiem kategorii odpowiadającej współczynnikowi nadmiarowemu.

Style. Dostępne są różne style wyświetlania, które są dostępne z poziomu listy rozwijanej **Styl**.

- **Diagram.** Jest to wykres, w którym najpierw wyświetlany jest wyraz wolny, a następnie sortowane są efekty od góry do dołu w kolejności, w jakiej zostały określone w ustawieniach Efekty stałe. W efektach zawierających

czynniki, współczynniki są posortowane rosnąco według wartości danych. Linie łączące w diagramie są kolorowane i ważone na podstawie istotności współczynnika, gdzie większa szerokość linii odpowiada bardziej istotnym współczynnikom (niższe wartości *p*). Jest to domyślny styl.

- **Tabela.** Pokazuje ona wartości, testy istotności i przedziały ufności dla poszczególnych współczynników modelu. Po wyrazie wolnym poszczególne efekty są posortowane od góry do dołu w kolejności, w jakiej zostały określone w ustawieniach Efekty stałe. W efektach zawierających czynniki, współczynniki są posortowane rosnąco według wartości danych.

Wielomianowy. Jeśli obowiązuje rozkład wielomianowy, wówczas lista rozwijana Wielomianowy kontroluje to, która kategoria zmiennych przewidywanych będzie wyświetlana. Kolejność wyświetlania wartości na liście jest określona przez specyfikację w ustawieniach Opcje budowania.

Wykładnicza. Wyświetlane są oszacowania współczynników wykładniczych i przedziały ufności dla niektórych typów modeli, w tym dla binarnej regresji logistycznej (rozkład dwumianowy i funkcja łączenia logit), nominalnej regresji logistycznej (rozkład wielomianowy i funkcja łączenia logit), negatywnej regresji dwumianowej (negatywna regresja dwumianowa i logarytmiczna funkcja łączenia) oraz modelu logarytmiczno-liniowego (rozkład Poissona i logarytmiczna funkcja łączenia).

Istotność. Dostępny jest suwak Istotność, który steruje widocznością współczynników w widoku. Współczynniki o wartościach istotności większych niż wartości suwaka, pozostają ukryte. Nie zmienia to modelu, ale pozwala po prostu na skoncentrowanie się na najistotniejszych współczynnikach. Domyślną wartością jest 1,00 tak, że na podstawie istotności żadne współczynniki nie są filtrowane.

Kowariancje efektów losowych: W tym widoku przedstawiana jest macierz kowariancji efektów losowych (**G**).

Style. Dostępne są różne style wyświetlania, które są dostępne z poziomu listy rozwijanej **Styl**.

- **Wartości kowariancji.** Jest to mapa natężeń macierzy kowariancji, w której efekty są posortowane od góry do dołu w kolejności, w jakiej zostały określone w ustawieniach Efekty stałe. Kolory na diagramie corrgram odpowiadają wartościom komórek przedstawionych w kluczu. Jest to ustawienie domyślne.
- **Corrgram.** Jest to mapa natężeń macierzy kowariancji.
- **Skompresowane.** Jest to mapa natężeń macierzy kowariancji bez nagłówków wierszy i kolumn.

Bloki. Jeśli istnieje wiele bloków efektów losowych, wówczas istnieje lista rozwijana Blok przeznaczona do wyboru bloku do wyświetlenia.

Grupy. Jeśli blok efektów losowych obejmuje specyfikację grupy, wówczas istnieje lista rozwijana Grupa przeznaczona do wyboru poziomu grupy do wyświetlenia.

Wielomianowy. Jeśli obowiązuje rozkład wielomianowy, wówczas lista rozwijana Wielomianowy kontroluje to, która kategoria zmiennych przewidywanych będzie wyświetlana. Kolejność wyświetlania wartości na liście jest określona przez specyfikację w ustawieniach Opcje budowania.

Parametry kowariancji: Ten widok kontroluje oszacowania parametrów kowariancji i powiązane statystyki dla efektów reszty i efektów losowych. Są to zaawansowane, ale fundamentalne wyniki, które udostępniają informację na temat tego, czy struktura kowariancji jest odpowiednia.

Tabela podsumowań. Jest to krótki przegląd liczby parametrów w macierzach kowariancji reszt (**R**) i efektów losowych (**G**), ranga (liczba kolumn) w efekcie losowym (**X**) oraz macierzach projektów efektu losowego (**Z**), a także liczba obiektów zdefiniowanych przez zmienne obiektowe, które definiują strukturę danych.

Tabela parametrów kowariancji. Dla wybranego efektu wyświetlany jest wybrany efekt, oszacowanie, błąd standardowy oraz przedział ufności dla każdego parametru kowariancji. Liczba widocznych parametrów jest zależna od struktury kowariancji dla efektu, bloków efektów losowych, a także od liczby efektów w bloku. Jeśli widoczne jest, że parametry, które nie leżą na przekątnej, są nieistotne, być może oznacza to, że możliwe będzie użycie prostszej struktury kowariancji.

Efekty. Jeśli istnieją bloki efektów losowych, wówczas dostępna jest lista rozwijana Efekt, z której można wybrać efekt reszty lub efekt losowy do wyświetlenia. Efekt reszty jest zawsze dostępny.

Grupy. Jeśli blok efektów reszty lub losowych obejmuje specyfikację grupy, wówczas istnieje lista rozwijana Grupa przeznaczona do wyboru poziomu grupy do wyświetlenia.

Wielomianowy. Jeśli obowiązuje rozkład wielomianowy, wówczas lista rozwijana Wielomianowy kontroluje to, która kategoria zmiennych przewidywanych będzie wyświetlana. Kolejność wyświetlania wartości na liście jest określona przez specyfikację w ustawieniach Opcje budowania.

Oszacowane średnie: efekty istotne: Są to wykresy wyświetlane dla 10 „najbardziej istotnych" efektów o wszystkich czynnikach ustalonych, począwszy od interakcji trzeciego rzędu, następnie interakcje drugiego rzędu, a na koniec efekty główne. Na osi pionowej wykresu przedstawiona jest wartość zmiennej przewidywanej oszacowana przez model dla każdej wartości efektu głównego (lub efektu, który znajduje się pierwszy na liście w interakcji) na osi poziomej; osobna linia jest generowana dla każdej wartości drugiego efektu z listy w interakcji; osobny wykres jest generowany dla każdej wartości trzeciego efektu z listy w interakcji trzeciego rzędu; wszystkie pozostałe predyktory są utrzymywane jako stałe. Zapewnia on przydatną wizualizację efektów współczynników docelowej każdego predyktora. Należy zwrócić uwagę, że jeśli żaden predyktor nie jest istotny, nie jest generowana żadna oszacowana średnia.

Ufność. Przedstawiane są górne i dolne granice ufności dla średnich brzegowych — na podstawie poziomu ufności określonego w ramach opcji budowania.

Oszacowane średnie: efekty użytkownika: Są to tabele i wykresy dla żądanych przez użytkownika efektów o wszystkich czynnikach ustalonych.

Style. Dostępne są różne style wyświetlania, które są dostępne z poziomu listy rozwijanej **Styl**.

- **Diagram.** W przypadku tego stylu na osi pionowej wykresu liniowego przedstawiona jest wartość zmiennej przewidywanej oszacowana przez model dla każdej wartości efektu głównego (lub efektu, który znajduje się pierwszy na liście w interakcji) na osi poziomej; osobna linia jest generowana dla każdej wartości drugiego efektu z listy w interakcji; osobny wykres jest generowany dla każdej wartości trzeciego efektu z listy w interakcji trzeciego rzędu; wszystkie pozostałe predyktory są utrzymywane jako stałe.

  Jeśli zażądano kontrastów, wówczas wyświetlany jest inny wykres w celu porównania poziomów zmiennej kontrastu; na potrzeby interakcji wykres jest wyświetlany dla każdego poziomu kombinacji efektów innych niż zmienna kontrastu. W przypadku kontrastów **parami** wykres sieci odległości to graficzne odzwierciedlenie tabeli porównań, w przypadku której odległości między węzłami w sieci odpowiadają różnicom między próbami. Żółte linie odpowiadają różnicom istotnym pod względem statystycznym. Linie czarne odpowiadają różnicom nieistotnym. Ustawienie kursora na sieci powoduje wyświetlenie podpowiedzi zawierającej skorygowaną istotność różnicy między węzłami połączonymi za pomocą linii.

  W przypadku kontrastów **odchylenie** wyświetlany jest wykres słupkowy z oszacowaną przez model wartością zmiennej przewidywanej na osi pionowej oraz wartościami zmiennej kontrastu na osi poziomej; na potrzeby interakcji wyświetlany jest wykres dla każdego poziomu kombinacji efektów innych niż zmienna kontrastu. Słupki przedstawiają różnice między poszczególnymi poziomami zmiennej kontrastu oraz średnią ogólną, którą reprezentuje czarna linia pozioma.

  W przypadku kontrastów **prostych** wyświetlany jest wykres słupkowy z oszacowaną przez model wartością zmiennej przewidywanej na osi pionowej oraz wartościami zmiennej kontrastu na osi poziomej; na potrzeby interakcji wyświetlany jest wykres dla każdego poziomu kombinacji efektów innych niż zmienna kontrastu. Słupki przedstawiają różnice między poszczególnymi poziomami zmiennej kontrastu (za wyjątkiem ostatniego) oraz poziomem ostatnim, który reprezentuje czarna linia pozioma.
- **Tabela.** W tym stylu wyświetlana jest tabela zawierająca oszacowaną przez model wartość zmiennej przewidywanej, jej błąd standardowy oraz przedział ufności dla każdego poziomu kombinacji obowiązujących zmiennych; wszystkie pozostałe predyktory są utrzymywane jako stałe.

  Jeśli żądane były kontrasty, wówczas wyświetlana jest inna tabela z oszacowaniem, błędem standardowym, testem istotności i przedziałem ufności dla każdego kontrastu; na potrzeby interakcji istnieje osobny zestaw wierszy dla

każdego poziomu kombinacji efektów innych niż zmienna kontrastu. Dodatkowo wyświetlana jest tabela z wynikami testu ogólnego; na potrzeby interakcji istnieje osobny test ogólny dla każdego poziomu kombinacji efektów innych niż zmienna kontrastu.

Ufność. Ta opcja umożliwia przełączanie górnych i dolnych przedziałów ufności dla średnich brzegowych — na podstawie poziomu ufności określonego w ramach opcji budowania.

Układ. Ta opcja umożliwia przełączanie układu diagramu kontrastów par. Układ kołowy przedstawia mniej informacji na temat kontrastów niż układ sieciowy, ale eliminuje problem przecinających się linii.

Ustawienia: Podczas oceniania modelu powinny być wygenerowane zaznaczone elementy z tej karty. Przy ocenie modelu zawsze obliczana jest wartość przewidywana (dla wszystkich zmiennych przewidywanych) i ufność (dla jakościowych zmiennych przewidywanych). Obliczona ufność może być oparta na prawdopodobieństwie przewidywanej wartości (najwyższe przewidywane prawdopodobieństwo) lub różnicy między najwyższym przewidywanym prawdopodobieństwem a drugim co do wysokości przewidywanym prawdopodobieństwem.

- **Przewidywane prawdopodobieństwo dla przewidywanych zmiennych jakościowych.** Generuje przewidywane prawdopodobieństwa dla jakościowych zmiennych przewidywanych. Dla każdej kategorii tworzona jest jedna zmienna.
- **Oceny skłonności (ważne tylko dla przewidywanych zmiennych typu flaga).** W przypadku modeli z przewidywaną zmienną typu flaga (zwracających predykcje tak lub nie) można wyliczyć oceny skłonności definiujące prawdopodobieństwo prawdziwego wyniku określonego dla zmiennej przewidywanej. Model generuje surowe oceny skłonności; jeśli stosowane są podzbiory, model generuje także skorygowane oceny skłonności na podstawie podzbioru testowego.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Węzeł GLE

Model GLE znajduje zmienną zależną, która jest związana liniowo z czynnikami i współzmiennymi za pomocą określonej funkcji łączenia. Model pozwala ponadto, aby zmienna zależna nie miała rozkładu normalnego. Dzięki bardzo ogólnej postaci wzoru modelu obejmuje on wiele modeli statystycznych, takich jak regresja liniowa dla odpowiedzi o rozkładzie normalnym, modele logistyczne dla danych binarnych, modele logarytmiczno-liniowe dla danych o liczebności i wiele innych modeli statystycznych

Przykłady. Firma transportowa może używać uogólnionych modeli liniowych do dopasowania regresji Poissona w celu zliczenia uszkodzeń dla kilku typów statków zbudowanych w różnym okresie, a model wynikowy może ułatwić określenie, które typy statków są najbardziej podatne na uszkodzenia.

Firma zajmująca się ubezpieczeniami samochodów może używać uogólnionych modeli liniowych w celu dopasowania regresji gamma do roszczeń związanych z uszkodzeniami samochodów, a model wynikowy może pomóc w ustaleniu czynników, jakie wpłynęły na wysokość większości roszczeń.

Badacze w dziedzinie medycyny mogą używać uogólnionych modeli liniowych do dopasowania regresji komplementarnej log-log do danych przeżycia obciętych przedziałowych w celu ustalenia predykcji czasu potrzebnego do ponownego wystąpienia określonego stanu zdrowia.

Działanie modeli GLE polega na budowaniu równania, które tworzy relację pomiędzy wartościami zmiennych wejściowych a wartościami zmiennych wyjściowych. Po wygenerowaniu modelu może być on używany do oszacowania wartości dla nowych danych.

W przypadku przewidywanej zmiennej jakościowej dla każdego rekordu obliczane jest prawdopodobieństwo członkostwa dla każdej możliwej kategorii wyjściowej. Jako predykowana wartość wyjściowa dla tego rekordu przypisywana jest kategoria zmiennej przewidywanej o najwyższym prawdopodobieństwie.

Wymagania. Wymagana jest co najmniej jedna zmienna wejściowa i dokładnie jedna zmienna przewidywana (której poziom pomiaru może być określony jako *Ilościowa*, *Jakościowa* lub *Flaga*) z co najmniej dwoma kategoriami. Typy zmiennych używanych w modelu muszą być w pełni zrealizowane jako instancje.

Przewidywana

Te ustawienia definiują zmienną przewidywaną, jej rozkład, a także jej relację z predyktorami przez funkcję łączenia.

Przewidywana Zmienna przewidywana jest wymagana. Może mieć dowolny poziom pomiaru, a poziom pomiaru zmiennej przewidywanej wpływa na to, które rozkłady i funkcje łączenia są odpowiednie.

- **Użyj predefiniowanej zmiennej przewidywanej** Wybierz tę opcję, aby wykorzystać ustawienia zmiennej przewidywanej z wcześniejszego węzła typu (lub z karty Typy wcześniejszego węzła źródłowego).
- **Użyj zmiennej przewidywanej użytkownika** Wybierz tę opcję, aby ręcznie przypisać zmienną przewidywaną.
- **Użyj liczby prób jako mianownika** Gdy odpowiedź jest liczbą zdarzeń występujących w zbiorze prób, zmienna przewidywana zawiera liczbę zdarzeń. Można wybrać dodatkową zmienną zawierającą liczbę prób. Na przykład podczas testowania nowego pestycydu można wystawiać mrówki na działanie pestycydu w różnych stężeniach, a następnie rejestrować liczbę mrówek zabitych i liczbę mrówek w każdej próbie. W tym przypadku zmienna rejestrująca liczbę mrówek zabitych powinna być określona jako zmienna przewidywana (zdarzenia), a zmienna rejestrująca liczbę mrówek w każdej próbie powinna być określona jako zmienna prób. Jeśli liczba mrówek jest taka sama dla każdej próby, wówczas liczba prób może być określona przy użyciu wartości stałej.

  Liczba prób powinna być większa niż lub równa liczbie zdarzeń w każdym rekordzie. Zdarzenia powinny być nieujemnymi liczbami całkowitymi, a próby powinny być dodatnimi liczbami całkowitymi.
- **Dostosuj kategorię odniesienia**. W przypadku przewidywanej zmiennej jakościowej można wybrać kategorię odniesienia. To może wpłynąć na niektóre wyniki, takie jak oszacowania parametrów, ale nie powinno zmienić dopasowania modelu. Na przykład, jeśli zmienna przewidywana przyjmuje domyślnie wartości 0, 1 i 2, wówczas procedura ustawia ostatnią kategorię (o najwyższej wartości) — czyli 2 — jako kategorię odniesienia. W tej sytuacji oszacowania parametrów należy interpretować jako mające związek z wiarygodnością kategorii 0 lub 1 *względem* wiarygodności kategorii 2. W wypadku określenia kategorii użytkownika i obecności zdefiniowanych etykiet zmiennej przewidywanej można określić kategorię odniesienia, wybierając wartość z listy. To może być wygodne, jeśli podczas określania modelu użytkownik nie pamięta dokładnie sposobu zakodowania konkretnej zmiennej.

Rozkład zmiennej przewidywanej i powiązanie (Funkcja łączenia) z modelem liniowym Na podstawie wartości predyktorów model oczekuje, że rozkład wartości zmiennej przewidywanej będzie zgodny z określonym kształtem, a w przypadku wartości zmiennej przewidywanej, że będą powiązane liniowo z predyktorami przez określoną funkcję łączenia. Dostępne są skróty dla kilku typowych modeli. Jeśli istnieje konkretna kombinacja rozkładu i funkcji łączenia, dla której użytkownik planuje znaleźć dopasowanie, a która nie jest dostępna na liście skrótów, wówczas można również wybrać ustawienie **Użytkownika**.

- **Model liniowy** Określa rozkład normalny z łączem tożsamości, które jest użyteczne, gdy zmienna przewidywana może zostać przewidziana z użyciem modelu regresji liniowej lub modelu ANOVA.
- **Regresja gamma** Określa rozkład gamma z logarytmiczną funkcją łączenia, który powinien być używany, gdy zmienna przewidywana zawiera wszystkie wartości dodatnie i wykazuje skośność w stronę wyższych wartości.

- **Model logliniowy** Określa rozkład Poissona z logarytmiczną funkcją łączenia, który powinien być używany, gdy zmienna przewidywana reprezentuje liczbę wystąpień w stałym okresie czasu.
- **Ujemna regresja dwumianowa** Określa rozkład ujemny dwumianowy z logarytmiczną funkcją łączenia, który powinien być używany, gdy zmienna przewidywana i mianownik reprezentują liczbę prób wymaganych do zaobserwowania *k* sukcesów.
- **Regresja Tweedie'ego** Określa rozkład Tweedie'ego z tożsamościową, logarytmiczną lub wykładniczą funkcją łączenia. Opcja użyteczna do modelowania odpowiedzi będących kombinacją zer i dodatnich wartości rzeczywistych. Rozkłady te nazywa się także rozkładami *złożonymi Poissona*, *złożonymi gamma* i *Poissona-gamma*.
- **Wielomianowa regresja logistyczna** Określa rozkład wielomianowy, który powinien być używany, gdy zmienna przewidywana jest odpowiedzią wielokategoryjną. Używa skumulowanej funkcji łączenia logit (wyniki porządkowe) lub uogólnionej funkcji łączenia logit (wielokategoryjne odpowiedzi nominalne).
- **Binarna regresja logistyczna** Określa rozkład dwumianowy z funkcją łączenia logit, który powinien być używany, gdy zmienna przewidywana jest odpowiedzią binarną przewidywaną przez model regresji logistycznej.
- **Binarny probit** Określa rozkład dwumianowy z funkcją łączenia probit, który powinien być używany, gdy zmienna przewidywana jest odpowiedzią binarną z bazowym rozkładem binarnym.
- **Przeżycia obcięte przedziałowe** Określa rozkład dwumianowy z funkcją łączenia komplementarny log-log, która jest użyteczna w analizie przeżycia, gdy niektóre obserwacje nie mają zdarzenia kończącego.
- **Użytkownika** Umożliwia użytkownikowi samodzielne określenie kombinacji rozkładu i funkcji łączenia.

Rozkład

Ten wybór określa **Rozkład** zmiennej przewidywanej. Możliwość określenia rozkładu innego niż normalny i nietożsamościowej funkcji łączenia jest istotnym ulepszeniem uogólnionego modelu liniowego w porównaniu do modelu liniowego. Istnieje wiele możliwych kombinacji rozkład-funkcja łączenia, a kilka z nich może być odpowiednich dla dowolnego zbioru danych, dlatego wybór może być zależny od rozważań teoretycznych apriori lub tego, która kombinacja wydaje się zapewniać najlepsze dopasowanie.

- **Automatycznie** Jeśli nie wiesz, którego rozkładu użyć, wybierz tę opcję; węzeł przeanalizuje dane do oszacowania i zastosuje najlepszą metodę rozkładu.
- **Dwumianowy** Ten rozkład jest odpowiedni tylko dla zmiennej przewidywanej, która reprezentuje odpowiedź binarną lub liczbę zdarzeń.
- **Gamma** Ten rozkład jest odpowiedni dla zmiennej przewidywanej z wartościami w skali dodatniej, które wykazują skośność w stronę większych wartości dodatnich. Jeśli wartość danych jest mniejsza niż lub równa 0 lub występuje brak wartości, wówczas dana obserwacja nie jest wykorzystywana w analizie.
- **Odwrócony Gaussa** Ten rozkład jest odpowiedni dla zmiennej przewidywanej z wartościami w skali dodatniej, które wykazują skośność w stronę większych wartości dodatnich. Jeśli wartość danych jest mniejsza niż lub równa 0 lub występuje brak wartości, wówczas dana obserwacja nie jest wykorzystywana w analizie.
- **Wielomianowy** Ten rozkład jest odpowiedni dla zmiennej przewidywanej, która reprezentuje odpowiedź wielokategoryjną. Forma modelu będzie zależna od poziomu pomiaru zmiennej przewidywanej.

  Wynikiem dla **nominalnej** zmiennej przewidywanej będzie nominalny model wielomianowy, w którym szacowany jest osobny zestaw parametrów modelu dla każdej kategorii zmiennej przewidywanej (z wyjątkiem kategorii odniesienia). Oszacowania parametrów dla konkretnego predyktora przedstawiają związek między predyktorem a wiarygodnością dla każdej kategorii zmiennej przewidywanej, względem kategorii odniesienia.

  Wynikiem dla **porządkowej** zmiennej przewidywanej będzie porządkowy model wielomianowy, w którym tradycyjny składnik stałej jest zastępowany przez zestaw parametrów **progowych**, które odnoszą się do prawdopodobieństwo skumulowanego kategorii zmiennej przewidywanej.
- **Ujemny dwumianowy** W regresji ujemnej dwumianowej używany jest ujemny rozkład dwumianowy z logarytmiczną funkcją łączenia, który powinien być używany, gdy zmienna przewidywana reprezentuje liczbę wystąpień o wysokiej wariancji.
- **Normalny** Jest odpowiedni w przypadku ilościowej zmiennej przewidywanej, której wartości przyjmują symetryczny rozkład w kształcie dzwona, z centralną wartością średnią.

- **Poissona** Ten rozkład może być traktowany jako seria wystąpień zdarzenia będącego przedmiotem zainteresowania w ustalonym okresie i jest odpowiedni dla zmiennych o nieujemnych liczbach całkowitych. Jeśli wartość danych nie jest liczbą całkowitą, jest mniejsza od 0 lub występuje brak danych, wówczas dana obserwacja nie jest wykorzystywana w analizie.
- **Tweedie** Ten rozkład jest odpowiedni dla zmiennych, które mogą być reprezentowane przez poissonowsko mieszane rozkłady gamma; rozkład ten jest mieszany, to znaczy że łączy właściwości rozkładu ciągłego (nieujemne wartości rzeczywiste) i dyskretnego (prawdopodobieństwo dodatnie dla pojedynczej wartości, 0). Zmienna zależna musi być liczbowa, z wartościami danych większymi niż lub równymi zero. Jeśli wartość danych jest mniejsza niż zero lub występuje brak danych, wówczas dana obserwacja nie jest wykorzystywana w analizie. Wartość stała parametru rozkładu Tweedie'go może być dowolną liczbą większą niż jeden i mniejszą niż dwa.

Funkcje łączenia

Funkcja łączenia to transformacja zmiennej przewidywanej, która umożliwia estymację modelu. Dostępne są następujące funkcje:

- **Automatycznie** Jeśli nie wiesz, której funkcji łączenia użyć, wybierz tę opcję; węzeł przeanalizuje dane do oszacowania i zastosuje najlepszą funkcję łączenia.
- **Tożsamość** $f(x)=x$. Zmienna przewidywana nie jest transformowana. Ta funkcja łączenia może być używana z dowolnym rozkładem, z wyjątkiem rozkładu wielomianowego.
- **Komplementarny log-log** $f(x)=\log(-\log(1-x))$. Jest odpowiednia tylko w przypadku rozkładu dwumianowego i wielomianowego.
- **Cauchit** $f(x) = \tan(\pi\ (x - 0.5))$. Jest odpowiednia tylko w przypadku rozkładu dwumianowego i wielomianowego.
- **Logarytm** $f(x)=\log(x)$. Ta funkcja łączenia może być używana z dowolnym rozkładem, z wyjątkiem rozkładu wielomianowego.
- **Komplementarny log** $f(x)=\log(1-x)$. Ma zastosowanie tylko w przypadku rozkładu dwumianowego.
- **Logit** $f(x)=\log(x\ /\ (1-x))$. Jest odpowiednia tylko w przypadku rozkładu dwumianowego i wielomianowego.
- **Ujemny log-log** $f(x)=-\log(-\log(x))$. Jest odpowiednia tylko w przypadku rozkładu dwumianowego i wielomianowego.
- **Probit** $f(x)=\Phi^{-1}(x)$, gdzie Φ^{-1} jest odwrotnością funkcji skumulowanego rozkładu standardowego normalnego. Jest odpowiednia tylko w przypadku rozkładu dwumianowego i wielomianowego.
- **Wykładnik** $f(x)=x^{\alpha}$, jeśli $\alpha \neq 0$. $f(x)=\log(x)$, jeśli $\alpha=0$. α jest wymaganą specyfikacją liczbową i musi być liczbą rzeczywistą. Ta funkcja łączenia może być używana z dowolnym rozkładem, z wyjątkiem rozkładu wielomianowego.

Parametr dla rozkładu Tweedie'ego Opcja dostępna tylko w przypadku wybrania przycisku opcji **Regresja Tweedie'ego** lub metody Tweedie w polu **Rozkład**. Wybierz wartość w zakresie od 1 do 2.

Efekty modelu

Współczynniki efektów stałych są zwykle traktowane jako zmienne, których wszystkie wartości badane są reprezentowane w zbiorze danych, i mogą być używane podczas oceniania. Domyślnie zmienne z predefiniowaną rolą w danych wejściowych, które nie są określone w innym miejscu w oknie dialogowym, są wprowadzane w części modelu dotyczącej efektów stałych. Zmienne jakościowe (flaga, nominalne i porządkowe) są używane jako współczynniki w modelu, a zmienne ilościowe są używane jako współzmienne.

Wprowadź efekty do modelu, zaznaczając jedną zmienną lub większą liczbę zmiennych na liście źródłowej i przeciągając do listy efektów. Typ tworzonego efektu jest zależny od tego, do którego obszaru aktywnego zostanie przeciągnięte zaznaczenie.

- **Główne** Zmienne odrzucane są wyświetlane jako osobne efekty główne u dołu listy efektów.
- **2-rzędu** Wszystkie możliwe pary zmiennych odrzucanych pojawiają się jako interakcje drugiego rzędu u dołu listy efektów.
- **3-rzędu** Wszystkie możliwe trójki zmiennych odrzucanych pojawiają się jako interakcje trzeciego rzędu u dołu listy efektów.

- ***** Kombinacja wszystkich zmiennych odrzucanych pojawia się jako pojedyncza interakcja u dołu listy efektów.

Przyciski po prawej stronie kreatora efektów umożliwiają wykonywanie różnych działań.

Tabela 12. Opisy przycisków kreatora efektów

Ikona	Opis
	Umożliwia usuwanie składników z modelu efektów stałych poprzez wybranie składników przeznaczonych do usunięcia i kliknięcie przycisku usuwania.
	Umożliwia reorganizację składników w modelu efektów stałych poprzez wybranie składników przeznaczonych do reorganizacji i kliknięcie strzałki w górę lub w dół.
	Umożliwia dodawanie składników zagnieżdżonych przy użyciu okna dialogowego Dodaj składnik zdefiniowany przez użytkownika poprzez kliknięcie przycisku Dodaj składnik zdefiniowany przez użytkownika.

Uwzględnij wyraz wolny Wyraz wolny jest zwykle uwzględniany w modelu. Jeśli można założyć, że dane przechodzą przez początek układu współrzędnych, wyraz wolny można wyłączyć z modelu.

Dodaj składnik zdefiniowany przez użytkownika

W tej procedurze można zbudować składniki zagnieżdżone dla modelu. Składniki zagnieżdżone są przydatne do modelowania efektu czynników lub współzmiennych, których wartości nie wchodzą w interakcje z poziomami innego czynnika. Na przykład sieć sklepów spożywczych może analizować zwyczaje zakupowe swoich klientów w kilku sklepach. Ponieważ każdy klient bywa regularnie tylko w jednym z tych sklepów, efekt Klient jest *zagnieżdżony* w efekcie Lokalizacja sklepu.

Ponadto można uwzględnić efekty interakcji, takie jak składniki wielomianowe z tą samą współzmienną, lub dodać wiele poziomów zagnieżdżenia do składnika zagnieżdżonego.

Ograniczenia. W odniesieniu do składników zagnieżdżonych obowiązują następujące ograniczenia:

- Wszystkie czynniki w interakcji muszą być unikalne. A zatem, jeśli *A* jest czynnikiem, to niedozwolone jest określenie *A*A*.
- Wszystkie czynniki w efekcie zagnieżdżonym muszą być unikalne. A zatem, jeśli *A* jest czynnikiem, to niedozwolone jest określenie *A(A)*.
- Efekt nie może być zagnieżdżony w obrębie współzmiennej. A zatem, jeśli *A* jest czynnikiem, a *X* jest współzmienną, to określenie *A(X)* jest niedozwolone.

Tworzenie zagnieżdżonego składnika

1. Wybierz czynnik lub współzmienną zagnieżdżony/-ą w innym czynniku, a następnie kliknij przycisk ze strzałką.
2. Kliknij opcję **(W)**.
3. Wybierz czynnik, w którym zagnieżdżony jest poprzedni czynnik lub współzmienna, a następnie kliknij przycisk ze strzałką.
4. Kliknij opcję **Dodaj składnik**.

Opcjonalnie można uwzględnić efekty interakcji lub dodać wiele poziomów zagnieżdżania do zagnieżdżonego składnika.

Waga i przesunięcie

Waga analizy Parametr skali to oszacowanie parametru modelu w odniesieniu do wariancji odpowiedzi. Wagi analiz są wartościami „znanymi", które mogą się różnić między obserwacjami. Jeśli określona jest zmienna **Waga analizy**, to parametr skali, który jest powiązany z wariancją odpowiedzi, jest dzielony przez wartości wagi analizy dla każdej obserwacji. W analizie nie są używane rekordy z wartościami wagi analizy mniejszymi od zera lub równymi zeru ani obserwacje brakujące.

Przesunięcie Składnik przesunięcia jest predyktorem *strukturalnym*. Jego wskaźnik nie jest szacowany przez model, ale przyjmuje się, że ma wartość 1; dlatego wartości przesunięcia są po prostu dodawane do predyktora liniowego zmiennej przewidywanej. Jest to szczególnie przydatne w modelach regresji Poissona, w których każda obserwacja może mieć inny poziom ekspozycji na badane zdarzenie.

Na przykład przy modelowaniu częstości wypadków wśród poszczególnych kierowców należy pamiętać o istotnej różnicy między kierowcą, który spowodował jeden wypadek w ciągu trzech lat, a kierowcą, który spowodował jeden wypadek w ciągu 25 lat. Liczba wypadków może być modelowana jako odpowiedź o rozkładzie Poissona lub odpowiedź o rozkładzie ujemnym dwumianowym z logarytmiczną funkcją łączenia, jeśli logarytm naturalny doświadczenia kierowcy jest uwzględniony w składniku przesunięcia.

Inne kombinacje typów rozkładu i funkcji łączenia będą wymagały przekształcenia zmiennej przesunięcia.

Opcje budowania

Te opcje określają niektóre bardziej zaawansowane kryteria używane do budowania modelu.

Porządek sortowania Te elementy sterujące określają porządek kategorii dla zmiennych przewidywanych i czynników (jakościowych zmiennych wejściowych) na potrzeby określenia „ostatniej" kategorii. Ustawienie porządku sortowania zmiennych przewidywanych jest ignorowane, jeśli zmienne przewidywane nie jest jakościowe lub jeśli w ustawieniach "Przewidywana" na stronie 220 określona jest niestandardowa kategoria odniesienia.

Ustawienia po estymacji Te ustawienia określają sposób obliczania niektórych wyników modelu na potrzeby wyświetlania.

- **Poziom ufności (%)** Jest to poziom ufności używany do wyliczania oszacowań przedziałów współczynników modelu. Należy podać wartość większą od 0 i mniejszą od 100. Domyślna wartość to 95.
- **Stopnie swobody** Określa sposób obliczania stopni swobody dla testów istotności. Wybierz opcję **Ustalone dla wszystkich testów (metoda resztowa)**, jeśli próba jest wystarczająco duża lub dane są zrównoważone, albo model używa kowariancji prostszego typu — na przykład przekątna lub tożsamość skalowana. Jest to ustawienie domyślne. Wybierz opcję **Różne pomiędzy testami (przybliżenie Satterthwaite'a)**, jeśli próba jest niewielka, dane są niezrównoważone lub w modelu używany jest skomplikowany typ kowariancji — na przykład nieustrukturalizowana.
- **Testy efektów stałych i współczynników**. Jest to metoda obliczania macierzy kowariancji oszacowań parametrów. Wybierz mocne oszacowanie, jeśli martwi Cię możliwość naruszenia założeń modelu.

Wykryj wpływające wartości odstające Dla wszystkich rozkładów z wyjątkiem wielomianowego można wybrać tę opcję, aby rozpoznać wpływające wartości odstające.

Przeprowadź analizę trendów W przypadku wykresu rozrzutu wybranie tej opcji powoduje przeprowadzenie analizy trendu.

Oszacowanie

Metoda Wybierz metodę estymacji metodą maksymalnej wiarygodności; dostępne są następujące opcje:

- Ocena Fishera
- Metoda Newtona-Raphsona
- Metoda hybrydowa

Maksymalna liczba iteracji w ocenie Fishera Podaj nieujemną liczbę całkowitą. Wartość 0 określa metodę Newtona-Raphsona. Wartości większe od 0 określają użycie algorytmu oceny Fishera aż do iteracji o numerze *n*, gdzie *n* jest podaną liczbą całkowitą, a następnie metody Newtona-Raphsona.

Metoda parametru skali Wybierz metodę oszacowania parametru skali; dostępne są następujące opcje:

- Estymacja metodą największej wiarygodności
- Wartość ustalona. Można również określić **Wartość**, która ma być używana.
- Dewiancja
- Chi-kwadrat Pearsona

Metoda ujemna dwumianowa Wybierz metodę oszacowania parametru pomocniczego w metodzie ujemnej dwumianowej; dostępne są następujące opcje:

- Estymacja metodą największej wiarygodności
- Wartość ustalona. Można również określić **Wartość**, która ma być używana.

Zbieżność parametru: Zbieżność jest zakładana, jeśli maksymalna zmiana bezwzględna lub względna w oszacowaniach parametru jest mniejsza niż podana wartość, która musi być nieujemna. Kryterium nie jest stosowane, jeśli określona wartość jest równa 0.

Zbieżność logarytmu wiarygodności Zbieżność jest zakładana, jeśli zmiana bezwzględna lub względna w funkcji logarytmu wiarygodności jest mniejsza niż podana wartość, która musi być nieujemna. Kryterium nie jest stosowane, jeśli określona wartość jest równa 0.

Zbieżność Hessego W przypadku specyfikacji **Wartości bezwzględne** zakładana jest zbieżność, jeśli statystyka oparta na macierzy Hessego jest mniejsza niż określona wartość. W przypadku specyfikacji **Względne** zbieżność jest zakładana, jeśli statystyka jest mniejsza niż iloczyn wartości określonej i wartości bezwzględnej logarytmu wiarygodności. Kryterium nie jest stosowane, jeśli określona wartość jest równa 0.

Maksymalna liczba iteracji Możliwe jest określenie maksymalnej liczby iteracji, jakie wykona algorytm. W algorytmie wykorzystywany jest proces z podwójną iteracją, który obejmuje pętlę wewnętrzną i zewnętrzną. Wartość określona dla maksymalnej liczby iteracji ma zastosowanie do obu pętli. Podaj nieujemną liczbę całkowitą. Domyślną wartością jest 100.

Tolerancja osobliwości Ta wartość jest stosowana jako tolerancja podczas kontroli osobliwości. Podaj wartość dodatnią.

Uwaga: Domyślnie używana jest **zbieżność parametru**, gdzie sprawdzana jest maksymalna **bezwzględna** zmiana przy tolerancji 1E-6. To ustawienie może zwracać wyniki różniące się od wyników uzyskiwanych w wersjach wcześniejszych niż wersja 17. W celu odtworzenia wyników z wersji wcześniejszych niż wersja 17 należy użyć opcji **Względne** dla kryterium zbieżności parametru i zachować domyślną wartość tolerancji równą 1E-6.

Wybór modelu

Użyj wyboru modelu lub regularyzacji Zaznacz to pole wyboru, aby aktywować elementy sterujące na tym panelu.

Metoda Wybierz metodę wyboru modelu lub (w wypadku stosowania opcji **Grzbietowa**) regularyzację, która ma być używana. Do wyboru dostępne są następujące opcje:

- **Lasso** Metoda znana również jako regularyzacja L1. Jest szybsza niż Krokowa postępująca w wypadku dużej liczby predyktorów. Ta metoda zapobiega przeuczeniu poprzez narzucenie kary na parametry (zmniejszenie wartości parametrów). Niektóre parametry mogą zostać zredukowane do zera, przez co metoda wyboru nabiera charakteru zmiennego.
- **Grzbietowa** Metoda znana również jako regularyzacja L2. Zapobiega przeuczeniu poprzez narzucenie kary na parametry (zmniejszenie wartości parametrów). Zmniejsza wszystkie parametry proporcjonalnie tak samo, ale żadnego z nich nie eliminuje i nie ma charakteru zmiennego.

- **Elastyczna sieć** Metoda znana również jako regularyzacja L1 + L2. Zapobiega przeuczeniu poprzez narzucenie kary na parametry (zmniejszenie wartości parametrów). Niektóre parametry mogą zostać zredukowane do zera, przez co metoda wyboru nabiera charakteru zmiennego.
- **Krokowa postępująca** W metodzie tej działanie rozpoczyna się bez efektów w modelu. Następnie metoda dodaje oraz usuwa każdorazowo po jednym efekcie do momentu, aż nie można już nic dodać ani usunąć żadnego efektu zgodnie z kryterium krokowym.

Automatycznie wykrywaj interakcje drugiego rzędu Wybierz tę opcję, aby automatycznie wykrywać interakcje drugiego rzędu.

Parametry kary

Te opcje są dostępne tylko wtedy, gdy w polu **Metoda** wybrano opcję Lasso lub Elastyczna sieć.

Automatycznie wybierz parametry kary Jeśli nie wiesz, które parametry kary ustawić, zaznacz to pole wyboru, a węzeł automatycznie wybierze i zastosuje kary.

Parametr kary Lasso Wprowadź parametr kary, który ma być używany, gdy w polu **Metoda** wyboru modelu wybrana jest opcja Lasso.

Parametr kary sieci elastycznej 1 Wprowadź parametr kary L1, który ma być używany, gdy w polu **Metoda** wyboru modelu wybrana jest opcja Sieć elastyczna.

Parametr kary sieci elastycznej 2 Wprowadź parametr kary L2, który ma być używany, gdy w polu **Metoda** wyboru modelu wybrana jest opcja Sieć elastyczna.

Krokowa postępująca

Te opcje są dostępne tylko wtedy, gdy w polu **Metoda** wybrano opcję Krokowa postępująca.

Uwzględnij efekty z wartościami p niemniejszymi niż Określ minimalne prawdopodobieństwo, jakie muszą mieć efekty, aby były uwzględniane w obliczeniach.

Usuń efekty z wartościami p większymi niż Określ maksymalne prawdopodobieństwo, jakie mogą mieć efekty, aby były uwzględniane w obliczeniach.

Modyfikuj maksymalną liczbę efektów w ostatecznym modelu Zaznacz to pole wyboru, aby aktywować pole **Maksymalna liczba efektów**.

Maksymalna liczba efektów Określa maksymalną liczbę efektów dla metody krokowej postępującej.

Modyfikuj maksymalną liczbę kroków Zaznacz to pole wyboru, aby aktywować pole **Maksymalna liczba kroków**.

Maksymalna liczba kroków Określa maksymalną liczbę kroków dla metody krokowej postępującej.

Opcje modelu

Nazwa modelu Można automatycznie generować nazwę modelu na podstawie zmiennej przewidywanej lub podać nazwę **Użytkownika**. Automatycznie wygenerowana nazwa jest nazwą zmiennej docelowej. Jeśli istnieje więcej niż jedna zmienna przewidywana, nazwa modelu składa się z listy nazw zmiennych połączonych ampersandami. Na przykład, jeśli zmienne przewidywane to field1, field2 i field3, nazwa modelu będzie miała postać: *field1 & field2 & field3*.

Oblicz ważność predyktora W przypadku modeli generujących odpowiednią miarę ważności możliwe jest wyświetlenie wykresu wskazującego ważność względną każdego predyktora w procesie estymacji modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Należy zauważyć, że obliczenie ważności predyktora może potrwać dłużej dla niektórych modeli, szczególnie w przypadku pracy z dużymi zbiorami danych, i domyślnie ta opcja dla niektórych modeli jest wyłączona.

Aby uzyskać więcej informacji, zobacz “Ważność predyktorów” na stronie 43.

Model użytkowy GLE

Wynik modelu użytkowego GLE

Po utworzeniu modelu GLE w wynikach dostępne są następujące informacje.

Tabela Informacje o modelu

Tabela Informacje o modelu zawiera kluczowe informacje o modelu. Tabela określa niektóre ustawienia modelu wysokiego poziomu, takie jak:

- Nazwa zmiennej przewidywanej wybranej w węźle Typ lub na karcie Zmienne węzła GLE.
- Wartości procentowe kategorii przewidywanych: modelowane i referencyjne.
- Rozkład prawdopodobieństwa i powiązana funkcja łączenia.
- Zastosowana metoda budowania modelu.
- Liczba predyktorów na wejściu i w modelu końcowym.
- Procentowo określona dokładność klasyfikacji.
- Typ modelu.
- Procentowo określona dokładność modelu, jeśli zmienna przewidywana jest ilościowa.

Podsumowanie rekordów

Tabela podsumowań zawiera informację o liczbie rekordów użytych do dopasowywania modelu oraz o liczbie rekordów wykluczonych. Szczegółowe informacje obejmują liczbę i odsetek rekordów uwzględnionych i wykluczonych, a także liczbę nieważoną, jeśli zastosowano ważenie częstości.

Ważność predyktorów

Wykres Ważność predyktorów przedstawia ważność pierwszych 10 danych wejściowych (predyktorów) w modelu jako wykres słupkowy.

W przypadku, gdy na wykresie jest więcej niż 10 zmiennych, można zmienić wybór predyktorów uwzględnianych na wykresie, korzystając z suwaka pod wykresem. Wskaźniki na suwaku mają stałą szerokość, a każdy znak na suwaku reprezentuje 10 zmiennych. Wskaźniki można przemieszczać wzdłuż suwaka, wyświetlając w ten sposób 10 kolejnych lub poprzednich zmiennych, uporządkowanych według ważności predyktora.

Dwukrotne kliknięcie wykresu powoduje otwarcie osobnego okna dialogowego, w którym można edytować ustawienia wykresu. Można na przykład zmodyfikować cechy, takie jak wielkość wykresu, a także rozmiar i kolor używanych czcionek. Po zamknięciu tego osobnego okna dialogowego do edycji zmiany są odzwierciedlane na wykresie wyświetlanym na karcie Wynik.

Wykres Reszta względem przewidywanych

Wykres ten można wykorzystać do rozpoznania wartości odstających lub diagnozowania nieliniowości bądź niestałej wariancji błędu. Na idealnym wykresie punkty będą rozproszone losowo wokół linii zerowej.

Oczekiwany jest wzór, w którym rozkład reszt dewiancji standaryzowanej względem wartości przewidywanych predyktora liniowego ma średnią wartość równą zero i stały zakres. Oczekiwanym wzorem jest linia pozioma przebiegająca przez zero.

Ustawienia modelu użytkowego GLE

Karta Ustawienia modelu użytkowego GLE umożliwia określenie opcji surowej skłonności oraz generowania kodu SQL podczas oceny modelu. Ta karta jest dostępna tylko po dodaniu modelu użytkowego do strumienia.

Wylicz surowe oceny skłonności W przypadku modeli zawierających zmienne przewidywane typu flaga można zażądać surowych ocen skłonności, które wskazują wiarygodność wyniku prawda określonego dla zmiennej przewidywanej. Te oceny są stosowane dodatkowo obok standardowych wartości predykcji i ufności. Skorygowane oceny skłonności są niedostępne.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji:

- **Domyślne: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Po podłączeniu do bazy danych z zainstalowanym składnikiem Scoring Adapter generuje kod SQL, korzystając ze składnika Scoring Adapter oraz powiązanych funkcji zdefiniowanych przez użytkownika (UDF) i ocenia model w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Węzeł Model Coxa

Regresja Coxa tworzy model predykcyjny dla danych w czasie do zdarzenia. Model generuje funkcję przeżycia przewidującą prawdopodobieństwo, że zdarzenie będące przedmiotem zainteresowania wystąpiło w określonym czasie (*t*) dla danych wartości zmiennych predyktora. Kształt funkcji przeżycia i współczynniki regresji dla predyktorów są szacowane na podstawie obserwowanych obiektów; model może być następnie zastosowany do nowych obserwacji, które zawierają pomiary dla zmiennych predyktora. Należy zauważyć, że informacje z ocenzurowanych obiektów, czyli takich, które nie są przedmiotem zainteresowania podczas obserwacji, są przydatne podczas szacowania modelu.

Przykład. W ramach strategii zapobiegania odejściom klientów, operator telekomunikacyjny jest zainteresowany modelowaniem „czasu do odejścia” w celu określenia czynników charakterystycznych dla klientów, którzy szybko zmieniają operatora. W tym celu wybierana jest losowa próba klientów i z bazy danych pobierane są informacje o czasie współpracy (od jak dawna są lub jak długo byli klientami — w przypadku klientów już nieaktywnych) oraz różne dane demograficzne.

Wymagania. W węźle Model Coxa potrzebna jest dokładnie jedna zmienna przewidywana oraz zmienna czasu przeżycia. Zmienna przewidywana powinna być zakodowana w taki sposób, aby wartość „fałsz” oznaczała przeżycia, a wartość „prawda” oznaczała wystąpienie interesującego nas zdarzenia; zmienna ta musi mieć poziom pomiaru *Flaga* i być składowana jako łańcuch lub liczba całkowita. (W razie potrzeby sposób składowania można przekształcić za pomocą węzła wypełniania lub węzła wyliczeń). Zmienne o roli *Łącznie* lub *Brak* są ignorowane. Typy zmiennych używanych w modelu muszą być w pełni zrealizowane jako instancje. Czas przeżycia może być dowolnym polem liczbowym.

Uwaga: Jeśli puste łańcuchy w zmiennych jakościowych służą jako dane wejściowe do budowy modelu, podczas oceniania modelu regresji Coxa zostaje zgłoszony błąd. Nie należy używać pustych łańcuchów jako danych wejściowych.

Data i czas. Pola typu Data i czas nie mogą być używane bezpośrednio do określenia czasu przeżycia. Jeśli dostępne są zmienne Data i czas, należy wykorzystać je do utworzenia zmiennej zawierającej czasy przeżycia jako różnicę między datą wprowadzenia do badania a datą obserwacji.

Analiza Kaplana-Meiera. Regresję Coxa można stosować bez zmiennych wejściowych. Jest ona wówczas równoważna analizie Kaplana-Meiera.

Opcje zmiennych węzła Model Coxa

Czas przeżycia. Aby możliwe było wykonanie węzła, wybierz zmienną liczbową (z poziomem pomiaru *Ilościowa*). Czas przeżycia określa czas istnienia rekordu objętego predykcją. Na przykład przy modelowaniu czasu do odejścia

klienta byłaby to zmienna zawierająca czas współpracy klienta z operatorem. Data rozpoczęcia współpracy ani data odejścia nie wpływa na model; istotny jest tylko czas między tymi datami.

Czas przeżycia traktuje się jako liczbę bezwymiarową. Należy dopilnować, aby wszystkie zmienne wejściowe były zgodne z przyjętą skalą czasu przeżycia. Na przykład, jeśli badanie ma mierzyć czas odejścia w miesiącach, to jako zmiennej wejściowej należałoby użyć sprzedaży miesięcznej, a nie sprzedaży rocznej. Jeśli dane zawierają daty początku i końca współpracy, a nie czas trwania współpracy, to należy przekodować ta daty na czas trwania w węźle poprzedzającym Model Coxa.

Pozostałe pola w tym oknie dialogowym są polami standardowymi używanymi w programie IBM SPSS Modeler. Więcej informacji można znaleźć w temacie “Opcje zmiennych węzła modelowania” na stronie 31.

Opcje modelu węzła Model Coxa

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Utwórz modele rozdzielone. Tworzy osobny model dla każdej możliwej wartości zmiennych wejściowych, jakie są określone jako zmienne podziału. Więcej informacji można znaleźć w “Budowanie modeli rozdzielonych” na stronie 28.

Metoda. Dostępne są następujące opcje związane z wprowadzaniem predyktorów do modelu:

- **Wprowadzanie.** Jest to metoda domyślna, która wprowadza wszystkie składniki bezpośrednio do modelu. Podczas budowania modelu nie jest wykonywany wybór zmiennych.
- **Krokowa.** Metoda Krokowa wyboru zmiennych buduje model w sposób krokowy, jak wskazuje jej nazwa. Początkowy model jest najprostszym możliwym modelem bez składników (z wyjątkiem stałej). Na każdym kroku składniki, które nie zostały jeszcze dodane do modelu, są oceniane, a jeśli najlepszy składnik znacząco zwiększa jakość predykcji modelu, jest dodawany. Ponadto składniki, które aktualnie znajdują się w modelu, są ponownie oceniane w celu ustalenia, czy dowolny z nich może zostać usunięty bez znaczącego pogorszenia jakości modelu. Jeśli tak jest, składniki są usuwane. Proces jest powtarzany, a inne składniki są dodawane i/lub usuwane. Jeśli nie można już dodać składników w celu poprawy modelu ani nie można usunąć żadnych składników bez pogorszenia jakości modelu, tworzony jest ostateczny model.
- **Krokowa wsteczna.** Metoda Krokowa wsteczna jest przeciwieństwem metody Krokowa. W przypadku tej metody model początkowy zawiera wszystkie składniki jako predyktory. Na każdym kroku składniki w modelu są oceniane, a dowolne składniki, które mogą być usuwane bez znaczącego umniejszania wartości modelu, są usuwane. Ponadto usunięte wcześniej składniki są oceniane ponownie, aby określić, czy najlepszy z nich powoduje znaczące zwiększenie jakości predykcji modelu. Jeśli tak jest, są dodawane ponownie do modelu. Jeśli nie można usunąć więcej składników bez znaczącego zmniejszenia wartości modelu i nie można dodać żadnych składników w celu poprawy modelu, tworzony jest ostateczny model.

Uwaga: metody automatyczne, takie jak Krokowa i Krokowa wsteczna, to wysoce adaptacyjne metody uczenia, które wykazują silną tendencję do przeuczania na podstawie danych uczących. Gdy te metody są stosowane, szczególnie ważne jest sprawdzenie poprawności modelu wynikowego z użyciem nowych danych albo próby testowej utworzonej z użyciem węzła podziału na podzbiory.

Grupy. Wybranie zmiennej grupowej powoduje, że węzeł obliczy osobny model dla każdej kategorii zmiennej. Może to być dowolna zmienna jakościowa (Flaga lub Nominalna) składowana jako łańcuch lub liczba całkowita.

Typ modelu. Istnieją dwie opcje definiowania składników w modelu. Modele **efektów głównych** obejmują tylko pojedyncze zmienne wejściowe i nie testują interakcji (efektów multiplikatywnych) między zmiennymi wejściowymi. Modele **użytkownika** uwzględniają tylko składniki (efekty główne i interakcje) określone przez użytkownika. W przypadku wyboru tej opcji należy użyć listy Składniki modelu, aby dodać lub usunąć składniki w modelu.

Składniki modelu. W przypadku budowania modelu Użytkownika konieczne będzie jawne określenie składników w modelu. Ta lista przedstawia bieżący zestaw składników dla modelu. Przyciski po prawej stronie listy Składniki modelu umożliwiają dodawanie i usuwanie składników modelu.

- Aby dodać składniki do modelu, kliknij przycisk *Dodaje nowe składniki modelu*.
- W celu usuwania składników należy wybrać żądane składniki i kliknąć przycisk *Usuwa wybrane składniki modelu*.

Dodawanie składników do modelu regresji Coxa

W przypadku żądania niestandardowego modelu regresji Coca można dodawać składniki do modelu, klikając przycisk *Dodaje nowe składniki modelu* na karcie Model. Zostanie otwarte nowe okno dialogowe, w którym można określać składniki.

Typ dodawanego składnika. Istnieje kilka sposobów dodawania składników do modelu w zależności od wybranych zmiennych wejściowych na liście Dostępne zmienne.

- **Interakcja jednostkowa.** Umożliwia wstawienie składnika reprezentującego interakcję wszystkich wybranych zmiennych.
- **Efekty główne.** Wstawia jeden składnik efektu głównego (samą zmienną) dla każdej wybranej zmiennej wejściowej.
- **Wszystkie interakcje 2. rzędu.** Umożliwia wstawienie składnika interakcji 2. rzędu (jest to iloczyn zmiennych wejściowych) dla każdej możliwej pary wybranych zmiennych wejściowych. Na przykład, jeśli wybierzesz zmienne wejściowe *A*, *B*, oraz *C* na liście Dostępne zmienne, ta metoda spowoduje wstawienie składników *A * B*, *A * C* oraz *B * C*.
- **Wszystkie interakcje 3. rzędu.** Umożliwia wstawienie składnika interakcji 3. rzędu (jest to iloczyn zmiennych wejściowych) dla każdej możliwej kombinacji wybranych zmiennych wejściowych, pobieranych po trzy jednocześnie. Na przykład, jeśli wybierzesz zmienne wejściowe *A*, *B*, *C* oraz *D* na liście Dostępne zmienne, ta metoda spowoduje wstawienie składników *A * B * C*, *A * B * D*, *A * C * D* oraz *B * C * D*.
- **Wszystkie interakcje 4. rzędu.** Umożliwia wstawienie składnika interakcji 4. rzędu (jest to iloczyn zmiennych wejściowych) dla każdej możliwej kombinacji wybranych zmiennych wejściowych, pobieranych po cztery jednocześnie. Na przykład, jeśli wybierzesz zmienne wejściowe *A*, *B*, *C*, *D* oraz *E* na liście Dostępne zmienne, ta metoda spowoduje wstawienie składników *A * B * C * D*, *A * B * C * E*, *A * B * D * E*, *A * C * D * E* oraz *B * C * D * E*.

Dostępne zmienne. Ta opcja przedstawia listę dostępnych zmiennych wejściowych, która może być używana w celu konstruowania składników modelu. Lista może zawierać zmienne niebędące dozwolonymi zmiennymi wejściowymi, zatem należy dopilnować, aby wszystkie składniki modelu obejmowały tylko zmienne wejściowe.

Podgląd. Przedstawia składniki, które zostaną dodane do modelu w przypadku kliknięcia opcji **Wstaw** — odpowiednio do wybranych powyżej zmiennych i typów składników.

Wstaw. Umożliwia wstawienie składników do modelu (na podstawie aktualnie wybranych zmiennych oraz typu składnika) i powoduje zamknięcie okna dialogowego.

Opcje zaawansowane węzła Model Coxa

Zbieżność. Te opcje umożliwiają sterowanie parametrami zbieżności modelu. Podczas wykonywania modelu ustawienia zbieżności kontrolują liczbę powtórzonych uruchomień różnych parametrów w celu sprawdzenia ich dopasowania. Im częściej parametry są wypróbowywane, tym bliższe będą wyniki (co oznacza, że wyniki uzyskają zbieżność). Więcej informacji można znaleźć w temacie “Kryteria zbieżności węzła Model Coxa” na stronie 231.

Wynik. Te opcje umożliwiają zażądanie dodatkowych statystyk i wykresów, w tym krzywej przeżycia, które będą wyświetlane w zaawansowanych wynikach modelu wygenerowanego przez węzeł. Więcej informacji można znaleźć w temacie “Zaawansowane opcje wyników węzła Model Coxa” na stronie 231.

Krokowa. Te opcje umożliwiają określanie kryteriów dodawania i usuwania zmiennych w przypadku krokowej metody estymacji. (Przycisk jest wyłączony, jeśli wybrana jest metoda Wprowadzanie). Więcej informacji można znaleźć w temacie “Kryteria metod krokowych węzła Model Cox” na stronie 231.

Kryteria zbieżności węzła Model Coxa

Maksymalna liczba iteracji. Umożliwia określenie maksymalnej liczby iteracji dla modelu. Wartość ta wpływa na czas poszukiwania rozwiązania.

Zbieżność logarytmu wiarygodności. Iteracje zatrzymują się, gdy logarytm wiarygodności jest niższy niż ta wartość. Kryterium nie jest stosowane, jeśli wartość wynosi 0.

Zbieżność parametru. Iteracje zatrzymują się, jeśli zmiana bezwzględna lub względna w estymacjach parametru jest mniejsza niż ta wartość. Kryterium nie jest stosowane, jeśli wartość wynosi 0.

Zaawansowane opcje wyników węzła Model Coxa

Statystyki. Możliwe jest uzyskanie statystyk parametrów modelu, w tym przedziałów ufności dla exp(*B*) i korelacji oszacowań. Można zażądać tych statystyk w każdym lub tylko w ostatnim kroku.

Wyświetl funkcję bazową. Umożliwia wyświetlenie funkcji linii bazowej hazardu i skumulowanego przeżycia przy średniej współzmiennych.

Wykresy

Wykresy mogą okazać się pomocne w ewaluacji oszacowanego modelu i interpretacji wyników. Można wykreślić funkcję przeżycia, hazardu, log-minus-log i jeden-minus-przeżycie.

- *Przeżycie*. Umożliwia wyświetlenie funkcji skumulowanego przeżycia na skali liniowej.
- *Hazard*. Umożliwia wyświetlenie funkcji skumulowanego przeżycia na skali liniowej.
- **Log minus log.** Przedstawia skumulowane przeżycie po zastosowaniu transformacji ln(-ln) do oszacowania.
- *Jeden minus przeżycie*. Umożliwia wykreślenie funkcji Jeden minus przeżycie na skali liniowej.

Kreśli osobną linię dla każdej wartości. Ta opcja jest dostępna wyłącznie w przypadku zmiennych jakościowych.

Wartość przedstawiana na wykresach. Ponieważ funkcje te zależą od wartości predyktorów, należy jako predyktorów użyć wartości stałych, aby wykreślić funkcje względem czasu. Domyślnie jako stałej używa się średniej wartości danego predyktora, ale w tabeli można wprowadzić własne wartości. W przypadku predyktorów jakościowych używane jest kodowanie wskaźnikami, zatem istnieje współczynnik regresji dla każdej kategorii z wyjątkiem ostatniej. Zatem predyktor jakościowy ma wartość średnią dla kontrastu każdego wskaźnika, a wartość ta równa jest odsetkowi obserwacji w kategorii odpowiadających kontrastowi wskaźnika.

Kryteria metod krokowych węzła Model Cox

Kryterium usuwania. W celu uzyskania bardziej odpornego modelu należy wybrać opcję **Iloraz wiarygodności**. Aby skrócić czas wymagany do zbudowania modelu, można spróbować użyć opcję **Walda**. Dodatkowa opcja **Warunek** zapewnia testowanie usuwania na podstawie prawdopodobieństwa ilorazu wiarygodności wyliczonego na podstawie ocen parametrów warunkowych.

Wartości graniczne istotności dla kryteriów. Ta opcja pozwala określenie kryteriów wyboru na podstawie prawdopodobieństwa statystycznego (wartości *p*) skojarzonego z poszczególnymi zmiennymi. Zmienne będą dodawane do modelu, pod warunkiem że powiązana wartość *p* jest mniejsza niż wartość **Wprowadzanie** i będą usuwane, jeśli wartość *p* jest większa niż wartość **Usunięcie**. Wartość **Wprowadzanie** musi być mniejsza niż wartość **Usunięcie**.

Opcje ustawień węzła Model Coxa

Przewidywanie przeżycia w przyszłości. Określ co najmniej jeden czas w przyszłości. Przeżycie, tj. informacja o tym, czy dana obserwacja prawdopodobnie przetrwa co najmniej określony czas (od teraz) bez wystąpienia zdarzenia końcowego, przewidywane jest dla każdego rekordu i każdej wartości czasu, tj. powstaje po jednej predykcji dla każdej wartości czasu. Należy pamiętać, że przeżyciu odpowiada wartość „fałsz” zmiennej przewidywanej.

- **Regularne przedziały.** Wartości czasu przeżycia są generowane na podstawie określonej wartości **Przedział czasowy** i **Liczba okresów czasu do oceny**. Na przykład, jeśli zażądamy 3 okresów do oceny, a przedział między

każdym z czasów wyniesie 2, to przeżycie zostanie przewidziane dla następujących czasów w przyszłości: 2, 4, 6. Każdy rekord jest ewaluowany z tymi samymi wartościami czasu.

- **Zmienna czasu.** Czasy przeżycia dla każdego rekordu podane są w wybranej zmiennej czasu (generowana jest jedna zmienna predykcyjna), zatem każdy rekord może być poddawany ewaluacji z różnymi czasami.

Poprzedni czas przeżycia. Określ dotychczasowy czas przeżycia rekordu, na przykład czas dotychczasowej współpracy klienta z operatorem, jako zmienną. Ocena prawdopodobieństwa przeżycia w przyszłości będzie uwarunkowana poprzednim czasem przeżycia.

Uwaga: wartości przyszłego i dotychczasowego czasu przeżycia muszą mieścić się w przedziale czasów przeżycia występujących w danych użytych do uczenia modelu. Rekordy, których czas wykracza poza ten przedział, otrzymują ocenę null.

Dołącz wszystkie prawdopodobieństwa. Określa, czy prawdopodobieństwa dla poszczególnych kategorii zmiennych wyjściowych są dodawane do poszczególnych rekordów przetwarzanych przez węzeł. Jeśli ta opcja nie zostanie wybrana, wówczas zostanie dodane tylko prawdopodobieństwo przewidywanej kategorii. Prawdopodobieństwa są obliczane dla każdego czasu w przyszłości.

Oblicz funkcję skumulowanego hazardu. Określa, czy do każdego rekordu ma być dodawana wartość skumulowanego hazardu. Skumulowany hazard jest obliczany dla każdego czasu w przyszłości.

Model użytkowy Coxa

Modele regresji Coxa odzwierciedlają równania oszacowane przez węzły Coxa. Zawierają one wszystkie informacje przechwytywane przez model, a także informacje dotyczące struktury i wydajności modelu.

W przypadku uruchomienia strumienia zawierającego model użytkowy regresji Coxa węzeł dodaje dwie nowe zmienne zawierające predykcję modelu i powiązane prawdopodobieństwo. Nazwy nowych zmiennych są wywodzone z nazwy zmiennej przewidywanej uzupełnionej o przedrostek *$C-* dla przewidywanej kategorii i *$CP-* dla powiązanego prawdopodobieństwa oraz o przyrostek z numerem przyszłego przedziału czasu lub nazwą zmiennej czasu definiującej przedział. Na przykład, jeśli dla zmiennej przewidywanej *churn* są zdefiniowane dwa przedziały czasowe w regularnych odstępach, nowe zmienne otrzymają nazwy *$C-churn-1*, *$CP-churn-1*, *$C-churn-2* i *$CP-churn-2*. Jeśli przyszłe czasy są zdefiniowane za pomocą zmiennej *tenure*, to nowe zmienne będą miały nazwy *$C-churn_tenure* i *$CP-churn_tenure*.

Jeśli w węźle modelu Coxa wybrano opcję **Dołącz wszystkie prawdopodobieństwa**, to dla każdego czasu w przyszłości zostaną dodane jeszcze dwie zmienne zawierające prawdopodobieństwo przeżycia i porażki dla każdego rekordu. Nazwy tych dodatkowych zmiennych oparte są na nazwie zmiennej przewidywanej, która uzupełniana jest o przedrostek *$CP-<wartość fałsz>-* dla prawdopodobieństwa przeżycia i *$CP-<wartość prawda>-* dla prawdopodobieństwa zajścia zdarzenia, a także o przyrostek z numerem przedziału czasu w przyszłości. Na przykład dla zmiennej przewidywanej, w której „fałsz” zakodowany jest jako 0, a „prawda” jako 1, a dwa przyszłe przedziały czasowe zdefiniowane są w regularnych odstępach, nowe zmienne będą miały nazwy *$CP-0-1*, *$CP-1-1*, *$CP-0-2* i *$CP-1-2*. Jeśli czasy w przyszłości są zdefiniowane za pomocą jednej zmiennej czasu *tenure*, to nowe zmienne otrzymałyby nazwy *$CP-0-1* i *$CP-1-1*, ponieważ istnieje tylko jeden przyszły przedział czasu.l

Jeśli w węźle modelu Coxa wybrano opcję **Oblicz funkcję skumulowanego hazardu**, to dla każdego czasu w przyszłości zostanie dodana zmienna zawierająca funkcję skumulowanego hazardu dla każdego rekordu. Nazwy tych dodatkowych zmiennych tworzone są na podstawie nazwy zmiennej przewidywanej, która uzupełniana jest o przedrostek *$CH-* oraz o przyrostek zawierający numer przyszłego przedziału czasu lub nazwę zmiennej czasu, która definiuje przedział. Na przykład dla zmiennej przewidywanej o nazwie *churn* i dwóch przyszłych przedziałów czasu zdefiniowanych w regularnych odstępach nowe zmienne będą miały nazwy *$CH-churn-1* i *$CH-churn-2*. Jeśli czasy w przyszłości są zdefiniowane za pośrednictwem zmiennej czasu *tenure*, nowa zmienna otrzyma nazwę *$CH-churn-1*.

Ustawienia wyników regresji Coxa

Z wyjątkiem opcji generowania kodu SQL karta Ustawienia modelu użytkowego zawiera te same elementy sterujące, co karta Ustawienia węzła modelu. Wartości domyślne elementów sterujących modelu użytkowego zależą od wartości ustawionych w węźle modelu. Więcej informacji można znaleźć w temacie “Opcje ustawień węzła Model Coxa” na stronie 231.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

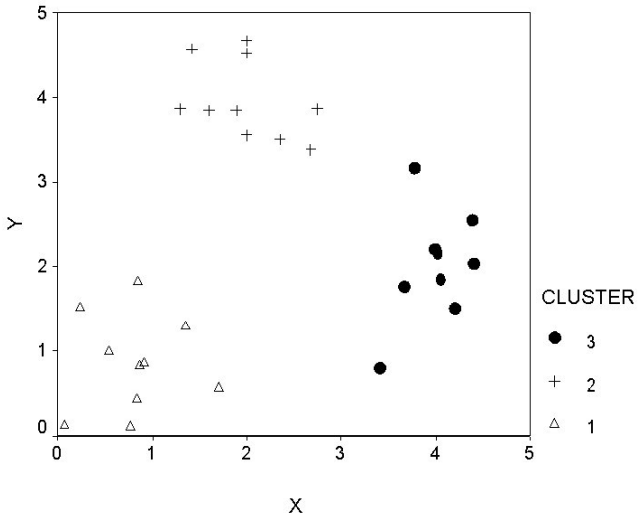
Zaawansowane wyniki regresji Coxa

Zaawansowane wyniki regresji Coxa zawierają szczegółowe informacje o oszacowanym modelu i jego działaniu, w tym krzywą przeżycia. Większość informacji zawartych w zaawansowanych wynikach ma charakter ściśle techniczny i do ich interpretacji niezbędna jest szczegółowa znajomość metody regresji Coxa.

Rozdział 11. Modele skupień

Modele skupień koncentrują się na identyfikacji grup o podobnych rekordach i oznaczaniu rekordów etykietami zgodnie z grupą, do której należą. Nie wykorzystują przy tym żadnej wcześniejszej wiedzy o grupach i ich cechach. Nie musimy nawet wiedzieć dokładnie, ilu grup szukamy. To właśnie odróżnia modele skupień od innych technik uczenia maszynowego — brak wstępnie zdefiniowanego wyniku czy zmiennej przewidywanej dla modelu objętego predykcją. Modele takie często nazywane są modelami **uczenia nienadzorowanego**, ponieważ nie istnieje zewnętrzna norma oceny jakości klasyfikacji. W przypadku takich modeli nie ma *poprawnych* czy *niepoprawnych* odpowiedzi. Ich wartość określana jest przez zdolność do przechwytywania interesujących skupień w danych i oferowania użytecznych opisów tych skupień.

Metody analizy skupień oparte są na pomiarze odległości między rekordami i między skupieniami. Rekordy przypisywane są do skupień w taki sposób, by odległości między rekordami w tym samym klastrze były jak najmniejsze.



Rysunek 44. Prosty model skupień

Dostępne są następujące metody analizy skupień:

Węzeł Metoda k-średnich grupuje zbiór danych w osobne grupy (lub skupienia). Metoda ta definiuje stałą liczbę skupień, w sposób iteracyjny przypisuje rekordy do skupień i dopasowuje centra skupień do chwili, gdy dalsze pokrycie nie będzie miało wpływu na ulepszenie modelu. Zamiast prób predykcji danych wynikowych *k*-średnia korzysta z procesu znanego jako nienadzorowane uczenie w celu ujawnienia wzorców w zbiorze zmiennych wejściowych.

Węzeł Dwustopniowa korzysta z dwustopniowej metody grupowania. Pierwszy krok stanowi pojedynczy przebieg danych z myślą o kompresji surowych danych wejściowych w łatwy w zarządzaniu zestaw podgrup. Drugi krok korzysta z hierarchicznej metody grupowania w celu progresywnego scalania podgrup w coraz większe grupy. Metoda Dwustopniowa oferuje korzyści wynikające z automatycznego szacowania optymalnej liczby grup na potrzeby danych szkoleniowych. Pozwala ona skutecznie obsługiwać mieszane typy zmiennych i duże zbiory danych.

Węzeł Sieć Kohonena generuje typ sieci neuronowej, którą można wykorzystać do grupowania zbioru danych w osobne grupy. Po pełnym przeszkoleniu sieci rekordy podobne do siebie powinny znajdować się blisko siebie na mapie wyników, podczas gdy rekordy różne od siebie powinny znajdować się daleko od siebie. Na podstawie liczby obserwacji przechwyconych przez każdą jednostkę w modelu użytkowym można rozpoznać silne jednostki. Może to dać pojęcie o odpowiedniej liczbie skupień.

Węzeł Hierarchical Density-Based Spatial Clustering (HDBSCAN)© korzysta z algorytmu uczenia nienadzorowanego, aby wyszukać skupienia lub regiony o dużej gęstości w zbiorze danych. Węzeł HDBSCAN w programie SPSS Modeler eksponuje tylko funkcje podstawowe i najczęściej używane parametry biblioteki HDBSCAN. Węzeł jest zaimplementowany w języku Python i można go użyć do skupiania zbioru danych w osobne grupy, jeśli nie wiemy z góry, co to są za grupy.

Modele skupień są często używane do tworzenia grup lub segmentów, które mogą stanowić dane wejściowe do dalszej analizy. Typowym przykładem jest segmentacja rynku stosowana przez marketerów do podziału całego rynku na homogeniczne podgrupy. Każdy segment ma specyficzne cechy wpływające na powodzenie działań marketingowych kierowanych do tego segmentu. Optymalizując strategię marketingową przy wykorzystaniu technik eksploracji danych, można zwykle znacząco udoskonalić model, rozpoznając segmenty i wykorzystując informacje o segmentach w modelach predykcyjnych.

Węzeł Kohonena

Sieć Kohonena to rodzaj sieci neuronowej, która przeprowadza grupowanie, znanej również jako **sieć k** lub **mapa samoorganizująca**. Sieć tego typu może być używana do grupowania zbioru danych w osobne grupy, jeśli nie wiadomo, czym są te grupy na początku. Rekordy są grupowane, tak aby rekordy należące do jednej grupy lub skupienia były do siebie podobne, a rekordy należące do różnych grup były do siebie niepodobne.

Podstawowymi jednostkami są **neurony**, które znajdują się na dwóch warstwach: na **warstwie wejściowej** i **warstwie wyjściowej** (zwanej również **mapą wyników**). Wszystkie neurony wejściowe są połączone ze wszystkimi neuronami wyjściowymi, a z połączeniami tymi powiązana jest **siła** lub **waga**. W czasie uczenia poszczególne jednostki konkurują z pozostałymi, aby „wygrać” dany rekord.

Mapa wynikowa jest dwuwymiarową siatką neuronów bez połączeń między jednostkami.

Dane wejściowe są przedstawiane na warstwie wejściowej, a wartości są umieszczane na warstwie wyjściowej. Neuron wyjściowy z najsilniejszą odpowiedzią jest **zwycięzcą** i stanowi odpowiedź dla zmiennej wejściowej.

Początkowo wszystkie wagi mają charakter losowy. Kiedy jednostka wygrywa rekord, jej wagi (wraz z wagami innych sąsiednich jednostek, zbiorczo zwanymi **sąsiedztwem**) zostają skorygowane, tak aby były lepiej dopasowane do wzorca wartości predykcyjnych dla danego rekordu. Wyświetlane są wszystkie rekordy wejściowe, a wagi zostają odpowiednio zaktualizowane. Ten proces jest wielokrotnie powtarzany, dopóki zmiany są niewielkie. W czasie uczenia wagi jednostek siatki są korygowane do momentu, aż utworzą dwuwymiarową „mapę” skupień (stąd termin **mapa samoorganizująca**).

Po pełnym przeszkoleniu sieci rekordy podobne do siebie powinny znajdować się blisko siebie na mapie wyników, podczas gdy rekordy bardzo różniące się powinny znajdować się daleko od siebie.

W przeciwieństwie do większości metod uczenia w produkcie IBM SPSS Modeler w sieciach Kohonena *nie* są stosowane zmienne przewidywane. Sposób uczenia bez zmiennej przewidywanej jest nazywany **uczeniem nienadzorowanym**. Zamiast prób przewidzenia danych wynikowych sieci Kohonena próbują ujawnić schematy w zestawie zmiennych wejściowych. Zwykle sieć Kohonena obejmuje kilka jednostek, które podsumowują wiele obserwacji (jednostek **silnych**), oraz kilka jednostek, które w rzeczywistości nie odpowiadają żadnym obserwacjom (**słabe** jednostki). Jednostki silne (i czasami inne jednostki sąsiadujące z nimi w siatce) reprezentują prawdopodobne centra skupień.

Inne zastosowanie sieci Kohonena polega na **redukcji wymiarów**. Charakterystyka przestrzenna sieci dwuwymiarowej zapewnia odwzorowanie predyktorów źródłowych *k* na dwa predyktory pochodne, które zachowują relacje podobieństwa do predyktorów źródłowych. W niektórych przypadkach takie działanie może przynieść korzyści takie same, jak analiza czynnikowa lub PCA.

Należy zwrócić uwagę na to, że metoda obliczania domyślnego rozmiaru siatki wynikowej uległa zmianie w porównaniu do metody z poprzednich wersji produktu IBM SPSS Modeler. Nowa metoda zwykle powoduje uzyskiwanie mniejszych warstw wynikowych, których nauka przebiega szybciej, a uogólnianie zwraca lepsze wyniki. Jeśli okaże się, że rozmiar domyślny zapewnia słabe wyniki, należy podjąć próbę powiększenia rozmiaru siatki wynikowej na karcie Zaawansowane. Więcej informacji można znaleźć w temacie “Opcje zaawansowane węzła Kohonena” na stronie 238.

Wymagania. W celu uczenia sieci Kohonena wymagana jest jedna lub większa liczba zmiennych z rolą ustawioną na *Dane wejściowe*. Zmienne z rolą ustawioną na wartość *Zmienna przewidywana*, *Łącznie* lub *Brak* są ignorowane.

Mocne strony. W celu zbudowania modelu sieci Kohonena nie są wymagane dane przynależne do grupy. Nie jest wymagana nawet znajomość liczby poszukiwanych grup. Sieci Kohonena na początku zawierają duże liczby jednostek i w miarę postępu uczenia jednostki wykazują tendencję do tworzenia naturalnych skupień w danych. W celu zidentyfikowania jednostek silnych można przyjrzeć się liczbom obserwacji przechwyconych przez poszczególne jednostki w modelu użytkowym, co umożliwia przybliżone ustalenie właściwej liczby skupień.

Opcje modelu węzła Kohonena

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Kontynuuj uczenie istniejącego modelu. Domyślnie każdorazowe wykonanie węzła Kohonena powoduje utworzenie całkowicie nowej sieci. W przypadku wyboru tej opcji uczenie jest kontynuowane z ostatnią siecią pomyślnie wygenerowaną przez węzeł.

Pokaż wykres sprzężenia zwrotnego. W przypadku wyboru tej opcji podczas uczenia wyświetlana jest wizualna reprezentacja tablicy dwuwymiarowej. Siłę każdego węzła reprezentuje kolor. Kolor czerwony oznacza jednostkę, która wygrywa wiele rekordów (**silnia** jednostka), a kolor biały oznacza jednostkę, która wygrywa niewiele rekordów lub w ogóle nie wygrywa rekordów (**słaba** jednostka). Możliwe, że sprzężenie zwrotne nie będzie wyświetlane, jeśli czas wykorzystany na zbudowanie modelu jest względnie krótki. Należy zwrócić uwagę na to, że ta właściwość może wydłużyć czas uczenia. W celu przyspieszenia uczenia należy usunąć zaznaczenie tej opcji.

Zatrzymywanie uczenia. Domyślnie kryterium zatrzymywania zatrzymuje uczenie w zależności od parametrów wewnętrznych. Jako kryterium zatrzymywania można określić czas. Wprowadź czas (w minutach) uczenia sieci.

Ustaw wartość początkową generatora liczb losowych. Jeśli nie ustawiono wartości początkowej generatora liczb losowych, sekwencja losowych wartości używana do zainicjowania wag sieci będzie różna dla każdego wykonania węzła. W wyniku tego węzeł może utworzyć różne modele dla różnych przebiegów, nawet jeśli ustawienia węzła i wartości danych będą dokładnie takie same. Wybierając tę opcję można ustawić konkretną wartość początkową generatora liczb losowych, dzięki czemu model będzie dokładnie powtarzalny. Konkretna wartość początkowa generatora liczb losowych zawsze będzie powodowała wygenerowanie takiej samej sekwencji wartości losowych, w wyniku czego wykonanie węzła zawsze spowoduje wygenerowanie takiego samego modelu.

Uwaga: Jeśli używana jest opcja **Ustaw wartość początkową generatora liczb losowych** w przypadku rekordów odczytanych z bazy danych, przed przeprowadzeniem próby konieczne może być sortowanie węzła, aby po każdym wykonaniu węzła uzyskany wynik był taki sam. Wynika to z faktu, że wartość początkowa generatora liczb losowych zależy od kolejności rekordów, która w relacyjnej bazie danych nie musi pozostawać jednakowa.

Uwaga: Jeśli wymagane jest uwzględnienie zmiennych nominalnych w modelu, ale pojawiły się problemy z pamięcią podczas budowania modelu lub budowanie modelu trwa zbyt długo, należy rozważyć zarejestrowanie dużych zmiennych nominalnych, aby zmniejszyć liczbę wartości, lub należy rozważyć użycie innej zmiennej z mniejszą liczbą wartości zamiast zmiennej typu dużego zestawu. Na przykład, jeśli pojawiły się problemy ze zmienną *product_id* zawierającą wartości dla pojedynczych produktów, można rozważyć usunięcie jej z modelu i dodanie zamiast niej mniej szczegółowej zmiennej *product_category*.

Optymalizuj ze względu na. Można tutaj wybrać opcje służące do optymalizacji wydajności podczas tworzenia modelu.

- Zaznaczenie opcji **Szybkość** uniemożliwia przenoszenie danych na dysk, co przekłada się na większą wydajność.
- Zaznaczenie opcji **Pamięć** sprawia, że algorytm przenosi dane na dysk kosztem szybkości przetwarzania. Ta opcja jest wybrana domyślnie.

  Uwaga: Podczas uruchamiania trybu analizy rozproszonej to ustawienie może być przesłonięte przez opcje administratora określone w pliku options.cfg.

Dołącz etykietę grupy. Ta opcja jest wybierana domyślnie dla nowych modeli ale nie jest wybierana w przypadku modeli załadowanych z wcześniejszych wersji produktu IBM SPSS Modeler. Ta opcja powoduje utworzenie pojedynczej jakościowej zmiennej oceny, która jest tworzona dla k-średnich oraz dla węzłów dwustopniowych. Ta zmienna łańcuchowa jest używana w węźle Auto Grupowanie podczas obliczania miar rangowania dla różnych typów modeli. Więcej informacji można znaleźć w temacie “Węzeł Auto Grupowanie” na stronie 76.

Opcje zaawansowane węzła Kohonena

Użytkownikom, którzy posiadają szczegółową wiedzę na temat sieci Kohonena, opcje zaawansowane umożliwiają precyzyjne dostosowywanie procesu uczenia. W celu uzyskania dostępu do opcji zaawansowanych należy ustawić opcję Tryb na wartość **Zaawansowany** na karcie Zaawansowany.

Szerokość i długość. Określ rozmiar (szerokość i długość) dwuwymiarowej mapy wyjściowej jako liczbę jednostek wynikowych wzdłuż każdego wymiaru.

Zmiana współczynnika uczenia. Wybierz liniową lub wykładniczą zmianę współczynnika uczenia. **Współczynnik uczenia** to czynnik wagi, który zmniejsza się w miarę upływu czasu, co oznacza, że na początku sieć koduje wielkoskalowe właściwościach danych, a następnie skupia się na drobniejszych właściwościach.

Faza 1 i faza 2. Uczenie sieci Kohonena jest podzielone na dwie fazy. Faza 1 jest fazą szacowania zgrubnego i jest używana do przechwytywania większych wzorców w danych. Faza 2 jest fazą precyzyjnego dostosowywania i jest używana do modyfikowania mapy w celu modelowania drobniejszych właściwości danych. Dla każdej fazy istnieją trzy parametry:

- **Sąsiedztwo.** Ustawia rozmiar początkowy (promień) sąsiedztwa. Określa to liczbę „sąsiednich” jednostek, które są aktualizowane wraz z jednostką wygrywającą podczas uczenia. Podczas fazy 1 rozmiar sąsiedztwa zaczyna się od *sąsiedztwa z fazy 1* i zmniejsza się do (*sąsiedztwa z fazy 2* + 1). Podczas fazy 2 rozmiar sąsiedztwa zaczyna się od *sąsiedztwa z fazy 2* i zmniejsza się do 1,0. *Sąsiedztwo z fazy 1* powinno być większe niż *sąsiedztwo z fazy 2*.
- **Początkowe Eta.** Ustawia początkową wartość współczynnika uczenia **eta**. Podczas fazy 1 eta zaczyna się od *Początkowe Eta, faza 1* i zmniejsza się do wartości *Początkowe Eta, faza 2*. Podczas fazy 2 eta zaczyna się od *Początkowe Eta, faza 2* i zmniejsza się do 0. *Początkowe Eta, faza 1* powinno być większe niż *Początkowe Eta, faza 2*.
- **Epoki.** Ustawia liczbę epok dla każdej fazy uczenia. Każda faza jest kontynuowana przez określoną liczbę przejść.

Modele użytkowe Kohonena

Modele użytkowe Kohonena zawierają wszystkie informacje zebrane przez wytrenowaną sieć Kohonena oraz informacje o architekturze tej sieci.

Po uruchomieniu strumienia zawierającego model użytkowy Kohonena węzeł dodaje dwie nowe zmienne zawierające współrzędne *X* i *Y* tej jednostki w wynikowej siatce Kohonena, która najsilniej odpowiedziała na dany rekord. Nazwy nowych zmiennych wywiedzione są z nazwy modelu poprzedzonej przedrostkami *$KX-* i *$KY-*. Na przykład, jeśli model nosi nazwę *Kohonen*, to nowe zmienne otrzymałyby nazwy *$KX-Kohonen* i *$KY-Kohonen*.

Aby uzyskać bliższe informacje o kodowaniu danej sieci Kohonena, kliknij kartę Model w przeglądarce modeli użytkowych. Spowoduje to otwarcie Przeglądarki skupień z graficzną reprezentacją skupień, zmiennych i poziomów istotności. Więcej informacji można znaleźć w temacie “Przeglądarka skupień — Zakładka modelu” na stronie 250.

Jeśli chcesz przedstawić skupienia w formie siatki, możesz wyświetlić wyniki sieci Kohonena, wykreślając zmienne *$KX-* i *$KY-* za pomocą węzła wykresu. (Aby uniknąć nakładania się rekordów wszystkich jednostek na wykresie, należy w węźle wykresu wybrać opcje **Pobudzenie X** i **Pobudzenie Y**). Na wykres można także nałożyć zmienną symboliczną, aby zbadać, w jaki sposób sieć Kohonena pogrupowała dane.

Inną techniką analizy sieci Kohonena jest indukcja reguł mająca na celu ujawnienie cech wyróżniających skupienia odnalezione przez sieć. Więcej informacji można znaleźć w temacie “Węzeł C5.0” na stronie 106.

Ogólne informacje na temat korzystania z przeglądarki modelu zawiera sekcja “Przeglądanie modeli użytkowych” na stronie 42

Podsumowanie modelu Kohonena

Karta Podsumowanie dla modelu użytkowego Kohonena zawiera informacje o architekturze lub topologii sieci. Długość i szerokość dwuwymiarowej mapy elementów Kohonena (warstwa wyjściowa) są przedstawione jako **$KX-** *model_name* oraz **$KY-** *model_name*. Dla warstwy wejściowej i wynikowej przedstawiona jest liczba jednostek w poszczególnych warstwach.

Węzeł Metoda k-średnich

Węzeł Metoda k-średnich oferuje metodę **analizy skupień**. Może ona posłużyć do skupiania zbioru danych w osobne grupy, jeśli nie wiemy z góry, co to są za grupy. W przeciwieństwie do większości metod uczenia w produkcie IBM SPSS Modeler w modelach K-średnich *nie* są stosowane zmienne przewidywane. Sposób uczenia bez zmiennej przewidywanej jest nazywany **uczeniem nienadzorowanym**. Zamiast prób przewidzenia danych wynikowych K-średnie próbują ujawnić wzorce w zestawie zmiennych wejściowych. Rekordy są grupowane w taki sposób, aby rekordy w ramach grupy lub skupienia były do siebie podobne, zaś rekordy z różnych grup były do siebie niepodobne.

Metoda K-średnich polega na definiowaniu zestawu początkowych centrów skupień wyliczanych na podstawie danych. Następnie każdy rekord jest przypisywany do skupienia, do którego jest on najbardziej podobny, w oparciu o wartości zmiennych wejściowych rekordu. Po przypisaniu wszystkich obserwacji centra skupień są aktualizowane tak, aby odzwierciedlały nowy zestaw rekordów przypisanych do każdego skupienia. Rekordy są następnie sprawdzane ponownie pod kątem tego, czy powinny one zostać przypisane do innego skupienia. Proces przypisywania rekordów/iteracji skupień jest kontynuowany aż do osiągnięcia maksymalnej liczby iteracji lub do chwili, gdy zmiana między daną a następną iteracją będzie mniejsza od zadanego progu.

Uwaga: Model wynikowy zależy do pewnego stopnia od kolejności danych uczących. Zmiana kolejności danych i ponowna budowa modelu może prowadzić do utworzenia innego końcowego modelu skupień.

Wymagania. Do uczenia modelu K-średnich wymagana jest jedna lub więcej zmiennych z rolą ustawioną na *Dane wejściowe*. Zmienne z rolą ustawioną na wartość *Wynik*, *Łącznie* lub *Brak* są ignorowane.

Mocne strony. Do zbudowania modelu K-średnich nie są wymagane dane o przynależności do grupy. Model K-średnich jest często najszybszą metodą skupiania w przypadku dużych zbiorów danych.

Opcje modelu węzła K-średnie

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Określona liczba grup. Określ liczbę skupień do wygenerowania. Domyślną wartością jest 5.

Utwórz zmienną odległości. Jeśli ta opcja zostanie wybrana, model użytkowy będzie obejmował zmienną zawierającą odległość każdego rekordu od środka przypisanego do niego skupienia.

Etykieta grupy. Określ format dla wartości w generowanej zmiennej przynależności do skupień. Informacja o przynależności do skupień może być dostępna jako **Łańcuch** z podanym **Przedrostkiem etykiety** (na przykład "Grupa 1", "Grupa 2" itd.) lub jako **Liczba**.

Uwaga: Jeśli wymagane jest uwzględnienie zmiennych nominalnych w modelu, ale pojawiły się problemy z pamięcią podczas budowania modelu lub budowanie modelu trwa zbyt długo, należy rozważyć zarejestrowanie dużych zmiennych nominalnych, aby zmniejszyć liczbę wartości, lub należy rozważyć użycie innej zmiennej z mniejszą liczbą wartości zamiast zmiennej typu dużego zestawu. Na przykład, jeśli pojawiły się problemy ze zmienną *product_id* zawierającą wartości dla pojedynczych produktów, można rozważyć usunięcie jej z modelu i dodanie zamiast niej mniej szczegółowej zmiennej *product_category*.

Optymalizuj ze względu na. Można tutaj wybrać opcje służące do optymalizacji wydajności podczas tworzenia modelu.

- Zaznaczenie opcji **Szybkość** uniemożliwia przenoszenie danych na dysk, co przekłada się na większą wydajność.
- Zaznaczenie opcji **Pamięć** sprawia, że algorytm przenosi dane na dysk kosztem szybkości przetwarzania. Ta opcja jest wybrana domyślnie.

  Uwaga: Podczas uruchamiania trybu analizy rozproszonej to ustawienie może być przesłonięte przez opcje administratora określone w pliku options.cfg.

Zaawansowane opcje węzła K-średnie

Użytkownikom, którzy posiadają szczegółową wiedzę na temat grupowania *k*-średnich opcje zaawansowane umożliwiają precyzyjne dostosowywanie procesu uczenia. W celu uzyskania dostępu do opcji zaawansowanych należy ustawić opcję Tryb na wartość **Zaawansowany** na karcie Zaawansowany.

Zatrzymywanie uczenia. Określ kryterium zatrzymywania, które będzie stosowane podczas uczenia modelu. **Default** kryterium zatrzymywania to 20 iteracji lub zmiana < 0,000001 — w zależności od tego, co nastąpi jako pierwsze. Aby określić własne kryterium zatrzymywania, wybierz opcję **Użytkownika**.

- **Maksymalna liczba Iteracji.** Ta opcja umożliwia zatrzymanie uczenia modelu po upływie podanej liczbie iteracji.
- **Zmiana tolerancji.** Ta opcja umożliwia zatrzymanie uczenia modelu, gdy największa zmiana w centrum skupienia dla iteracji jest mniejsza niż podany poziom.

Kodowanie wartości dla zmiennych jakościowych. Podaj wartość w zakresie od 0 do 1,0, która będzie używana do rejestrowania zmiennych nominalnych jako grup zmiennych numerycznych. Wartością domyślną jest pierwiastek kwadratowy z 0,5 (około 0,707107), który zapewnia prawidłowe określanie wag zarejestrowanych zmiennych flagi. Wartości bliższe 1,0 będą określać dla zmiennych nominalnych wagi wyższe niż dla zmiennych numerycznych.

Wartościowa informacja z modelu K-średnie

Model użytkowy K-średnie zawiera wszystkie informacje przechwycone przez model skupień, a także informacje o danych uczących i procesie estymacji.

Po uruchomieniu strumienia zawierającego węzeł modelowania K-średnie węzeł dodaje dwie nowe zmienne zawierające przynależność do skupień i odległość od przypisanego środka skupienia dla tego rekordu. Nazwy nowych zmiennych są pochodnymi nazwy modelu z przedrostkiem *$KM-* w przypadku przynależności do skupień i przedrostkiem *$KMD-* w przypadku odległości od środka skupienia. Przykładowo, jeśli nazwa modelu to *Kmeans*, nazwy nowych zmiennych będą następujące: *$KM-Kmeans* i *$KMD-Kmeans*.

Jedną ze skutecznych technik analizy modelu K-średnich jest indukcja reguł mająca na celu ujawnienie cech wyróżniających skupienia odnalezione przez ten model. Więcej informacji można znaleźć w temacie “Węzeł C5.0” na stronie 106. Można także kliknąć zakładkę Model w przeglądarce modeli użytkowych, aby wyświetlić Przeglądarkę skupień z graficzną reprezentacją skupień, zmiennych i poziomów istotności. Więcej informacji można znaleźć w temacie “Przeglądarka skupień — Zakładka modelu” na stronie 250.

Ogólne informacje na temat korzystania z przeglądarki modelu zawiera sekcja “Przeglądanie modeli użytkowych” na stronie 42

Podsumowanie modelu K-średnie

Karta Podsumowanie dla modelu użytkowego K-średnich zawiera informacje o danych uczących, procesie szacowania, a także o skupieniach definiowanych przez model. Widoczna jest liczba skupień, a także przebieg iteracji. Jeśli wykonano węzeł analizy dołączony do tego węzła modelowania, informacje z tej analizy również będą wyświetlane w tej sekcji.

Węzeł Dwustopniowa

Węzeł Dwustopniowa oferuje jedną z metod **analizy skupień**. Może ona posłużyć do skupiania zbioru danych w osobne grupy, jeśli nie wiemy z góry, co to są za grupy. Podobnie jak w przypadku węzłów Kohonen i K-średnie, modele Dwustopniowa *nie* mają zmiennej przewidywanej. Węzeł Dwustopniowa nie próbuje przewidzieć wyniku, lecz ujawnia istniejące wzorce w zbiorze zmiennych wejściowych. Rekordy są grupowane w taki sposób, aby rekordy w ramach grupy lub skupienia były do siebie podobne, zaś rekordy z różnych grup były do siebie niepodobne.

Węzeł Dwustopniowa realizuje dwustopniową metodę analizy skupień. Pierwszy krok stanowi pojedynczy przebieg przez dane polegający na kompresji surowych danych wejściowych w łatwy w zarządzaniu zestaw podgrup. Drugi krok korzysta z hierarchicznej metody grupowania w celu progresywnego scalania podgrup w coraz większe grupy, bez konieczności wykonywania jeszcze jednego przejścia przez dane. Zaletą grupowania hierarchicznego jest brak konieczności wybierania liczby skupień przed rozpoczęciem całego procesu. Wiele metod hierarchicznej analizy skupień rozpoczyna od pojedynczych rekordów traktowanych jako skupienia początkowe, a następnie rekursywnie je scala, tworząc coraz większe grupy. Choć takie strategie często nie sprawdzają się przy pracy na dużych ilościach danych, wstępne grupowanie stosowane w węźle Dwustopniowa zapewnia dużą szybkość hierarchicznej analizy skupień nawet w przypadku obszernych zbiorów danych.

Uwaga: Model wynikowy zależy do pewnego stopnia od kolejności danych uczących. Zmiana kolejności danych i ponowna budowa modelu może prowadzić do utworzenia innego końcowego modelu skupień.

Wymagania. Do uczenia modelu Dwustopniowa wymagana jest jedna lub więcej zmiennych z rolą ustawioną na *Dane wejściowe*. Zmienne z rolą ustawioną na wartość *Zmienna przewidywana*, *Łącznie* lub *Brak* są ignorowane. Algorytm dwustopniowej analizy skupień nie obsługuje braków danych. Rekordy z wartościami pustymi w jakichkolwiek zmiennych wejściowych będą ignorowane podczas budowania modelu.

Mocne strony. Dwustopniowa analiza skupień obsługuje różne typy zmiennych jednocześnie i wydajnie radzi sobie z obszernymi zbiorami danych. Umożliwia również przetestowanie kilku wariantów grupowania i wybranie najlepszego z nich, zatem użytkownik nie musi z góry wiedzieć, ilu skupień zażądać. Węzeł Dwustopniowa można skonfigurować w taki sposób, aby automatycznie wykluczał **wartości odstające** lub skrajnie nietypowe obserwacje, które mogłyby zanieczyścić wyniki.

Ważne:

W produkcie IBM SPSS Modeler dostępne są dwie różne wersje węzła Dwustopniowa:

- **Dwustopniowa** to tradycyjny węzeł działający na serwerze IBM SPSS Modeler Server.
- **Dwustopniowa - AS** to węzeł działający tylko po nawiązaniu połączenia z serwerem IBM SPSS Analytic Server.

Opcje modelu węzła Dwustopniowe grupowanie

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Standaryzuj zmienne numeryczne. Domyślnie grupowanie dwustopniowe standaryzuje wszystkie numeryczne zmienne wejściowe do tej samej skali, ze średnią równą 0 i wariancją równą 1. Aby zachować oryginalne skalowanie zmiennych numerycznych, należy usunąć zaznaczenie tej opcji. Nie ma to wpływu na zmienne symboliczne.

Wyklucz wartości odstające. W przypadku zaznaczenia tej opcji rekordy, które wydają się nie pasować do konkretnej grupy, będą automatycznie wykluczane z analizy. Zapobiega to zniekształcaniu wyników w takich sytuacjach.

Wykrywanie wartości odstających odbywa się na etapie grupowania wstępnego. Jeśli ta opcja jest zaznaczona, podgrupy o niewielkiej liczbie rekordów w porównaniu z innymi podgrupami są uznawane za potencjalne wartości odstające, a drzewo podgrup jest budowane ponownie z wykluczeniem tych rekordów. O wielkości, poniżej której podgrupy są uznawane za zawierające potencjalne wartości odstające, decyduje wartość opcji **Procent**. Niektóre z rekordów zawierających te potencjalne wartości odstające można dodać do budowanych ponownie podgrup, o ile wykazują one wystarczające podobieństwo do któregokolwiek z nowych profili podgrup. Pozostałe z potencjalnych wartości odstających, których nie można połączyć, są uznawane za wartości odstające i dodawane do grupy „szumów”, a następnie wykluczane z hierarchicznej analizy skupień.

W przypadku *oceniania* danych za pomocą modelu Dwustopniowa korzystającego z obsługi wartości odstających nowe obserwacje powyżej pewnego progu odległości (w oparciu o logarytm wiarygodności) od najbliższego znaczącego skupienia są uznawane za wartości odstające i przypisywane do grupy „szum” z nazwą -1.

Etykieta grupy. Określ format dla generowanej zmiennej przynależności do skupień. Informacja o przynależności do skupień może być dostępna jako **Łańcuch** z podanym **Przedrostkiem etykiety** (na przykład "Cluster 1", "Cluster 2" itd.) lub jako **Liczba**.

Automatycznie wylicz liczbę grup. Dwustopniowa analiza skupień pozwala w sposób nagły analizować duże liczby rozwiązań dla grup w celu wybrania optymalnej liczby grup dla danych uczących. Należy podać zakres rozwiązań do wypróbowania, ustawiając wartości **Maksimum** i **Minimum** liczby grup. Analiza dwustopniowa określa optymalną liczbę grup w dwuetapowym procesie. W pierwszym etapie wybierana jest górna granica liczby grup w modelu w oparciu o zmianę w Bayesowskim kryterium informacyjnym (BIC) w miarę dodawania skupień. W drugim etapie wyznaczana jest zmiana w minimalnej odległości między grupami dla wszystkich modeli o liczbie grup mniejszej niż w rozwiązaniu minimum-BIC. Największa zmiana w odległości służy do identyfikowania końcowego modelu skupień.

Określ liczbę grup. Jeśli wiesz, ile grup należy uwzględnić w modelu, wybierz tę opcję i wprowadź liczbę grup.

Miara odległości. Wybrana tutaj opcja określa sposób wyliczenia podobieństwa dwóch skupień.

- **Logarytm wiarygodności.** Miara wiarygodności stosuje do zmiennych rozkład prawdopodobieństwa. Zakłada się, że zmienne ilościowe mają rozkład normalny, natomiast kategorialne rozkład wielomianowy. Zakłada się, że wszystkie zmienne są niezależne.
- **Euklidesowa.** Odległość euklidesowa jest odległością „w linii prostej” pomiędzy dwoma skupieniami. Można jej użyć tylko wówczas, gdy wszystkie zmienne są zmiennymi ilościowymi.

Kryterium grupowania. Wybrana tutaj opcja określa sposób ustalenia liczby skupień przez algorytm automatycznego grupowania. Dostępne opcje to Bayesowskie Kryterium Informacyjne (BIC) i Kryterium informacyjne Akaike (AIC).

Wartościowe informacje z modelu dwustopniowego skupienia

Model dwustopniowego skupienia zawiera wszystkie informacje zarejestrowane w modelu skupień, a także informacje o danych uczących i procesie estymacji.

Po uruchomieniu strumienia zawierającego model użytkowy dwustopniowego skupienia węzeł dodaje nową zmienną zawierającą informacje o przynależności do skupień dotyczące tego rekordu. Nazwa nowej zmiennej tworzona jest na podstawie nazwy modelu z przedrostkiem *$T-*. Na przykład, jeśli model nosi nazwę *TwoStep*, nowa zmienna będzie miała nazwę *$T-TwoStep*.

Jedną ze skutecznych technik analizy modelu dwustopniowego skupienia jest indukcja reguł mająca na celu ujawnienie cech wyróżniających skupienia odnalezione przez ten model. Więcej informacji można znaleźć w temacie “Węzeł C5.0” na stronie 106. Można także kliknąć zakładkę Model w przeglądarce modeli użytkowych, aby wyświetlić Przeglądarkę skupień z graficzną reprezentacją skupień, zmiennych i poziomów istotności. Więcej informacji można znaleźć w temacie “Przeglądarka skupień — Zakładka modelu” na stronie 250.

Ogólne informacje na temat korzystania z przeglądarki modelu zawiera sekcja “Przeglądanie modeli użytkowych” na stronie 42

Podsumowanie modelu Dwustopniowa

Karta Podsumowanie dla modelu użytkowego dwustopniowego grupowania zawiera liczbę znalezionych grup wraz z informacjami na temat danych uczących, procesu oszacowania i ustawień budowy.

Więcej informacji można znaleźć w temacie “Przeglądanie modeli użytkowych” na stronie 42.

Węzeł Dwustopniowa-AS

W produkcie IBM SPSS Modeler dostępne są dwie różne wersje węzła Dwustopniowa:

- **Dwustopniowa** to tradycyjny węzeł działający na serwerze IBM SPSS Modeler Server.
- **Dwustopniowa - AS** to węzeł działający tylko po nawiązaniu połączenia z serwerem IBM SPSS Analytic Server.

Dwustopniowa-AS analiza skupień

Procedura Dwustopniowa analiza skupień jest narzędziem eksploracyjnym służącym do ujawniania występowania w zbiorze danych naturalnych zgrupowań (lub skupień), które nie są widoczne w inny sposób. Algorytm zastosowany w tej procedurze posiada kilka wyjątkowych cech, które odróżniają go od tradycyjnych metod grupowania:

- **Obsługa zmiennych jakościowych i ilościowych.** Przy założeniu niezależności zmiennych, do zmiennych jakościowych i ilościowych można zastosować połączony rozkład wielomianowo-normalny.
- **Automatyczny wybór liczby skupień.** Przez porównanie wartości kryterium wyboru modelu dla różnych rozwiązań grupowania procedura może automatycznie określić optymalną liczbę skupień.
- **Skalowalność.** Przez utworzenie predyktorów (CF) podsumowującego rekordy algorytm Dwustopniowa umożliwia analizę dużych plików danych.

Na przykład producenci i sprzedawcy artykułów konsumpcyjnych regularnie stosują techniki grupowania do danych opisujących nawyki nabywcze, płeć, wiek, poziom przychodów i inne atrybuty swoich klientów. Przedsiębiorstwa te dostosowują swoje strategie marketingowe i produktowe do każdej grupy konsumentów celem zwiększenia sprzedaży i pozyskiwania lojalności klientów wobec danej marki produktów.

Karta Zmienne

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Użyj wstępnie zdefiniowanych ról. Wybrane są wszystkie zmienne o roli zmiennych wejściowych.

Użyj niestandardowych przypisań. Można tutaj dodawać i usuwać zmienne niezależnie od ich roli. Można wybierać pola o dowolnej roli i przenosić je na listę **Predyktory (Wejścia)** lub usuwać z tej listy.

Podstawowe

Liczba skupień

Dobierz automatycznie
Procedura określa najlepszą liczbę skupień w określonym przedziale. **Minimum** musi być większe od 1. Jest to ustawienie domyślne.

Ustalona liczba skupień
Procedura generuje określoną liczbę skupień. **Liczba** musi być większa od 1.

Kryterium grupowania

Wybrana tutaj opcja określa sposób ustalenia liczby skupień przez algorytm automatycznego grupowania.

Bayesowskie kryterium informacyjne (BIC)
Miara wybierania i porównywania modeli mieszanych tworzonych na podstawie -2 logarytmu wiarygodności. Mniejsze wartości oznaczają lepszy model. Wartość BIC „karze" także modele przeparametryzowane (na przykład złożone modele z dużą liczbą danych wejściowych), jednak silniej niż miara AIC.

Kryterium informacyjne Akaike (AIC)
Miara wybierania i porównywania modeli mieszanych tworzonych na podstawie -2 logarytmu wiarygodności. Mniejsze wartości oznaczają lepszy model. Wartość AIC „karze" modele przeparametryzowane (na przykład złożone modele z dużą liczbą danych wejściowych).

Metoda automatycznego grupowania

W razie wybrania opcji **Dobierz automatycznie** należy wybrać jedną z następujących metod grupowania, aby automatycznie ustalić liczbę skupień:

Użyj ustawienia kryterium grupowania
Zbieżność kryterium informacyjnego jest to iloraz liczby kryteriów informacyjnych odpowiadających dwóm bieżącym rozwiązaniom dla skupień do liczby kryteriów odpowiadających pierwszemu rozwiązaniu dla skupień. Używane jest kryterium wybrane w grupie Kryterium grupowania

Skok odległości
Skok odległości jest to iloraz odległości odpowiadających dwóm kolejnym rozwiązaniom dla skupień.

Maksimum
Łączy wyniki z metody zbieżności kryterium informacyjnego i metody skoku odległości, generując liczbę skupień odpowiadającą drugiemu skokowi.

Minimum
Łączy wyniki z metody zbieżności kryterium informacyjnego i metody skoku odległości, generując liczbę skupień odpowiadającą pierwszemu skokowi.

Metoda ważności predyktorów

Metoda ważności predyktorów określa, jak ważne są predyktory (zmienne) w rozwiązaniu dla skupień. Wyniki zawierają informacje o ogólnej ważności predyktorów oraz ważności każdego predyktora w każdym skupieniu. Predyktory o ważności mniejszej niż minimalna są wykluczane.

Użyj ustawienia kryterium grupowania.
Jest to metoda domyślna, bazująca na kryterium wybranym w grupie Kryterium grupowania.

Wielkość wpływu
Ważność predyktorów oparta jest na wielkości wpływu (efektu), a nie wartościach istotności.

Kryterium drzewa predyktorów

Te ustawienia określają sposób budowania drzewa predyktorów skupień. Przez utworzenie drzewa predyktorów skupień i podsumowanie rekordów algorytm Dwustopniowa jest w stanie analizować bardzo duże pliki danych. Innymi słowy, algorytm Dwustopniowa używa drzewa predyktorów skupień do budowania skupień, co pozwala mu na przetwarzanie dużej liczby obserwacji.

Miara odległości

Wybrana tutaj opcja określa sposób wyliczenia podobieństwa dwóch skupień.

Logarytm wiarygodności

Miara wiarygodności stosuje do zmiennych rozkład prawdopodobieństwa zmiennych. Zakłada się, że zmienne ilościowe mają rozkład normalny, natomiast kategorialne rozkład wielomianowy. Zakłada się, że wszystkie zmienne są niezależne.

Euklidesowa

Odległość euklidesowa jest odległością „w linii prostej" pomiędzy dwoma skupieniami. Miara euklidesowa kwadratowa i metoda Warda stosowane są do obliczania podobieństwa między skupieniami. Można jej użyć tylko wówczas, gdy wszystkie zmienne są zmiennymi ilościowymi.

Skupienia odstające

Uwzględnij grupy odstające

Powoduje tworzenie skupień obserwacji odstających od typowych skupień. Gdy ta opcja nie jest wybrana, wszystkie obserwacje umieszczane są w typowych skupieniach.

Liczba obserwacji w liściu drzewa predyktorów jest mniejsza niż.

Jeśli liczba obserwacji w liściu drzewa predyktorów jest mniejsza od określonej wartości, to liść uznawany jest za odstający. Wartość musi być liczbą całkowitą większą od 1. Większe wartości zwiększają prawdopodobieństwo tworzenia skupień odstających.

Górny procent wartości odstających.

Podczas budowania modelu skupień wartości odstające są rangowane według siły odstawania. Siła odstawania wymagana do tego, by obserwacja znalazła się w górnych dziesięciu procentach wartości odstających, jest progiem, powyżej którego obserwacje uznawane są za odstające. Wyższe wartości powodują, że więcej obserwacji zostanie uznanych za odstające. Wartość musi mieścić się w przedziale od 1 do 100.

Ustawienia dodatkowe

Początkowy próg zmiany odległości

Początkowy próg używany do rozwijania drzewa predyktorów. Jeśli wstawienie liścia do istniejącego liścia drzewa powodowałoby takie zagęszczenie obserwacji, że odległość między nimi byłaby niższa od progu, liść nie jest dzielony. Jeśli zagęszczenie jest wyższe od progu, liść jest dzielony.

Maksimum gałęzi z liściastymi węzłami

Maksymalna liczba węzłów podrzędnych, które węzeł liścia może posiadać.

Maksimum gałęzi z węzłami bez liści

Maksymalna liczba węzłów podrzędnych, jaką może mieć węzeł niebędący liściem.

Maksymalna głębokość drzewa

Maksymalna liczba poziomów, jaką może mieć drzewo skupień.

Korekta wagi na poziomie pomiaru

Zmniejsza wpływ zmiennych jakościowych poprzez zwiększenie wagi zmiennych ilościowych. Ta wartość jest mianownikiem używanym przy redukcji wagi zmiennych jakościowych. Na przykład domyślna wartość 6 powoduje, że waga zmiennych jakościowych jest mnożona przez 1/6.

Przydział pamięci

Maksymalna ilość pamięci w megabajtach (MB), jaką wykorzystuje algorytm grupowania. Jeśli procedura przekroczy taką wielkość maksymalną, dane niemieszczące się w pamięci będą przechowywane na dysku.

Opóźniony podział

Opóźnia ponowne budowanie drzewa predyktorów skupień. Algorytm grupowania wielokrotnie od początku buduje drzewa predyktorów, oceniając nowe obserwacje. Ta opcja może przyczynić się do wzrostu wydajności poprzez ograniczenie liczby operacji odbudowy drzewa.

Standaryzowanie

Algorytm grupowania operuje na zestandaryzowanych zmiennych ilościowych. Domyślnie wszystkie zmienne ilościowe są standaryzowane. Aby zaoszczędzić czas i zasoby obliczeniowe, można przenieść zmienne ilościowe, które są już zestandaryzowane, na listę **Nie standaryzuj**.

Dobór predyktorów

Na ekranie Dobór predyktorów można zdefiniować reguły wykluczania zmiennych. Można na przykład wykluczyć zmienne z licznymi brakami danych.

Reguły wykluczania zmiennych

Procent braków danych większy niż.

Zmienne z odsetkiem braków danych większym od podanej wartości będą wykluczone z analizy. Wartość musi być liczbą większą od zera i mniejszą od 100.

Liczba kategorii dla zmiennych jakościowych większa od.

Zmienne jakościowe z liczbą kategorii większą od podanej będą wykluczone z analizy. Wartość ta musi być dodatnią liczbą całkowitą większą od 1.

Zmienne z tendencją do przyjmowania pojedynczej wartości.

Współczynnik zmienności zmiennych ciągłych mniejszy niż.

Zmienne ilościowe ze współczynnikiem zmienności mniejszym od podanej wartości będą wykluczone z analizy. Współczynnik zmienności to iloraz odchylenia standardowego i średniej. Niższe wartości wskazują zwykle na mniejszą zmienność wartości. Wartość musi mieścić się w przedziale od 0 do 1.

Procent obserwacji w pojedynczej kategorii większy od.

Zmienne jakościowe z odsetkiem obserwacji w jednej kategorii większej od podanej wartości będą wykluczone z analizy. Wprowadzana wartość musi być liczbą większą od 0 i mniejszą od 100.

Adaptacyjny wybór predyktora

Ta opcja powoduje wykonanie dodatkowego przebiegu w celu znalezienia i usunięcia najmniej ważnych zmiennych.

Wyniki modelu

Podsumowanie tworzenia modelu

Specyfikacja modelu

Podsumowanie specyfikacji modelu, liczby skupień w ostatecznym modelu i zmienne wejściowe uwzględnione w ostatecznym modelu.

Podsumowanie rekordów

Liczba i odsetek rekordów (obserwacji) uwzględnionych w modelu i wykluczonych z modelu.

Wykluczone dane wejściowe

Dla każdej zmiennej wykluczonej z modelu podana jest przyczyna wykluczenia.

Ewaluacja

Jakość modelu

Tabela dobroci i ważności poszczególnych skupień oraz ogólnej dobroci dopasowania modelu.

Wykres słupkowy ważności predyktorów

Wykres słupkowy ważności predyktorów (zmiennych) obejmujący wszystkie skupienia. Predyktory (zmienne) z dłuższymi słupkami są ważniejsze od tych z krótszymi słupkami. Predyktory są także posortowane od największej do najmniejszej ważności (słupek na górze jest najważniejszy).

Chmura słów ważności predyktora

Chmura słów ważności predyktorów (zmiennych) obejmująca wszystkie skupienia. Predyktory (zmienne) z większym tekstem są ważniejsze od tych z mniejszym tekstem.

Skupienia odstające

Te opcje są wyłączone, jeśli wybrano opcję nieuwzględniania wartości odstających.

Interaktywna tabela i wykres

Tabela i wykres siły odstawania i względnego podobieństwa skupień odstających do zwykłych skupień. Wybieranie różnych wierszy w tabeli powoduje wyświetlanie na wykresie informacji o różnych skupieniach odstających.

Tabela przestawna

Tabela siły odstawania i względnego podobieństwa skupień odstających do zwykłych skupień. Ta tabela zawiera te same informacje, co prezentacja interaktywna. Tabela ta obsługuje wszystkie standardowe funkcje przestawiania i edytowania tabel.

Maksymalna liczba

Maksymalna liczba skupień odstających, która może być ujęta w wynikach. Jeśli liczba skupień odstających przekracza dwadzieścia, to zostanie wyświetlona tabela przestawna.

Interpretacja

Profile ważności predyktorów między grupami

Interaktywna tabela i wykres.

Tabela i wykresy ważności predyktorów oraz środków skupień dla każdej zmiennej wejściowej używanej w rozwiązaniu dla skupień. Wybieranie różnych wierszy w tabeli powoduje wyświetlanie różnych wykresów. W przypadku zmiennych jakościowych wyświetlany będzie wykres słupkowy. W przypadku zmiennych ilościowych wyświetlany będzie wykres średnich i odchyleń standardowych.

Tabela przestawna.

Tabela ważności predyktorów oraz środków skupień dla każdej zmiennej wejściowej. Ta tabela zawiera te same informacje, co prezentacja interaktywna. Tabela ta obsługuje wszystkie standardowe funkcje przestawiania i edytowania tabel.

Ważność predyktorów w grupie

Dla każdego skupienia podany jest środek skupienia i ważność poszczególnych zmiennych wejściowych. Dla każdego skupienia istnieje osobna tabela.

Odległości grup

Wykres panelowy odległości między skupieniami. Dla każdego skupienia istnieje osobny panel.

Etykieta grupy

Tekst Etykieta każdego skupienia jest wartością określoną jako **Przedrostek**, po której następuje numer kolejny.

Liczba Etykieta każdego skupienia jest numerem kolejnym.

Opcje modelu

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Modele użytkowe Dwustopniowa-AS

Model użytkowy modelu Dwustopniowa-AS przedstawia szczegóły dotyczące modelu na karcie Model w oknie wynikowym. Więcej informacji na temat korzystania z okna raportów wynikowych można znaleźć w temacie „Praca z oknem wyników" w Podręczniku użytkownika programu Modeler (ModelerUsersGuide.pdf).

Model Dwustopniowa-AS zawiera wszystkie informacje zarejestrowane w modelu skupień, a także informacje o danych uczących i procesie estymacji.

Po uruchomieniu strumienia zawierającego model użytkowy Dwustopniowa-AS węzeł dodaje nową zmienną zawierającą informacje o przynależności do skupień dotyczące tego rekordu. Nazwa nowej zmiennej tworzona jest na podstawie nazwy modelu z przedrostkiem ***$AS-***. Na przykład, jeśli model nosi nazwę TwoStep, nowa zmienna będzie miała nazwę ***$AS-TwoStep***.

Jedną ze skutecznych technik analizy modelu Dwustopniowa-AS jest indukcja reguł mająca na celu ujawnienie cech wyróżniających skupienia odnalezione przez ten model. Więcej informacji można znaleźć w temacie “Węzeł C5.0” na stronie 106.

Ogólne informacje na temat korzystania z przeglądarki modelu zawiera sekcja “Przeglądanie modeli użytkowych” na stronie 42

Ustawienia modelu użytkowego Dwustopniowa-AS

Karta Ustawienia zawiera dodatkowe opcje dla modelu użytkowego Dwustopniowa-AS.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę, wykorzystując natywny kod SQL** Jeśli ta opcja jest wybrana, generowany jest natywny kod SQL w celu oceny modelu w bazie danych.

  Uwaga: Ta opcja może szybciej zwracać wyniki, ale rozmiar i złożoność natywnego kodu SQL wzrastają wraz ze wzrostem złożoności modelu.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Węzeł K-średnie-AS

K-średnie to jeden z najpowszechniej używanych algorytmów grupowania. Grupuje on punkty danych w określoną z góry liczbę grup.[1] Węzeł K-średnie-AS w programie SPSS Modeler jest zaimplementowany w środowisku Spark.

Aby uzyskać szczegółowe informacje na temat algorytmów K-średnich, patrz https://spark.apache.org/docs/2.2.0/ml-clustering.html.

Należy zwrócić uwagę, że węzeł K-średnie-AS automatycznie wykonuje kodowanie one-hot (kodowanie z gorącą jedynką) dla zmiennych kategorialnych.

[1] "Clustering." *Apache Spark*. MLlib: Main Guide. WWW. 3 października 2017 r.

Węzeł K-średnie-AS — Zmienne

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Użyj wstępnie zdefiniowanych ról. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej z wcześniejszego węzła Typ. Ta opcja jest wybrana domyślnie.

Użyj niestandardowych przypisań. Aby ręcznie przypisać zmienne wejściowe, wybierz tę opcję, a następnie zmienną lub zmienne wejściowe. Działanie tej opcji jest podobne, jak ustawienie roli zmiennej na **Zmienna wejściowa** w węźle Typy.

Węzeł K-średnie-AS — Opcje budowania

Na karcie Opcje budowania można określić opcje budowania dla węzła K-średnie-AS, w tym zwykłe opcje budowania modelu, opcje inicjowania środków grup i opcje zaawansowane dotyczące iteracji obliczeń i wartości startowej generatora liczb losowych. Aby uzyskać więcej informacji, patrz JavaDoc for K-Means on SparkML.[1]

Regular

Nazwa modelu. Nazwa zmiennej generowanej po ocenie z przypisaniem do określonej grupy. Wybierz opcję **Automatycznie** (ustawienie domyślne) lub wybierz opcję **Użytkownika** i wpisz nazwę.

Liczba grup. Określ liczbę skupień do wygenerowania. Wartość domyślna to **5**, a wartość minimalna to **2**.

Inicjowanie

Tryb inicjowania. Określ metodę inicjowania środków grup. Ustawienie domyślne to **K-średnie||**. Aby uzyskać szczegółowe informacje o obu metodach, patrz Scalable K-Means++. [2]

Kroki inicjowania. Jeśli wybrany jest tryb inicjowania **K-średnie||**, określ liczbę kroków inicjowania. Wartość domyślna to **2**.

Zaawansowane

Ustawienia zaawansowane. Wybierz tę opcję, jeśli chcesz ustawić poniższe opcje zaawansowane.

Maks. liczba iteracji. Określ, maksymalną liczbę iteracji, jaka ma być wykonywana podczas poszukiwania środków grup. Wartość domyślna to **20**.

Tolerancja. Określ tolerancję zbieżności dla algorytmów iteracyjnych. Wartość domyślna to **1.0E-4**.

Ustaw wartość startową generatora liczb losowych. Wybierz tę opcję i kliknij przycisk **Utwórz**, aby wygenerować wartość początkową dla generatora liczb losowych.

Wyświetl

Wyświetl wykres. Wybierz tę opcję, jeśli wyniki mają zawierać wykres.

W poniższej tabeli przedstawiono relację między ustawieniami w oknie dialogowym węzła K-średnie-AS w programie SPSS Modeler a parametrami algorytmu K-średnie w środowisku Spark.

Tabela 13. Właściwości węzła odwzorowane na parametry Spark

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr SparkML algorytmu K-średnie
Zmienne wejściowe	właściwości	
Liczba skupień	clustersNum	k

Tabela 13. Właściwości węzła odwzorowane na parametry Spark (kontynuacja)

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr SparkML algorytmu K-średnie
Tryb inicjowania	initMode	initMode
Kroki inicjowania	initSteps	initSteps
Maks. liczba iteracji	maxIter	maxIter
Tolerancja	toleration	tol
Losowa wartość początkowa	randomSeed	seed

[1] "Class KMeans." *Apache Spark*. JavaDoc. WWW. 3 października 2017 r.

[2] Bahmani, Moseley, et al. "Scalable K-Means++." 28 września 2012 r. http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf.

Przeglądarka skupień

Modele skupień są zwykle wykorzystywane do znajdowania grup (lub skupień), lub podobnych rekordów bazujących na badanych zmiennych w sytuacjach wysokiego podobieństwa między elementami tej samej grupy oraz niskiego podobieństwa między elementami różnych grup. Wyniki można wykorzystać do identyfikacji powiązań, które nie są widoczne w inny sposób. Na przykład dzięki analizie skupień preferencji klientów, poziomu dochodów oraz nawyków nabywczych, możliwe jest zidentyfikowanie typów klientów, którzy z większym prawdopodobieństwem odpowiedzą na określoną kampanię marketingową.

Istnieją dwa podejścia do interpretowania wyników na ekranie skupień:

- Zbadaj skupienia w celu określenia charakterystyki niepowtarzalnego skupienia. *Czy jedno skupienie zawiera wszystkich pożyczkobiorców o wysokich dochodach? Czy to skupienie zawiera więcej rekordów niż inne skupienia?*
- Zbadaj zmienne w skupieniach w celu określenia sposobu rozłożenia wartości po skupieniach. *Czy wykształcenie danej osoby determinuje przynależność do skupienia? Czy wysoka ocena kredytowa powoduje rozróżnienie w zakresie przynależności do jednego lub drugiego skupienia?*

Korzystając z głównych widoków oraz różnych, połączonych widoków w Przeglądarce skupień można uzyskać wiedzę, która pomoże odpowiedzieć na te pytania.

Poniższe, wartościowe informacje z modeli skupień można wygenerować w aplikacji IBM SPSS Modeler:

- Wartościowa informacja z modelu Sieć Kohonena
- Wartościowa informacja z modelu K-średnie
- Wartościowa informacja z modelu dwustopniowego skupienia

Aby zobaczyć informacje na temat wartościowych informacji z modeli skupień, kliknij prawym przyciskiem myszy na węzeł modelu, aby wybrać opcję **Przeglądaj** z menu kontekstowego (lub **Edytuj** dla węzłów w strumieniu). Alternatywnie, w przypadku używania węzła modelowania Automatyczne skupianie, dwukrotnie kliknij wymagane wartościowe informacje o skupieniu w wartościowych informacjach Automatycznego skupiania. Więcej informacji można znaleźć w temacie “Węzeł Auto Grupowanie” na stronie 76.

Przeglądarka skupień — Zakładka modelu

Zakładka Model dla modeli skupień stanowi graficzne przedstawienie statystyki podsumowującej i rozkładów dla zmiennych między skupieniami; nosi ona nazwę **Przeglądarka skupień**.

Uwaga: zakładka Model jest niedostępna dla modeli wbudowanych w wersjach programu IBM SPSS Modeler starszych niż wersja 13.

Przeglądarka skupień składa się z dwóch paneli, widoku głównego z lewej strony i powiązanego lub dodatkowego widoku z prawej strony. Istnieją dwa główne widoki:

- Podsumowanie modelu (domyślny). Więcej informacji można znaleźć w temacie “Widok podsumowania modelu”.
- Grupy. Więcej informacji można znaleźć w temacie “Widok skupień”.

Istnieją cztery połączone/dodatkowe widoki:

- Ważność predyktora. Więcej informacji można znaleźć w temacie “Widok ważności predyktora skupień” na stronie 253.
- Rozmiary grup (domyślne). Więcej informacji można znaleźć w temacie “Widok rozmiarów skupień” na stronie 253.
- Rozkład komórek. Więcej informacji można znaleźć w temacie “Widok rozkładu komórek” na stronie 253.
- Porównanie skupień. Więcej informacji można znaleźć w temacie “Widok porównania skupień” na stronie 253.

Widok podsumowania modelu

Widok Podsumowanie modelu przedstawia przegląd lub podsumowanie modelu skupień, w uwzględnieniem miary Silhouette spójności i odrębności, która jest zacieniowana w celu wskazania słabych, dostatecznych, lub dobrych wyników. Ten przegląd pozwala na szybkie sprawdzenie, czy jakość jest słaba, kiedy to można zdecydować o powrocie do węzła modelowania w celu korekty ustawień modelu skupień, aby uzyskać lepszy wynik.

Wynik słaby, dostateczny lub dobry bazuje na pracy Kaufmana oraz Rousseeuwa (1990), dotyczącej interpretacji struktur skupień. W widoku podsumowania modelu, dobry wynik równa się danym, które odzwierciedlają ocenę Kaufmana oraz Rousseeuwa jako raczej sensowny lub silny dowód struktury skupienia, dostateczny odzwierciedla ich ocenę słabego dowodu, a słaby odzwierciedla ich ocenę braku istotnego dowodu.

Miara silhouette uśrednia poprzez wszystkie rekordy (B–A)/maks (A,B), gdzie A oznacza odległość rekordu od środka grup, a B oznacza odległość rekordu od najbliższego środka grup, do którego rekord ten nie należy. Współczynnik silhouette o wartości 1 oznacza, że wszystkie obserwacje znajdują się bezpośrednio w centrach ich skupień. Wartość 1 oznacza, że wszystkie obserwacje znajdują się w środkach grup innych grup. Wartość 0 oznacza, że średnio obserwacje znajdują się w równej odległości od centrum ich własnego skupienia i od najbliższego, innego skupienia.

Podsumowanie obejmuje tabelę zawierającą następujące informacje:

- **Algorytm.** Używany algorytm grupowania, na przykład, "TwoStep" (dwustopniowy).
- **Zmienne wejściowe.** Liczba zmiennych, znanych również jako **wejścia** lub **predyktory**.
- **Grupy.** Liczba skupień w rozwiązaniu.

Widok skupień

Widok skupień zawiera siatkę skupień według predyktora, która zawiera nazwy, rozmiary i profile poszczególnych skupień.

Kolumny w siatce zawierają następujące informacje:

- **Grupowanie.** Numery skupień utworzone przez algorytm.
- **Etykieta.** Dowolne etykiety zastosowane do każdego skupienia (pole to jest domyślnie puste). Dwukrotnie kliknij komórkę, aby wprowadzić etykietę opisującą zawartość skupienia; na przykład, „Nabywcy luksusowych samochodów”.
- **Opis.** Dowolny opis zawartości skupienia (pole to jest domyślnie puste). Dwukrotnie kliknij komórkę, aby wprowadzić opis skupienia; na przykład, „wiek ponad 55 l., profesjonaliści, zarabiający powyżej 100 000 USD”.
- **Rozmiar.** Rozmiar każdego skupienia jako wartość procentowa całego przykładowego skupienia. Każda komórka rozmiaru w siatce przedstawia pasek pionowy, który pokazuje procent rozmiaru w ramach skupienia, procent rozmiaru w formacie liczbowym oraz liczebność obserwacji skupień.
- **Właściwości.** Pojedyncze wejścia lub predyktory, posortowane domyślnie według całkowitej istotności. Jeśli jakieś kolumny mają równe rozmiary, są one wyświetlane w kolejności rosnącej wg numerów skupień.

Całkowita ważność właściwości jest oznaczona kolorem cieniowania tła komórki; najistotniejsza właściwość jest najciemniejsza; najmniej istotna właściwość nie jest cieniowana. Przewodnik nad tabelą wskazuje ważność przypisaną do każdego koloru komórki właściwości.

Po najechaniu myszką na komórkę wyświetla się pełna nazwa / etykieta właściwości i wartość istotności dla komórki. Możliwe jest wyświetlenie dalszych informacji, zależnie od widoku i typu zmiennej. W widoku Środki grup uwzględniana jest statystyka komórki oraz jej wartość, na przykład: „Średnia: 4.32”. Dla zmiennych jakościowych komórki wyświetlają nazwę najczęstszej (modalnej) kategorii i jej wartość procentową.

W Widoku skupień można wybrać różne sposoby wyświetlania informacji o skupieniach:

- Transponuj grupy i zmienne Więcej informacji można znaleźć w temacie “Transponowanie skupień i zmiennych”.
- Sortowanie zmiennych. Więcej informacji można znaleźć w temacie “Sortowanie zmiennych”.
- Sortowanie skupień. Więcej informacji można znaleźć w temacie “Sortowanie skupień”.
- Wybór zawartości komórki. Więcej informacji można znaleźć w temacie “Zawartość komórki”.

Transponowanie skupień i zmiennych: Domyślnie skupienia wyświetlają się jako kolumny, a funkcje wyświetlają się jako wiersze. Aby odwrócić ten sposób wyświetlania, kliknij przycisk **Transponuj grupy i zmienne** z lewej strony przycisków **Sortowanie zmiennych według**. Może się to okazać potrzebne w sytuacji, gdy wyświetlonych jest wiele skupień. W wyniku tego zmniejszy się zakres do przewijania w poziomie w celu obejrzenia danych.

Sortowanie zmiennych: Przyciski **Sortowanie zmiennych według** umożliwiają wybór sposobu wyświetlania komórek zmiennych:

- **Całkowita ważność.** Jest to domyślny porządek sortowania. Zmienne są posortowane w kolejności malejącej według całkowitej istotności, a porządek sortowania jest taki sam we wszystkich skupieniach. Jeśli jakieś zmienne mają powiązane wartości istotności, powiązane zmienne są zestawione rosnąco według nazw zmiennych.
- **Istotność wewnątrzgrupowa.** Zmienne są posortowane według ich istotności dla każdego skupienia. Jeśli jakieś zmienne mają powiązane wartości istotności, powiązane zmienne są zestawione rosnąco według nazw zmiennych. Jeśli wybierze się tę opcję, wówczas zwykle zmienia się porządek sortowania w skupieniach.
- **Nazwa.** Zmienne są posortowane według nazwy w kolejności alfabetycznej.
- **Kolejność danych.** Zmienne są posortowane według ich kolejności w zbiorze danych.

Sortowanie skupień: Domyślnie skupienia są posortowane w porządku malejącym według rozmiaru. Przyciski **Sortowanie grup według** umożliwiają posortowanie skupień według nazw w kolejności alfabetycznej, lub jeśli utworzono niepowtarzalne etykiety, alfanumerycznej kolejności etykiet.

Zmienne posiadające taką samą etykietę są posortowane według nazwy skupienia. Jeśli skupienia są posortowane według etykiety i użytkownik dokona edycji etykiety skupienia, porządek sortowanie zostanie automatycznie zaktualizowany.

Zawartość komórki: Przyciski **Komórki** umożliwiają zmianę sposobu wyświetlania zawartości komórki dla zmiennych i zmiennych ewaluacyjnych.

- **Środki grup.** Domyślnie w komórkach wyświetlają się nazwy/etykiety zmiennych oraz tendencja centralna dla każdej kombinacji skupień/zmiennych. Dla zmiennych ciągłych wyświetla się średnia oraz tryb (najczęściej występująca kategoria) z procentem kategorii dla zmiennych jakościowych.
- **Rozkłady bezwzględne.** Pokazuje nazwy/etykiety zmiennych oraz rozkłady bezwzględne zmiennych w ramach każdego skupienia. Dla zmiennych jakościowych, ekran ten pokazuje wykresy słupkowe, na które są nałożone kategorie uporządkowane w kolejności rosnącej wartości danych. Dla zmiennych ilościowych, ekran ten pokazuje gładki wykres gęstości, który używa tych samych punktów końcowych i odstępów dla każdego skupienia.

  Ten stały, czerwony ekran pokazuje rozkład skupień, podczas gdy bledszy ekran przedstawia całkowite dane.
- **Rozkłady względne.** Pokazują przyszłe nazwy/etykiety i rozkłady względne w komórkach. Zasadniczo ekrany te są podobne do tych, które są wyświetlane dla rozkładów bezwzględnych, z tym, że zamiast nich wyświetlane są rozkłady względne.

Ten stały, czerwony ekran wyświetla rozkład skupień, podczas gdy bledszy ekran przedstawia całkowite dane.

- **Widok podstawowy.** Tam, gdzie jest wiele skupień, zobaczenie wszystkich szczegółów może być trudne bez przewijania. Aby zmniejszyć ilość przewijania, należy wybrać ten widok, aby zmienić sposób wyświetlania na bardziej pomniejszoną wersję tabeli.

Widok ważności predyktora skupień

Widok ważności predyktora skupień pokazuje względną ważność każdej zmiennej w ocenie modelu.

Widok rozmiarów skupień

Widok rozmiarów skupień przedstawia wykres kołowy zawierający każde skupienie. W każdym kawałku przedstawiony jest rozmiar procentowy każdego skupienia; najedź myszą na każdy kawałek, aby wyświetlić liczbę w tym kawałku.

Pod wykresem znajduje się tabela zawierająca następujące informacje o rozmiarach:

- Rozmiar najmniejszego skupienia (liczebność i wartość procentowa całości).
- Rozmiar największego skupienia (liczebność i wartość procentowa całości).
- Proporcja rozmiaru największego skupienia do najmniejszego skupienia.

Widok rozkładu komórek

Widok rozkładu komórek przedstawia rozszerzony, bardziej szczegółowy wykres rozkładu danych dla dowolnej komórki zmiennej wybranej w tabeli w panelu głównym Grupy.

Widok porównania skupień

Widok porównania skupień składa się z układu w postaci siatki, ze zmiennymi w wierszach oraz wybranymi skupieniami w kolumnach. Widok ten pomaga lepiej zrozumieć czynniki składające się na skupienia; umożliwia on również przeglądanie różnic między skupieniami nie tylko w porównaniu z całkowitymi danymi, ale również w porównaniu z samymi skupieniami.

Aby wybrać skupienia do wyświetlenia kliknij na górną część kolumny skupienia w panelu głównym Grupy. Użyj opcji Ctrl+kliknięcie lub Shift+kliknięcie, aby zaznaczyć lub odznaczyć więcej niż jedno skupienie do porównania.

Uwaga: do wyświetlenia można wybrać do pięciu skupień.

Grupy są przedstawione w kolejności, w jakiej zostały wybrane, podczas gdy kolejność zmiennych jest określona przez opcję **Sortowanie zmiennych według**. Po wybraniu opcji **Istotność wewnątrzgrupowa**, zmienne są zawsze posortowane według całkowitej istotności.

Wykresy w tle przedstawiają całkowite rozkłady wszystkich zmiennych:

- Zmienne jakościowe są przedstawione jako wykresy punktowe, gdzie rozmiar punktu oznacza najczęstszą/modalną kategorię dla każdego skupienia (według zmiennej).
- Zmienne ilościowe są wyświetlane jako wykresy skrzynkowe, które przedstawiają całkowite mediany i rozstępy ćwiartkowe.

Na widoki w tle nałożone są wykresy skrzynkowe dla wybranych skupień:

- Dla zmiennych ilościowych, znaczniki w postaci kwadratowych punktów oraz linie poziome wskazują rozstęp mediany i rozstęp ćwiartkowy dla każdego skupienia.
- Każdemu skupieniu odpowiada inny kolor, pokazany u góry widoku.

Nawigacja w Przeglądarce skupień

Przeglądarka skupień jest ekranem interaktywnym. Można:

- Wybrać zmienną lub skupienie, aby zobaczyć więcej szczegółów.
- Porównać skupienia w celu wyboru interesujących nas elementów.
- Zmienić ekran.

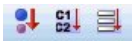
- Zamienić osie wykresu.
- Generuj, Wylicz, Filtruj i Wybierz węzły przy pomocy menu Generowanie.

Używanie pasków narzędzi

Informacjami pojawiającymi się w lewym i w prawym panelu można sterować przy pomocy opcji paska narzędzi. Można zmieniać orientacje ekranu (z góry na dół, od lewej do prawej, od prawej do lewej) przy pomocy elementów sterujących paska narzędzi. Ponadto można również przywrócić przeglądarkę do ustawień domyślnych i otworzyć okno dialogowe w celu określenia treści widoku skupień na panelu głównym.

Opcje **Sortowanie zmiennych według**, **Sortowanie skupień według**, **Komórki** oraz **Ekran** są dostępne tylko po wybraniu widoku **Grupy** na panelu głównym. Więcej informacji można znaleźć w temacie “Widok skupień” na stronie 251.

Tabela 14. Ikony na pasku narzędzi

Ikona	Temat
	Patrz Transponuj grupy i zmienne
	Patrz Sortowanie zmiennych według
	Patrz Sortowanie skupień według
	Patrz Komórki

Generowanie węzłów z Modeli skupień

Menu Generowanie umożliwia tworzenie nowych węzłów na podstawie modelu skupień. Opcja ta jest dostępna w zakładce Model generowanego modelu i pozwala na generowanie węzłów na podstawie bieżącego wyświetlenia lub wyboru (to znaczy, wszystkich widocznych lub wybranych skupień). Można na przykład wybrać pojedynczą zmienną, a następnie wygenerować Węzeł filtra, aby odrzucić wszystkie inne (niewidoczne) zmienne. Wygenerowane węzły są umieszczane, niepołączone, w obszarze roboczym. Ponadto można wygenerować kopię wartościowych informacji na palecie modeli. Pamiętaj, aby połączyć węzły i dokonać wszelkich zmian przed rozpoczęciem generowania.

- **Utwórz węzeł modelowania.** Tworzy węzeł modelowania w obszarze roboczym strumienia. Może to być przydatne na przykład w sytuacji, gdy mamy strumień, w którym chcemy użyć tych ustawień modelu, ale nie mamy już węzła modelowania służącego do generowania ustawień.
- **Model do palety.** Tworzy wartościowe informacje na palecie Modelu. Opcja ta jest przydatna w sytuacjach, gdy kolega/koleżanka wysłał(a) użytkownikowi strumień zawierający model, a nie sam model.
- **Węzeł filtra.** Tworzy nowy Węzeł filtra, służący do filtrowania zmiennych, które nie są używane przez model skupień i/lub są niewidoczne w bieżącym widoku Przeglądarki skupień. Jeśli nad tym Węzłem skupienia znajduje się Węzeł typu, wszystkie zmienne w roli *Zmienna przewidywana* są odrzucane przez wygenerowany Węzeł filtra.
- **Węzeł filtra (od wyboru).** Tworzy nowy węzeł filtra do filtrowania zmiennych na podstawie wyborów dokonanych w Przeglądarce skupień. Wybierz wiele zmiennych, korzystając z metody Ctrl+kliknięcie. Zmienne wybrane w Przeglądarce skupień są odrzucane w dół, ale można zmienić to zachowanie edytując Węzeł filtra przed wykonaniem filtrowania.
- **Węzeł wyboru.** Tworzy nową opcję Węzeł wyboru do wybierania rekordów na podstawie ich przynależności do dowolnego skupienia, widocznego w bieżącym widoku Przeglądarki skupień. Automatycznie generowany jest warunek wyboru.

- **Węzeł wyboru (od wyboru).** Tworzy nową opcję Węzeł wyboru do wybierania rekordów na podstawie ich przynależności do skupień wybranych w Przeglądarce skupień. Wybierz wiele skupień, korzystając z metody Ctrl+kliknięcie.
- **Węzeł wyliczenia.** Tworzy nowy Węzeł wyliczenia, który wylicza zmienną flagi, która przypisuje rekordy o wartości *Prawda* lub *Fałsz* na podstawie przynależności do wszystkich skupień widocznych w Przeglądarce skupień. Automatycznie generowany jest warunek wyliczenia.
- **Węzeł wyliczenia (od wyboru).** Tworzy nowy Węzeł wyliczenia, który wylicza zmienną flagi na podstawie jej przynależności do skupień wybranych w Przeglądarce skupień. Wybierz wiele skupień, korzystając z metody Ctrl+kliknięcie.

Poza generowaniem węzłów można również utworzyć wykresy w menu Generuj. Więcej informacji można znaleźć w temacie “Tworzenie wykresów na podstawie modeli skupień”.

Kontrolowanie wyświetlania widoku skupień

Aby kontrolować, co się wyświetla w widoku skupień na panelu głównym, kliknij przycisk **Wyświetl**; otworzy się okno dialogowe Wyświetl.

Właściwości. Wybrane domyślnie. Aby ukryć wszystkie zmienne wejściowe, odznacz to pole wyboru.

Zmienne ewaluacyjne. Wybierz zmienne ewaluacyjne (zmienne niesłużące do tworzenia modelu skupień, ale wysyłane do przeglądarki modelu w celu oceny skupień do wyświetlenia; domyślnie nie wyświetlają się żadne skupienia. *Uwaga* Zmienna ewaluacyjna musi być łańcuchem z więcej niż jedną wartością. To pole wyboru jest niedostępne, jeśli nie są dostępne żadne zmienne ewaluacyjne.

Opisy grup. Wybrane domyślnie. Aby ukryć wszystkie opisy skupień, odznacz to pole wyboru.

Rozmiary grup. Wybrane domyślnie. Aby ukryć wszystkie komórki rozmiarów skupień, odznacz to pole wyboru.

Maksymalna liczba kategorii. Podaj maksymalną liczbę kategorii, które mają się wyświetlać na wykresie zmiennych jakościowych; domyślna ilość to 20.

Tworzenie wykresów na podstawie modeli skupień

Modele skupień dostarczają bardzo wielu informacji, które jednak nie zawsze podane są w formie łatwo dostępnej dla użytkowników biznesowych. Aby przedstawić dane w postaci odpowiedniej do uwzględnienia w raportach biznesowych, prezentacjach itp., można tworzyć wykresy na podstawie wybranych danych. Na przykład w Widoku skupień można utworzyć wykres dla wybranego skupienia, tj. wykres przedstawiający tylko obserwacje należące do tego skupienia.

Uwaga: Wykres z Widoku skupień można wygenerować tylko wtedy, gdy model użytkowy jest połączony z innymi węzłami w strumieniu.

Tworzenie wykresu

1. Otwórz model użytkowy zawierający Widok skupień.
2. Na karcie Model wybierz opcję *Grupy* z listy rozwijanej **Widok**.
3. W widoku głównym wybierz skupienia (grupy), na podstawie których chcesz utworzyć wykres.
4. Z menu Utwórz wybierz polecenie **Wykres (z wyboru)**; zostanie wyświetlona karta podstawowej wizualizacji.

   Uwaga: Po wyświetleniu wizualizacji opisanym sposobem dostępna będzie tylko karta Podstawowe i Zaawansowane.
5. Korzystając z ustawień na karcie podstawowej lub zaawansowanej wizualizacji, określ informacje, które mają być widoczne na wykresie.
6. Kliknij przycisk OK, aby utworzyć wykres.

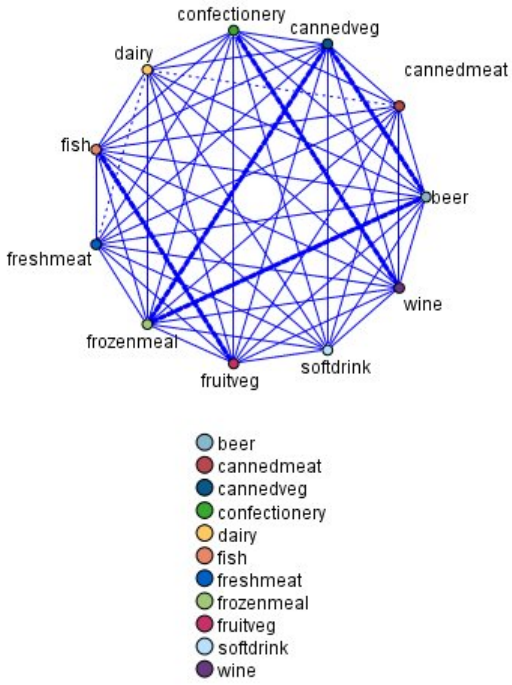
Nagłówek wykresu zawiera informację o typie modelu oraz skupieniu lub skupieniach, które wybrano do uwzględnienia na wykresie.

Rozdział 12. Reguły asocjacyjne

Reguły asocjacyjne kojarzą konkretny wniosek (np. decyzję o zakupie konkretnego produktu) ze zbiorem warunków (np. zakupem kilku innych produktów). Na przykład reguła

```
beer <= cannedveg & frozenmeal (173, 17.0%, 0.84)
```

mówi, że *beer* często występuje wtedy, gdy jednocześnie występują *cannedveg* i *frozenmeal*. Ta reguła jest w 84% niezawodna i ma zastosowanie do 17% danych, czyli 173 rekordów. Algorytmy reguł asocjacyjnych automatycznie znajdują związki, które można byłoby znaleźć ręcznie przy użyciu technik wizualizacji (por. węzeł sieciowy).



Rysunek 45. Węzeł sieciowy przedstawiający związki między elementami w koszyku zakupów

Przewagą algorytmów reguł asocjacyjnych wobec bardziej standardowych algorytmów drzew decyzyjnych (C5.0 i C&RTs) jest fakt, że dozwolone są w nich związki między *dowolnymi* atrybutami. Algorytm drzewa decyzyjnego pozwala utworzyć reguły z tylko jednym wnioskiem, podczas gdy algorytmy powiązań próbują znaleźć wiele reguł, z których każda może mieć inny wniosek.

Wadą algorytmów asocjacyjnych jest fakt, że próbują one znaleźć wzorce w potencjalnie bardzo dużej przestrzeni wyszukiwania i w związku z tym ich wykonanie może trwać znacznie dłużej niż wykonanie algorytmu drzewa decyzyjnego. Algorytmy te do znajdowania reguł stosują metodę **generowania i testowania** — początkowo generują proste reguły i walidują je względem zbioru danych. Dobre reguły są zachowywane, a wszystkie reguły, z zachowaniem różnych ograniczeń, podlegają specjalizacji. **Specjalizacja** polega na dodawaniu warunków do reguły. Uzyskane nowe reguły są walidowane względem danych i proces znów zapisuje najlepsze i najbardziej interesujące reguły. Użytkownik zwykle nakłada jakieś ograniczenie na liczbę poprzedników dozwolonych w regule. Ponadto w celu ograniczenia potencjalnie dużej przestrzeni wyszukiwania stosowane są różne techniki oparte na teorii informacji i systemach efektywnego indeksowania.

Po zakończeniu przetwarzania prezentowana jest tabela najlepszych reguł. W odróżnienia od drzewa decyzyjnego ten zestaw reguł asocjacyjnych nie może być używany bezpośrednio do generowania predykcji, tak jak model standardowy

(np. drzewo decyzyjne lub sieć neuronowa). Wynika to z faktu, że reguły mogą prowadzić do wielu różnych wniosków. Konieczny jest kolejny poziom transformacji, który przekształci reguły asocjacyjne w zestaw reguł klasyfikacji. Dlatego reguły asocjacyjne wygenerowane przez algorytmy asocjacyjne nazywamy **modelami surowymi**. Mimo że użytkownik może przeglądać takie modele surowe, nie można ich używać wprost jako modeli klasyfikacyjnych, jeśli użytkownik nie nakaże systemowi wygenerowania modelu klasyfikacyjnego z modelu surowego. Można to zrobić za pośrednictwem opcji menu Utwórz w przeglądarce.

Obsługiwane są dwa algorytmy reguł asocjacyjnych:

Węzeł Apriori pozwala wyodrębnić zestaw reguł na podstawie danych, pobierając reguły o najwyższej możliwej zawartości informacji. Apriori oferuje pięć różnych metod reguł wybierania i korzysta ze złożonego schematu indeksowania do efektywnego przetwarzania dużych zbiorów danych. W przypadku dużych problemów czas uczenia Apriori jest zwykle krótszy. Brak jest arbitralnego limitu co do liczby reguł do utrzymania, możliwa jest obsługa reguł z maksymalnie 32 predykcjami. Apriori wymaga, aby wszystkie zmienne wejściowe i wyjściowe były zmiennymi jakościowymi, lecz oferuje wyższą wydajność z uwagi na optymalizację pod kątem tego typu danych.

Węzeł Sekwencje wykrywa reguły asocjacyjne w danych sekwencyjnych lub zorientowanych czasowo. Sekwencja to lista zbiorów elementów z tendencją do występowania w przewidywalnej kolejności. Na przykład klient dokonujący zakupu brzytwy i balsamu po goleniu przy następnej wizycie w sklepie może dokonać zakupu kremu po goleniu. Węzeł Sekwencje bazuje na algorytmie reguł asocjacyjnych CARMA, który korzysta z efektywnej metody dwu przejść do znajdowania sekwencji.

Dane tabelaryczne a dane transakcyjne

Dane używane przez modele reguł asocjacyjnych mogą mieć format transakcyjny lub tabelaryczny, zgodnie z opisem poniżej. Są to opisy ogólne; specyficzne wymagania mogą być inne, co omówiono w dokumentacji dla każdego typu modelu. Należy zwrócić uwagę, że podczas przeprowadzania oceny modeli dane poddawane ocenie muszą mieć taki sam format, jak dane użyte do utworzenia modelu. Modele utworzone z zastosowaniem danych tabelarycznych mogą być używane do oceniania wyłącznie danych tabelarycznych; modele utworzone przy użyciu danych transakcyjnych mogą przeprowadzać ocenę tylko danych transakcyjnych.

Format transakcyjny

Dane transakcyjne są zapisywane w postaci osobnego rekordu dla każdej transakcji lub pozycji. Jeśli klient dokonuje kilku zakupów, każdy będzie zapisany w osobnym folderze, wraz z powiązanymi elementami dowiązanymi na podstawie id. klienta. Jest on również znany jako format **kasowy**.

Klient	Zakup
1	dżem
2	mleko
3	dżem
3	chleb
4	dżem
4	chleb
4	mleko

Węzły Apriori, CARMA i Sekwencje mogą korzystać z danych transakcyjnych.

Dane tabelaryczne

Dane tabelaryczne (znane również jako dane **z koszyka** lub **tabeli prawdy**) zawierają pozycje reprezentowane przez osobne flagi, a każda flaga reprezentuje obecność lub nieobecność konkretnej pozycji. Każdy rekord reprezentuje kompletny zestaw powiązanych pozycji. Zmienne flagi mogą być jakościowe lub numeryczne, choć dla niektórych modeli mogą istnieć bardziej specyficzne wymagania.

Klient	Dżem	Chleb	Mleko
1	T	F	F
2	F	F	T
3	T	T	F
4	T	T	T

Węzły Apriori, CARMA, GSAR i Sekwencje mogą korzystać z danych tabelarycznych.

---

węzeł Apriori

Węzeł Apriori wykrywa także reguły asocjacyjne w danych. Apriori oferuje pięć różnych metod reguł filtrowania i korzysta ze złożonego schematu indeksowania do efektywnego przetwarzania dużych zbiorów danych.

Wymagania. Do utworzenia zestawu reguł apriori potrzebna jest co najmniej jedna zmienna *wejściowa* i co najmniej jedna zmienna *przewidywana*. Zmienne wejściowe i przewidywane (o rolach zmiennej *wejściowej*, *przewidywanej* lub *obu tych rolach*) muszą być symboliczne. Zmienne o roli *Brak* są ignorowane. Przed wykonaniem węzła typy zmiennych muszą być zrealizowane jako instancje zmiennych. Dane mogą mieć format tabelaryczny lub transakcyjny. Więcej informacji można znaleźć w temacie “Dane tabelaryczne a dane transakcyjne” na stronie 258.

Mocne strony. W przypadku dużych problemów uczenie modeli apriori z reguły trwa krócej.. Brak jest arbitralnego limitu co do liczby zachowywanych reguł, możliwa jest obsługa reguł z maksymalnie 32 warunkami wstępnymi. Model Apriori oferuje pięć różnych metod szkolenia, co pozwala na bardziej elastyczne dopasowanie metod eksploracji danych do rozwiązywanego problemu.

Opcje modelu węzła Apriori

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Minimalne pokrycie poprzedników. Można określić kryterium pokrycia wpływające na zachowywanie reguł w zestawie reguł. **Pokrycie** oznacza odsetek rekordów w danych uczących, dla których poprzedniki (część „jeśli” reguły) są prawdziwe. (Uwaga: ta definicja pokrycia jest inna od stosowanej w węzłach CARMA i Kolejność. Więcej informacji można znaleźć w temacie “Opcje modelu węzła Sekwencje” na stronie 274.). W przypadku uzyskiwania reguł mających zastosowanie do bardzo małych podzbiorów danych należy podjąć próbę zwiększenia tego ustawienia.

Uwaga: Definicja pokrycia w przypadku modelu Apriori jest oparta na liczbie rekordów z poprzednikami. Inaczej jest w algorytmach CARMA i Kolejność, w których definicja pokrycia bazuje na liczbie rekordów ze wszystkimi elementami w regule (tj. zarówno poprzednikami, jak i następnikami). W wynikach dla modeli asocjacyjnych uwzględniona jest zarówno miara porycia poprzedników, jak i miara pokrycia reguł.

Minimalna ufność reguły. Można także określić kryterium ufności. **Ufność** wyznaczana jest na podstawie rekordów, dla których poprzedniki reguły są prawdziwe, i stanowi wśród nich odsetek rekordów, dla których następniki także są prawdziwe. Innymi słowy, jest to odsetek prawidłowych predykcji na podstawie reguły. Reguły o ufności niższej od podanego kryterium są odrzucane. Jeśli uzyskiwana liczba reguł jest zbyt duża, należy podjąć próbę zwiększenia tej wartości. Jeśli uzyskiwana liczba reguł jest za mała (lub nie są generowane żadne reguły), należy podjąć próbę zmniejszenia tej wartości.

Uwaga: W razie potrzeby można zaznaczyć wartość i zamiast niej wpisać własną wartość. Należy pamiętać, że zmniejszenie ufności poniżej 1,0 nie tylko spowoduje znaczne zwiększenie zapotrzebowania na pamięć, ale także potencjalne znaczące wydłużenie czasu budowania reguł.

Maksymalna liczba poprzedników. Dla dowolnej reguły można określić maksymalną liczbę warunków wstępnych. Jest to sposób na ograniczenie złożoności reguł. Jeśli reguły są zbyt złożone lub zbyt swoiste, należy podjąć próbę zmniejszenia tego ustawienia. Ustawienie to ma także silny wpływ na czas uczenia. Jeśli uczenie zestawu reguł trwa zbyt długo, należy podjąć próbę zmniejszenia tego ustawienia.

Tylko wartości true dla flag. Jeśli ta opcja zostanie wybrana dla danych w formacie tabelarycznym (tabela prawdy), to w wynikowych regułach zostaną ujęte tylko wartości „true". Może to ułatwić zrozumienie reguły. Ta opcja nie ma zastosowania do danych w formacie transakcyjnym. Więcej informacji można znaleźć w temacie "Dane tabelaryczne a dane transakcyjne" na stronie 258.

Uwaga: Węzeł budujący model CARMA ignoruje rekordy puste podczas budowania modelu, jeśli zmienna ma typ flagi, podczas gdy węzeł budujący model Apriori uwzględnia rekordy puste. Puste rekordy to rekordy, w których wszystkie zmienne używane w budowie modelu mają wartość fałszywą.

Optymalizuj ze względu na. Można tutaj wybrać opcje służące do optymalizacji wydajności podczas tworzenia modelu.

- Zaznaczenie opcji **Szybkość** uniemożliwia przenoszenie danych na dysk, co przekłada się na większą wydajność.
- Zaznaczenie opcji **Pamięć** sprawia, że algorytm przenosi dane na dysk kosztem szybkości przetwarzania. Ta opcja jest wybrana domyślnie.

  Uwaga: W trybie analizy rozproszonej to ustawienie może być przesłonięte przez opcje administratora określone w pliku *options.cfg*. Więcej informacji zawiera publikacja *IBM SPSS Modeler Server Administrator's Guide*.

Opcje zaawansowane węzła Apriori

Dla użytkowników, którzy orientują się w szczegółach działania algorytmu Apriori, dostępne są następujące opcje zaawansowane umożliwiające precyzyjne dostosowywanie procesu wywodzenia. W celu uzyskania dostępu do opcji zaawansowanych należy ustawić opcję Tryb na wartość **Zaawansowany** na karcie Zaawansowany.

Miara ewaluacyjna. Algorytm Apriori oferuje pięć metod oceny potencjalnych reguł.

- **Ufność reguły.** Metodą domyślną jest ocena reguł na podstawie ich ufności (lub dokładności). W przypadku tej miary pole **Dolna granica miary ewaluacyjnej** jest wyłączone, ponieważ jest zbędne wobec istnienia analogicznej opcji **Minimalna ufność reguły** na karcie Model. Więcej informacji można znaleźć w temacie "Opcje modelu węzła Apriori" na stronie 259.
- **Różnica ufności.** (Nazywana także **bezwzględną różnicą ufności względem poprzedniej**). Ta miara ewaluacyjna jest bezwzględną różnicą między ufnością reguły a jej poprzednią ufnością. Ta opcja pozwala uniknąć odchylenia w sytuacji, gdy wyniki nie są rozłożone równomiernie. Dzięki temu nie są zachowywane reguły „oczywiste". Wyobraźmy sobie, że 80% klientów naszej firmy kupuje najpopularniejszy produkt. Reguła, która przewidzi zakup tego popularnego produktu z 85-procentową dokładnością nie przynosi nam istotnej dodatkowej wiedzy, mimo że na skali bezwzględnej dokładność równa 85% wydaje się być wysoka. Dolną granicę miary ewaluacyjnej należy ustawić na minimalną różnicę ufności, przy której reguły mają być zachowywane.
- **Iloraz ufności.** (Nazywany także **różnicą współczynnika ufności względem 1**). Ta miara ewaluacyjna stanowi stosunek ufności reguły do poprzedniej ufności (lub, jeśli iloraz jest większy od 1, jego odwrotność) odjęty od 1. Podobnie jak Różnica metoda ta bierze pod uwagę nierównomierność rozkładu. Szczególnie dobrze nadaje się do znajdowania reguł przewidujących rzadkie zdarzenia. Załóżmy na przykład, że pewne rzadkie schorzenie występuje tylko u 1% pacjentów. Reguła zdolna przewidzieć to schorzenie w 10% przypadków jest znacznie lepsza niż losowe zgadywanie, mimo że na skali bezwzględnej dokładność na poziomie 10% wydaje się niska. Dolną granicę miary ewaluacyjnej należy ustawić na różnicę ufności, przy której reguły mają być zachowywane.
- **Różnica informacyjna.** (Nazywana także **różnicą informacyjną względem poprzedniej**). Ta miara oparta jest na **zysku informacyjnym**. Jeśli prawdopodobieństwo konkretnego następnika potraktujemy jako wartość logiczną (**bit**), to zysk informacyjny stanowi część, w jakiej można określić wartość tego bitu na podstawie poprzedników.

Różnica informacyjna jest to różnica między zyskiem informacyjnym przy danych poprzednikach a zyskiem informacyjnym przy znanej wyłącznie poprzedniej ufności następnika. Ważną cechą tej metody jest fakt, że uwzględnia ona pokrycie, tak że przy tym samym poziomie ufności preferowane są reguły pokrywające więcej rekordów. Dolną granicę miary ewaluacyjnej należy ustawić na różnicę informacyjną, przy której reguły mają być zachowywane.

Uwaga: Ponieważ skala tej miary jest nieco mniej intuicyjna niż inne, konieczne może być ustalenie dolnej granicy drogą eksperymentalną w celu uzyskania zadowalającego zestawu reguł.

- **Znormalizowany Chi-kwadrat.** (Nazywana także **znormalizowaną miarą chi-kwadrat**). Ta miara jest statystycznym indeksem związku między poprzednikami a następnikami. Jest znormalizowana w taki sposób, by przyjmowała wartość z przedziału od 0 do 1. Ta miara jest jeszcze silniej zależna od pokrycia niż różnica informacyjna. Dolną granicę miary ewaluacyjnej należy ustawić na różnicę informacyjną, przy której reguły mają być zachowywane.

  Uwaga: Podobnie jak w przypadku różnicy informacyjnej, skala tej miary jest nieco mniej intuicyjna niż inne, konieczne może być ustalenie dolnej granicy drogą eksperymentalną w celu uzyskania zadowalającego zestawu reguł.

Zezwalaj na reguły bez poprzedników. Wybranie tej opcji spowoduje, że dozwolone będą reguły zawierające tylko następnik (jeden element lub zbiór elementów). Jest to przydatne, gdy interesuje nas ustalenie wspólnych elementów lub zbiorów elementów. Na przykład cannedveg to reguła jednoelementowa bez poprzednika, która wskazuje, że w danych często występuje zakup towaru *cannedveg*. W niektórych przypadkach chcemy uwzględnić takie reguły, jeśli interesują nas po prostu predykcje o największej ufności. Ta opcja jest domyślnie wyłączona. Obowiązuje konwencja, zgodnie z którą pokrycie poprzedników przez reguły bez poprzedników wyrażone jest wartością 100%, a pokrycie reguły będzie takie samo, jak jej ufność.

Węzeł CARMA

Węzeł CARMA stosuje algorytm wykrywania reguł asocjacyjnych w danych. Reguły asocjacyjne są stwierdzeniami w postaci

jeśli *poprzednik(i)* **to** *następnik(i)*

Na przykład, jeśli klient sklepu internetowego kupi kartę bezprzewodową i wysokiej klasy router bezprzewodowy, to prawdopodobnie klient ten kupi także bezprzewodowy serwer muzyki, jeśli taki produkt zostanie mu zaoferowany. Model CARMA pozwala wyodrębnić zestaw reguł na podstawie danych bez konieczności określania zmiennych wejściowych lub przewidywanych. Oznacza to, że wygenerowane reguły są potencjalnie przydatne w szerszym spektrum zastosowań. Wygenerowane przez ten węzeł reguły można na przykład zastosować do zbudowania listy produktów lub usług (poprzedników), z których wynikać będzie decyzja o promowaniu konkretnego produktu (następnika) w tegorocznym sezonie świątecznym. Korzystając z programu IBM SPSS Modeler, można określić, którzy klienci kupili poprzedniki, i zbudować kampanię marketingową promującą następniki.

Wymagania. W odróżnieniu od węzła Apriori, węzeł CARMA nie wymaga zmiennych *wejściowych* ani *przewidywanych*. Wynika to wprost z zasady działania algorytmu i jest równoważne budowaniu modelu Apriori ze zmiennymi ustawionymi na rolę *Łącznie*. Filtrując już utworzony model, można nakazać wyświetlanie określonych elementów tylko jako poprzedników lub tylko jako następników. Na przykład za pomocą przeglądarki modelu można na przykład zbudować listy produktów lub usług (poprzedników), z których wynikać będzie decyzja o promowaniu konkretnego produktu w tegorocznym sezonie świątecznym.

W celu utworzenia zestawu reguł CARMA konieczne jest określenie zmiennej identyfikacyjnej i co najmniej jednej zmiennej zawartości. Zmienną identyfikacyjną może charakteryzować dowolna rola i dowolny poziom pomiaru. Zmienne o roli *Brak* są ignorowane. Przed wykonaniem węzła typy zmiennych muszą być zrealizowane jako instancje zmiennych. Podobnie jak w przypadku węzła Apriori, dane mogą mieć format tabelaryczny lub transakcyjny. Więcej informacji można znaleźć w temacie “Dane tabelaryczne a dane transakcyjne” na stronie 258.

Mocne strony. Węzeł CARMA działa w oparciu o algorytm reguł asocjacyjnych CARMA. W odróżnienia od węzła Apriori węzeł CARMA oferuje ustawienia tworzenia dotyczące pokrycia reguł (pokrycie zarówno poprzedników, jak i następników) zamiast pokrycia tylko poprzedników. Algorytm CARMA dopuszcza także reguły z więcej niż jednym

następnikiem. Podobnie jak w przypadku węzła Apriori, modele generowane przez węzeł CARMA można wstawiać do strumienia danych w celu utworzenie predykcji. Więcej informacji można znaleźć w temacie “Modele użytkowe” na stronie 37.

Opcje zmiennych węzła CARMA

Przed wykonaniem węzła CARMA należy określić zmienne wejściowe na karcie Zmienne węzła CARMA. Podczas gdy większość węzłów modelowania oferuje na karcie Zmienne identyczne opcje, węzeł CARMA zawiera kilka unikalnych ustawień. Wszystkie opcje zostały omówione poniżej.

Użyj ustawień węzła Typ. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej z wcześniejszego węzła Typ. Jest to ustawienie domyślne.

Użyj ustawień niestandardowych. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej określonych w tym miejscu, a nie w żadnym wcześniejszym węźle Typ. Po wybraniu tej opcji wprowadź informacje w odpowiednich polach, w zależności od tego, czy dane są odczytywane w formacie transakcyjnym, czy tabelarycznym.

Użyj formatu transakcyjnego. Ta opcja zmienia elementy sterujące zmiennej w pozostałej części tego okna dialogowego w zależności od tego, czy dane mają format transakcyjny czy tabelaryczny. W przypadku użycia wielu zmiennych z danymi transakcyjnymi zakłada się, że elementy określone w tych zmiennych dla konkretnego rekordu reprezentują elementy znajdujące się w pojedynczej transakcji i opatrzone pojedynczym znacznikiem czasu. Więcej informacji można znaleźć w temacie “Dane tabelaryczne a dane transakcyjne” na stronie 258.

Dane tabelaryczne

Jeśli opcja **Użyj formatu transakcyjnego** nie jest wybrana, wyświetlane są następujące pola.

- **Zmienne wejściowe.** Umożliwia wybór zmiennej wejściowej lub kilku zmiennych. Działanie jest podobne, jak w przypadku ustawienia roli zmiennej na *Zmienna wejściowa* w węźle Typ.
- **Podział.** To pole umożliwia określenie zmiennej używanej do podziału danych na osobne próby do uczenia, testowania i walidacji podczas budowania modelu. Korzystając z jednej próby do generowania modelu oraz innej do testowania go, można uzyskać wiarygodne informacje o tym, jak dobrze model pozwala uogólnić większe zbiory danych, podobne do danych bieżących. Jeśli korzystając z węzłów Typ lub Partycja, zdefiniowano wiele zmiennych dzielących na podzbiory, na karcie Zmienne każdego węzła modelowania korzystającego z tego podziału na podzbiory należy wybrać jedną zmienną dzielącą na podzbiory. (Jeśli obecna jest tylko jedna zmienna dzieląca na podzbiory, jest ona automatycznie używana po aktywowaniu dzielenia). Należy również pamiętać, że aby zastosować wybrany podział w analizie, dzielenie musi być również włączone na karcie Opcje modelu danego węzła. (Usunięcie zaznaczenia tej opcji umożliwia wyłączenie podziału bez zmiany ustawień zmiennych).

Dane transakcyjne

Jeśli opcja **Użyj formatu transakcyjnego** jest wybrana, wyświetlane są następujące pola.

- **Identyfikator.** W przypadku danych transakcyjnych należy wybrać z listy zmienną identyfikacyjną. Jako zmienna identyfikacyjna mogą być używane zmienne numeryczne lub symboliczne. Każda unikalna wartość tej zmiennej powinna wskazywać na określoną jednostkę analizy. Na przykład w aplikacji do obsługi koszyka zakupów każdy identyfikator może reprezentować jednego klienta. W przypadku aplikacji do analizy dzienników sieciowych każdy identyfikator może reprezentować komputer (wg adresu IP) lub użytkownika (wg danych logowania).
- **Wartości identyfikatorów są posortowane.** (Tylko węzły Apriori i CARMA) Jeśli dane zostały wstępnie posortowane tak, że wszystkie rekordy o tym samym identyfikatorze są zgrupowane w strumieniu danych, należy wybrać tę opcję w celu przyspieszenia przetwarzania. Jeśli dane nie zostały wstępnie posortowane (lub nie ma co do tego pewności), należy pozostawić tę opcję niezaznaczoną. Węzeł posortuje dane automatycznie.

  Uwaga: Jeśli dane nie są posortowane, a użytkownik wybierze tę opcję, model może zwrócić niepoprawne wyniki.
- **Zawartość.** Należy określić zmienne zawartości dla modelu. Zmienne te zawierają interesujące elementy w procesie modelowania sekwencji. Można określić wiele zmiennych typu flaga (jeśli dane mają format tabelaryczny) lub jedną zmienną nominalną (jeśli dane mają format transakcyjny).

Opcje modelu węzła CARMA

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Minimalne pokrycie reguł (%). Można określić kryterium pokrycia. **Pokrycie reguł** określa, jaki odsetek identyfikatorów w danych uczących zawiera całą regułę. (Uwaga: ta definicja pokrycia różni się od definicji pokrycia poprzedników używanej w węzłach Apriori). W celu skoncentrowania się na bardziej typowych regułach należy zwiększyć wartość tego ustawienia.

Minimalna ufność reguły (%). Można określić kryterium ufności wpływające na zachowywanie reguł w zestawie reguł. **Ufność** określa, dla jakiego odsetka identyfikatorów generowana jest poprawna predykcja (spośród wszystkich identyfikatorów, dla których ta reguła powoduje wygenerowanie predykcji). Jest ona obliczana jako iloraz liczby identyfikatorów, dla których znajduje się cała reguła, i liczby identyfikatorów, dla których w oparciu o dane uczące znajdują się poprzedniki. Reguły o ufności niższej od podanego kryterium są odrzucane. Jeśli uzyskiwana liczba reguł jest zbyt duża lub reguły nie są interesujące, należy podjąć próbę zwiększenia tej wartości. W przypadku uzyskiwania zbyt małej liczby reguł należy sprobować obniżyć wartość tego ustawienia.

Uwaga: W razie potrzeby można zaznaczyć wartość i zamiast niej wpisać własną wartość. Należy pamiętać, że zmniejszenie ufności poniżej 1,0 nie tylko spowoduje znaczne zwiększenie zapotrzebowania na pamięć, ale także potencjalne znaczące wydłużenie czasu budowania reguł.

Maksymalny rozmiar reguły. Można określić maksymalną liczbę odrębnych zestawów elementów (w odróżnieniu od pojedynczych elementów) w regule. Jeśli interesujące reguły są relatywnie krótkie, można zmniejszyć wartość tego ustawienia w celu przyspieszenia budowy zestawu reguł.

Uwaga: Węzeł budujący model CARMA ignoruje rekordy puste podczas budowania modelu, jeśli zmienna ma typ flagi, podczas gdy węzeł budujący model Apriori uwzględnia rekordy puste. Puste rekordy to rekordy, w których wszystkie zmienne używane w budowie modelu mają wartość fałszywą.

Opcje zaawansowane węzła CARMA

Dla użytkowników, którzy orientują się w szczegółach działania algorytmu Węzeł CARMA, dostępne są następujące opcje zaawansowane umożliwiające precyzyjne dostosowywanie procesu wywodzenia. W celu uzyskania dostępu do opcji zaawansowanych należy ustawić opcję Tryb na wartość **Zaawansowany** na karcie Zaawansowany.

Wyklucz reguły z wieloma następnikami. Wybranie tej opcji spowoduje wykluczenie następników podwójnych. Na przykład reguła bread & cheese & fish -> wine&fruit zawiera następnik dwuelementowy wine&fruit. Domyślnie takie reguły są uwzględniane przez algorytm.

Wartość przycinania. Aby oszczędniej gospodarować pamięcią, algorytm CARMA usuwa (**przycina**) zestawy rzadko występujących elementów ze swojej listy potencjalnych zestawów. Za pomocą tej opcji można zmienić częstotliwość przycinania. Wprowadzenie mniejszej wartości powoduje obniżenie wymagań algorytmu co do ilości pamięci (lecz może także wydłużyć niezbędny czas uczenia), natomiast wprowadzenie większej wartości przyspieszy proces uczenia (lecz jednocześnie może spowodować zwiększenie ilości potrzebnej pamięci). Domyślna wartość to 500.

Zmieniaj pokrycie. Wybierz tę opcje, aby poprawić wydajność poprzez wykluczenie rzadko występujących zestawów elementów, których nierównomierne uwzględnienie sprawia, że pozornie występują często. Pożądany efekt osiąga się, zaczynając od wyższego pokrycia i zmniejszając je do poziomu określonego na karcie Model. Wprowadź wartość **Szacowana liczba transakcji**, aby określić, jak szybko ma być zmniejszane pokrycie.

Zezwalaj na reguły bez poprzedników. Wybranie tej opcji spowoduje, że dozwolone będą reguły zawierające tylko następnik (jeden element lub zbiór elementów). Jest to przydatne, gdy interesuje nas ustalenie wspólnych elementów lub zbiorów elementów. Na przykład cannedveg to reguła jednoelementowa bez poprzednika, która wskazuje, że w danych często występuje zakup towaru *cannedveg*. W niektórych przypadkach chcemy uwzględnić takie reguły, jeśli interesują nas po prostu predykcje o największej ufności. Domyślnie ta opcja jest niezaznaczona.

Modele użytkowe reguł asocjacyjnych

Modele użytkowe reguł asocjacyjnych odzwierciedlają reguły wykryte przez jeden z następujących węzłów modelowania reguł asocjacyjnych:

- Apriori
- CARMA

Modele użytkowe zawierają informacje o regułach wyodrębnionych z danych podczas budowy modelu.

Uwaga: Ocena modeli użytkowych reguł asocjacyjnych może być nieprawidłowa, jeśli dane transakcyjne nie będą posortowane według identyfikatora.

Wyświetlanie wyników

Reguły wygenerowane przez modele asocjacyjne (Apriori i CARMA) oraz modele sekwencji można przeglądać na karcie Model w oknie dialogowym. Przeglądanie modelu użytkowego ujawnia informacje o regułach i udostępnia opcje filtrowania i sortowania wyników przed wygenerowaniem nowych węzłów i przed oceną modelu.

Ocenianie modelu

Udoskonalone modele użytkowe (Apriori, CARMA i Sequence) można dodać do strumienia i wykorzystać w ocenianiu. Więcej informacji można znaleźć w temacie “Używanie modeli użytkowych w strumieniach” na stronie 48. Okna dialogowe wartościowych informacji z modelu, które są używane podczas oceniania, zawierają dodatkową kartę Ustawienia. Więcej informacji można znaleźć w temacie “Ustawienia modelu użytkowego reguły asocjacyjnej” na stronie 267.

Surowego modelu użytkowego nie można bezpośrednio użyć do oceniania. Zamiast tego można jednak wygenerować zestaw reguł i użyć go do oceniania. Więcej informacji można znaleźć w temacie “Generowanie zestawu reguł z powiązanego modelu użytkowego” na stronie 269.

Szczegóły modelu użytkowego reguły asocjacyjnej

Na karcie Model dla modelu użytkowego reguły asocjacyjnej widoczna jest tabela zawierająca reguły wyodrębnione przez algorytm. Każdy wiersz w tabeli odpowiada jednej regule. Pierwsza kolumna zawiera następniki (część „to” reguły), natomiast następna kolumna zawiera poprzedniki (część „jeśli” reguły). Kolumny następników zawierają informacje o regułach, takie jak ufność, pokrycie i wzrost.

Reguły asocjacyjne prezentowane są zwykle w formacie takim, jak przedstawiony w poniższej tabeli.

Tabela 15. Przykład reguły asocjacyjnej

Następnik	Poprzednik
Drug = drugY	Sex = F BP = HIGH

Przykładową regułę należy interpretować następująco: *jeśli Sex = "F" i BP = "HIGH," to Drug prawdopodobnie będzie równy drugY*; innymi słowy *dla rekordów, w których płeć = K, a ciśnienie tętnicze jest wysokiej, lekiem będzie prawdopodobnie lek Y*. Korzystając z paska narzędzi w oknie dialogowym można wybrać do wyświetlenia dodatkowe informacje, takie jak ufność, pokrycie i instancje.

Menu Sortowanie. Przycisk menu Sortowanie na pasku narzędzi steruje regułami sortowania. Kierunek sortowania (rosnąco/malejąco) można zmienić za pomocą przycisku kierunku sortowania (strzałka w górę lub w dół).

Reguły można sortować wg następujących pól:

- Pokrycie

- Ufność
- Pokrycie reguł
- Następnik
- Ewaluacja
- Wzrost
- Wdrażalność

Pokaż/Ukryj menu. Przycisk Pokaż/Ukryj menu (przycisk kryteriów na pasku narzędzi) steruje opcjami wyświetlania reguł.

Rysunek 46. Przycisk Pokaż/Ukryj

Dostępne są następujące opcje wyświetlania:

- **Identyfikator reguły** wyświetla identyfikator reguły przypisany podczas budowania modelu. Identyfikator reguły umożliwia określenie reguł stosowanych w danej predykcji. Identyfikatory reguł umożliwiają również późniejsze scalenie dodatkowych informacji o regułach, takich jak wdrażalność, informacje o produkcie lub następniki.
- **Instancje** wyświetla informacje o liczbie unikalnych identyfikatorów, do których reguła ma zastosowanie, tj. dla których poprzedniki są prawdziwe. Na przykład w przypadku reguły bread -> cheese liczba rekordów w danych uczących, które zawierają poprzednik *bread*, nazywana jest **instancjami** lub wystąpieniami.
- **Pokrycie** wyświetla pokrycie poprzedników, tj. odsetek identyfikatorów, dla których poprzedniki są prawdziwe, na podstawie danych uczących. Na przykład, jeśli 50% danych uczących obejmuje zakup chleba (bread), to reguła bread -> cheese będzie miała pokrycie poprzedników równe 50%. *Uwaga*: Zdefiniowane w ten sposób pokrycie jest równoważne liczbie instancji, ale wyrażone w procentach.
- **Ufność** wyświetla stosunek pokrycia reguły do pokrycia poprzedników. Jest to odsetek identyfikatorów z określonymi poprzednikami, dla których następniki także są prawdziwe. Na przykład, jeśli 50% danych uczących zawiera element bread (co oznacza pokrycie poprzednika), lecz tylko 20% zawiera zarówno element bread, jak i cheese, wówczas ufność dla reguły bread -> cheese będzie równa ilorazowi: Pokrycie reguł / Pokrycie poprzedników lub, w tym przypadku, 40%.
- **Pokrycie reguł** wyświetla odsetek identyfikatorów, dla których prawdziwa jest cała reguła, poprzedniki i następniki. Na przykład, jeśli 20% danych uczących zawiera zarówno element bread, jak i cheese, wówczas pokrycie reguły dla reguły bread -> cheese wynosi 20%.
- **Ocena** jest uwzględniana, jeśli wybrano jedno z zaawansowanych kryteriów reguły asocjacyjnej (różnica ufności, iloraz ufności, różnica informacyjna lub znormalizowany Chi-kwadrat). Miary tych zaawansowanych kryteriów są porównywane z liczbą **dolnej granicy miary ewaluacyjnej** ustawioną przez użytkownika (i obowiązują tylko, jeśli wybrano zaawansowane kryterium reguły). Statystyka oceny ma następujące znaczenie dla każdego zaawansowanego kryterium reguły asocjacyjnej:
  - Różnica ufności: ufność a posteriori - ufność wstępna
  - Iloraz ufności: (ufność a posteriori - ufność wstępna)/ufność a posteriori
  - Różnica informacyjna: miara korzyści informacyjnej
  - Znormalizowany Chi-kwadrat: statystyka znormalizowanego Ch-kwadrat

  Każda z tych statystyk jest porównywana z liczbą **dolnej granicy miary ewaluacyjnej** ustawioną przez użytkownika, a jeśli statystyki przekraczają tę liczbę, wybierana jest reguła.
- **Wzrost** wyświetla stosunek ufności reguły i prawdopodobieństwa apriori posiadania następnika. Na przykład, jeśli 10% całej populacji kupuje chleb (bread), to reguła przewidująca, czy ludzie kupią chleb z ufnością 20% będzie miała wzrost równy 20/10 = 2. Inna reguła, która mówi, że ludzie kupią chleb z ufnością 11%, ma wzrost bliski 1, co oznacza że posiadane poprzednika nie wpływa znacząco na prawdopodobieństwo posiadania następnika. Co do zasady reguły ze wzrostem różnym od 1 będą bardziej interesujące niż reguły ze wzrostem bliskim 1.

- **Wdrażalność** to miara określająca, jaki odsetek danych uczących spełnia warunki poprzednika, ale nie spełnia następnika. W kontekście zakupu produktów miara ta oznacza zasadniczo, jaki odsetek wszystkich klientów posiada (lub kupił) poprzedniki, ale nie kupił jeszcze następnika. Wdrażalność jest zdefiniowana jako ((Pokrycie poprzedników wyrażone liczbą rekordów - Pokrycie reguł wyrażone liczbą rekordów) / Liczba rekordów) * 100, gdzie *Pokrycie poprzedników* oznacza liczbę rekordów, dla których poprzedniki są prawdziwe, a *Pokrycie reguł* oznacza liczbę rekordów, dla których prawdziwy jest zarówno poprzednik, jak i następnik.

Przycisk Filtr. Przycisk Filtr (ikona lejka) w menu umożliwia rozwinięcie dolnej części okna dialogowego, zawierającej panel, w którym wyświetlane są filtry aktywnych reguł. Filtry służą do zawężenia liczby reguł wyświetlanych na karcie Modele.

Rysunek 47. Przycisk Filtr

Aby utworzyć filtr, kliknij ikonę Filtr po prawej stronie rozwiniętego panelu. Spowoduje to otwarcie osobnego okna dialogowego, w którym można określić ograniczenia co do wyświetlania reguł. Należy zwrócić uwagę, że przycisk Filtr jest często używany w połączeniu z menu Utwórz, w pierwszej kolejności do filtrowania reguł, a następnie do generowania modelu zawierającego ten podzbiór reguł. Aby uzyskać więcej informacji, patrz “Określanie filtrów dla reguł” poniżej.

Przycisk Znajdź regułę. Przycisk Znajdź regułę (ikona lornetki) umożliwia wyszukiwanie reguł na podstawie identyfikatora. W sąsiednim polu podana jest liczba obecnie wyświetlanych reguł spośród liczby dostępnych reguł. Identyfikatory reguł są przypisywane przez model w kolejności wykrywania i są dodawane do danych podczas oceniania.

Rysunek 48. Przycisk Znajdź regułę

Aby zmienić kolejność identyfikatorów reguł:

1. W programie IBM SPSS Modeler można zmienić kolejność identyfikatorów reguł, najpierw sortując tabelę z regułami według żądanej miary, np. ufności lub wzrostu.
2. Następnie należy użyć opcji z menu Utwórz, aby utworzyć filtrowany model.
3. W oknie dialogowym Filtrowany model wybierz opcję **Zmień numerowanie reguł kolejno rozpoczynając od** i podaj numer początkowy.

Więcej informacji można znaleźć w “Generowanie modelu filtrowanego” na stronie 269.

Określanie filtrów dla reguł

Przy domyślnych ustawieniach algorytmy generujące reguły, takie jak Apriori, CARMA i Kolejność, mogą generować dużą liczbę reguł, które trudno będzie w praktyce wykorzystać. Aby zwiększyć przejrzystość zbioru reguł podczas przeglądania lub aby usprawnić ocenianie reguł, należy rozważyć odfiltrowanie reguł w taki sposób, aby interesujące następniki i poprzedniki były lepiej wyeksponowane. Korzystając z opcji filtrowania na karcie Model przeglądarki reguł, można otworzyć okno dialogowe do określania kwalifikacji filtrów.

Następniki. Wybierz opcję **Włącz filtr**, aby aktywować opcje reguł filtrowania poprzez uwzględnienie lub wykluczenie określonych następników. Wybierz opcję **Dołącz dowolne z**, aby utworzyć filtr wybierający te reguły, które zawierają co najmniej jeden z określonych następników. Możesz także wybrać opcję **Wyklucz**, aby utworzyć filtr wykluczający określone następniki. Następniki można wybierać za pomocą ikony selektora po prawej stronie okna listy. Spowoduje to otwarcie okna dialogowego z listą wszystkich następników obecnych w wygenerowanych regułach.

Uwaga: Następniki mogą zawierać więcej niż jeden element. Filtry sprawdzają tylko to, czy następnik zawiera jeden z określonych elementów.

Poprzedniki. Wybierz opcję **Włącz filtr**, aby aktywować opcje reguł filtrowania poprzez uwzględnienie lub wykluczenie określonych poprzedników. Poprzedniki można wybierać za pomocą ikony selektora po prawej stronie okna listy. Spowoduje to otwarcie okna dialogowego z listą wszystkich poprzedników obecnych w wygenerowanych regułach.

- Wybierz opcję **Dołącz wszystkie z**, aby zdefiniować filtr inkluzywny, wymagający obecności w regule wszystkich określonych poprzedników.
- Wybierz opcję **Dołącz dowolne z**, aby utworzyć filtr wybierający te reguły, które zawierają co najmniej jeden z określonych poprzedników.
- Wybierz opcję **Wyklucz**, aby utworzyć filtr wykluczający reguły zawierające określony poprzednik.

Ufność. Wybierz opcję **Włącz filtr**, aby aktywować opcje reguł filtrowania na podstawie poziomu ufności reguły. Do określenia przedziału ufności można użyć elementów sterujących **Min.** i **Maks.**. Podczas przeglądania wygenerowanych modeli ufność jest podana jako wartość procentowa. Podczas oceniania wyników ufność jest wyrażona jako liczba z przedziału od 0 do 1.

Pokrycie poprzedników. Wybierz opcję **Włącz filtr**, aby aktywować opcje reguł filtrowania na podstawie pokrycia poprzedników w regule. Pokrycie poprzedników określa odsetek danych uczących, który zawiera te same poprzedniki, co bieżąca reguła, przez co jest miarą analogiczną do wskaźnika popularności. Przedział używany do filtrowania reguł na podstawie pokrycia można określić za pomocą elementów sterujących **Min.** i **Maks.**.

Wzrost. Wybierz opcję **Włącz filtr**, aby aktywować opcje reguł filtrowania na podstawie miary wzrostu reguły. *Uwaga*: Filtrowanie według wzrostu jest dostępne tylko w przypadku modeli asocjacyjnych zbudowanych w wersji późniejszej niż 8.5 lub w przypadku wcześniejszych modeli zawierających miarę wzrostu. Modele sekwencji nie zawierają tej opcji.

Kliknięcie przycisku **OK** spowoduje zastosowanie wszystkich filtrów włączonych w tym oknie dialogowym.

Tworzenie wykresów reguł

Węzły asocjacyjne dostarczają bardzo wielu informacji, które jednak nie zawsze podane są w formie łatwo dostępnej dla użytkowników biznesowych. Aby przedstawić dane w postaci odpowiedniej do uwzględnienia w raportach biznesowych, prezentacjach itp., można tworzyć wykresy na podstawie wybranych danych. Z karty Model można utworzyć wykres dla wybranej reguły, tj. wykres przedstawiający tylko obserwacje objęte tą regułą.

1. Na karcie Model wybierz regułę, którą chcesz przedstawić.
2. Z menu Utwórz wybierz opcję **Wykres (z wyboru)**. Zostanie wyświetlona karta podstawowej wizualizacji.

   Uwaga: Po wyświetleniu wizualizacji opisanym sposobem dostępna będzie tylko karta Podstawowe i Zaawansowane.
3. Korzystając z ustawień na karcie podstawowej lub zaawansowanej wizualizacji, określ informacje, które mają być widoczne na wykresie.
4. Kliknij przycisk OK, aby utworzyć wykres.

W nagłówku wykresu podane są informacje o regule i uwzględnionych poprzednikach.

Ustawienia modelu użytkowego reguły asocjacyjnej

Karta Ustawienia służy do określania opcji oceniania modeli asocjacyjnych (Apriori i CARMA). Ta karta jest dostępna tylko po dodaniu modelu użytkowego do strumienia w celu przeprowadzania oceniania.

Uwaga: Okno dialogowe do przeglądania modelu surowego nie zawiera karty Ustawienia, ponieważ nie można go ocenić. Aby ocenić model „surowy”, należy najpierw wygenerować zestaw reguł. Więcej informacji można znaleźć w temacie “Generowanie zestawu reguł z powiązanego modelu użytkowego” na stronie 269.

Maksymalna liczba predykcji Określ maksymalną liczbę predykcji, które będą uwzględnione dla każdego zbioru elementów koszyka. Ta opcja jest używana z pozycjami Kryterium reguły w celu wygenerowania „najważniejszych” predykcji, w których *najważniejsza* oznacza najwyższy poziom ufności, pokrycia, wzrostu itp.

Kryterium reguły Wybierz miarę, która będzie używana w celu określenia siły reguł. Reguły są sortowane wg siły kryteriów wybranych w tej opcji w celu uzyskania najważniejszych predykcji dla zbioru elementów. Dostępne kryteria przedstawiono na poniższej liście

- Ufność
- Pokrycie
- Pokrycie reguł (Pokrycie * Ufność)
- Wzrost
- Wdrażalność

Zezwalaj na powtórzone predykcje Wybierz tę opcję, aby podczas oceniania uwzględnić wiele reguł posiadających ten sam następnik. Wybranie tej opcji pola umożliwi ocenę na przykład następujących reguł:

```
bread & cheese -> wine
cheese & fruit -> wine
```

Wyłącz tę opcję, aby wykluczyć powtórzone predykcje podczas oceniania.

Uwaga: Reguły z wieloma następnikami (bread & cheese & fruit -> wine & pate) są traktowane jako powtórzone predykcje, pod warunkiem że wszystkie następniki (wine & pate) zostały wcześniej przewidziane.

Ignoruj niedopasowane elementy koszyka Wybierz, aby ignorować obecność dodatkowych elementów w zestawie elementów. Na przykład, gdy ta opcja jest wybrana dla koszyka zawierającego [tent & sleeping bag & kettle], reguła tent & sleeping bag -> gas_stove będzie miała zastosowanie mimo obecności w koszyku dodatkowego elementu (kettle).

W pewnych okolicznościach dodatkowe elementy powinny zostać wykluczone. Na przykład prawdopodobne jest, że osoba, która kupuje tent (namiot), sleeping bag (śpiwór) i kettle (kociołek), posiada już kuchenkę gazową (gas stove), skoro kupuje kociołek. Innymi słowy, kuchenka gazowa może nie być trafną predykcją. W takich przypadkach należy usunąć zaznaczenie opcji **Ignoruj niedopasowane elementy koszyka**, aby mieć pewność, że poprzedniki będą dokładnie dopasowane do zawartości koszyka. Domyślnie elementy niedopasowane są ignorowane.

Sprawdź, czy predykcji nie ma w koszyku. Wybierz tę opcję, aby mieć pewność, że następniki nie są także obecne w koszyku. Na przykład, jeśli celem oceny jest uzyskanie rekomendacji na temat mebli domowych, wówczas mało prawdopodobne jest to, że osoba z koszykiem, który zawiera już stół do jadalni, kupi kolejny taki stół. W takich właśnie przypadkach należy wybrać tę opcję. Natomiast w przypadku produktów psujących się lub jednorazowych (takich jak ser, mleko dla dzieci lub chusteczka higieniczna) wartościowe mogą być reguły, w których następnik istnieje już w koszyku. W tym drugim przypadku najbardziej użyteczną opcją może być **Nie sprawdzaj koszyka pod kątem predykcji** (patrz niżej).

Sprawdź, czy predykcje są w koszyku Wybierz tę opcję, aby upewnić się, że następniki także są obecne w koszyku. Takie podejście jest stosowane przy próbie uzyskania wglądu w istniejących klientów lub istniejące transakcje. Na przykład może pojawić się potrzeba znalezienia reguł o największym wzroście, a następnie sprawdzenia, którzy klienci spełniają te reguły.

Nie sprawdzaj koszyka pod kątem predykcji Wybierz tę opcję, aby podczas oceny uwzględnić wszystkie reguły bez względu na obecność lub brak następników w koszyku.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz

powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.

- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Podsumowanie modelu użytkowego reguły asocjacyjnej

Na karcie Podsumowanie modelu użytkowego wyświetlana jest liczba wykrytych reguł oraz wartości minimum i maksimum pokrycia, wzrostu, ufności i wdrażalności reguł w zestawie reguł.

Generowanie zestawu reguł z powiązanego modelu użytkowego

Użytkowe modele asocjacyjne, takie jak Apriori i CARMA, mogą być używane do oceniania danych bezpośrednio lub można najpierw wygenerować dodatkowy zestaw reguł, znany jako **zestaw reguł**. Zestawy reguł są szczególnie przydatne podczas pracy z surowym modelem, którego nie można użyć bezpośrednio do oceniania. Więcej informacji można znaleźć w temacie “Modele surowe” na stronie 51.

Aby wygenerować zestaw reguł, należy wybrać opcję **Zestaw reguł** z menu Utwórz w przeglądarce modeli użytkowych. Można określić następujące opcje umożliwiające przekształcenie reguł na zestaw reguł:

Nazwa zestawu reguł. Umożliwia określenie nazwy nowo wygenerowanego węzła zestawu reguł.

Utwórz węzeł na. Decyduje o lokalizacji nowo wygenerowanego węzła zestawu reguł. Można wybrać **Obszar roboczy**, **Paleta modeli** lub **Łącznie**.

Zmienna przewidywana. Określa, jakie zmienne wyjściowe będą używane do wygenerowania węzła zestawu reguł. Należy wybrać jedną zmienną wyjściową z listy.

Minimalne wsparcie. Należy określić minimalne pokrycie reguł, od którego będą one zachowywane w wygenerowanym zestawie reguł. Reguły o pokryciu mniejszym od określonej wartości nie zostaną uwzględnione w nowym zestawie reguł.

Minimalna ufność. Należy określić minimalną ufność dla reguł, aby zostały zachowane w wygenerowanym zestawie reguł. Reguły, które będą miały niższą ufność od określonej wartości, nie zostaną uwzględnione w nowym zestawie reguł.

Wartość standardowa. Umożliwia określenie domyślnej wartości dla zmiennej przewidywanej, która jest przypisana do ocenianych rekordów, dla których żadna reguła nie ma zastosowania.

Generowanie modelu filtrowanego

Aby wygenerować model filtrowany z użytkowego modelu asocjacyjnego, takiego jak Apriori, CARMA lub węzeł zestawu reguł sekwencyjnych, należy wybrać opcję **Filtrowany model** z menu Utwórz w przeglądarce modeli użytkowych. Pozwoli to utworzyć model dodatkowego zestawu, który będzie zawierał tylko te reguły, które są aktualnie wyświetlane w przeglądarce. *Uwaga*: Nie można wygenerować modeli filtrowanych dla modeli surowych.

Dla reguł filtrowania można określić następujące opcje:

Nazwa nowego modelu. Umożliwia określenie nazwy nowego węzła modelu filtrowanego.

Utwórz węzeł na. Decyduje o lokalizacji nowego węzła modelu filtrowanego. Można wybrać **Obszar roboczy**, **Paleta modeli** lub **Łącznie**.

Numerowanie reguł. Umożliwia określenie sposobu numerowania identyfikatorów reguł w dodatkowym zestawie reguł zawartych w modelu filtrowanym.

- **Składuj oryginalne numery identyfikatorów reguł.** Opcję tę należy wybrać, aby zachowywana były oryginalna numeracja reguł. Domyślnie regułom nadawany jest identyfikator, który odpowiada ich kolejności wykrywania przez algorytm. Ta kolejność może różnić się w zależności od zastosowanego algorytmu.
- **Zmień numerowanie reguł kolejno rozpoczynając od.** Tę opcję należy wybrać, aby przypisać nowe identyfikatory do filtrowanych reguł. Nowe identyfikatory są przypisywane w oparciu o kolejność sortowania wyświetlaną w tabeli przeglądarki reguł na karcie Model, począwszy od określonego tutaj numeru. Numer początkowy dla identyfikatorów można określić za pomocą strzałek po prawej stronie.

Ocenianie reguł asocjacyjnych

Oceny wygenerowane w wyniku stosowania modeli użytkowych reguł asocjacyjnych do nowych danych są zwracane w osobnych zmiennych. Dla każdej predykcji dodawane są nowe zmienne, przy czym *P* oznacza predykcje, *C* oznacza ufność, a *I* oznacza identyfikator reguły. Organizacja tych zmiennych wynikowych zależy od tego, czy dane wejściowe mają format transakcyjny, czy tabelaryczny. Przegląd tych formatów zawiera sekcja “Dane tabelaryczne a dane transakcyjne” na stronie 258.

Załóżmy na przykład, że oceniamy dane koszyka przy użyciu modelu, który generuje predykcje na podstawie następujących trzech reguł:

```
Rule_15  bread&wine -> meat (ufność 54%)
Rule_22  cheese -> fruit (ufność 43%)
Rule_5  bread&cheese -> frozveg (ufność 24%)
```

Dane tabelaryczne. W danych tabelarycznych trzy predykcje (3 to liczba domyślna) będą zwracane w jednym rekordzie.

Tabela 16. Oceny w formacie tabelarycznym

Id.	Bread	Wine	Cheese	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	meat	0,54	15	fruit	0,43	22	frozveg	0,24	5

Dane transakcyjne. W przypadku danych transakcyjnych dla każdej predykcji generowany jest osobny rekord. Predykcje nadal dodawane są w osobnych kolumnach, ale oceny zwracane są w miarę obliczania. W efekcie generowane są rekordy z niekompletnymi predykcjami, tak jak przedstawiono to w poniższych przykładowych wynikach. W pierwszym rekordzie druga i trzecia predykcja (P2 i P3) jest pusta, podobnie jak skojarzone z nimi ufności i identyfikatory reguł. Jednak ostatni rekord, wygenerowany już po zwróceniu dalszych ocen, zawiera wszystkie trzy predykcje.

Tabela 17. Oceny w formacie transakcyjnym

Id.	Pozycja	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	bread	meat	0,54	14	$null$	$null$	$null$	$null$	$null$	$null$
Fred	cheese	meat	0,54	14	fruit	0,43	22	$null$	$null$	$null$
Fred	wine	meat	0,54	14	fruit	0,43	22	frozveg	0,24	5

Aby na potrzeby raportowania lub wdrożenia uwzględnić w wynikach tylko kompletne predykcje, należy użyć rekordu selekcji do wybrania kompletnych rekordów.

Uwaga: Nazwy zmiennych użyte w tych przykładach zostały skrócone dla poprawienia ich czytelności. W rzeczywistości zmienne wynikowe w modelach asocjacyjnych noszą nazwy podane w poniższej tabeli.

Tabela 18. Nazwy zmiennych wynikowych modeli asocjacyjnych

Nowa zmienna	Nazwa zmiennej w przykładzie
Predykcja	*$A-TRANSACTION_NUMBER-1*
Confidence (lub inne kryterium)	*$AC-TRANSACTION_NUMBER-1*

Tabela 18. Nazwy zmiennych wynikowych modeli asocjacyjnych (kontynuacja)

Nowa zmienna	Nazwa zmiennej w przykładzie
Identyfikator reguły	*$A-Rule_ID-1*

Reguły z wieloma następnikami

Algorytm CARMA dopuszcza reguły z więcej niż jednym następnikiem, na przykład:

```
bread -> wine&cheese
```

Podczas oceniania takich reguł z dwoma następnikami predykcje zwracane są w formacie przedstawionym w poniższej tabeli.

Tabela 19. Wyniki oceniania obejmujące predykcję z wieloma następnikami

Id.	Bread	Wine	Cheese	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	meat&veg	0,54	16	fruit	0,43	22	frozveg	0,24	5

W niektórych przypadkach konieczne jest podzielenie takich ocen przed wdrożeniem. Aby podzielić predykcję z wieloma następnikami, trzeba będzie przeanalizować zmienną za pomocą funkcji łańcuchowych języka CLEM.

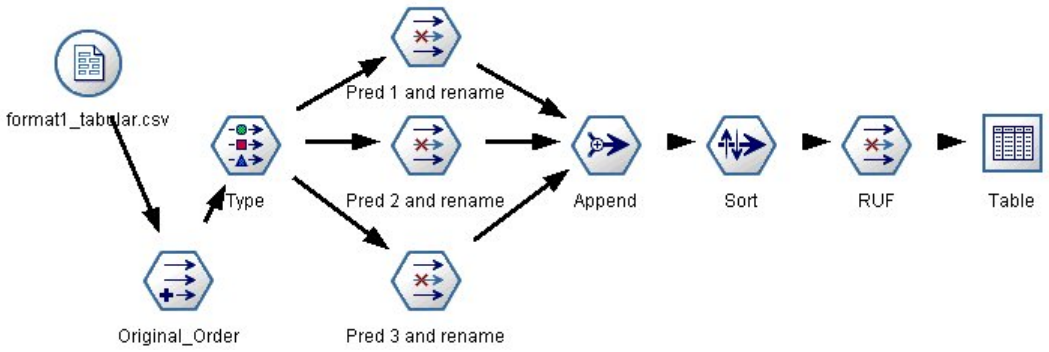
Wdrażanie modeli asocjacyjnych

Podczas oceniania modeli asocjacyjnych predykcje i ufności są generowane w osobnych kolumnach (gdzie *P* oznacza predykcję, *C* oznacza ufność, a *I* oznacza identyfikator reguły). Jest tak w przypadku, gdy dane wejściowe są tabelaryczne lub transakcyjne. Więcej informacji można znaleźć w temacie “Ocenianie reguł asocjacyjnych” na stronie 270.

Podczas przygotowywania ocen do wdrożenia może okazać się, że docelowa aplikacja wymaga przetransponowania danych wynikowych do formatu, w którym predykcje będą znajdowały się w wierszach, a nie w kolumnach (jedna predykcja na wiersz; ten format nazywany jest czasem formatem „kasowym” przez analogię do wydruku z kasy sklepowej).

Transpozycja ocen w tabelach

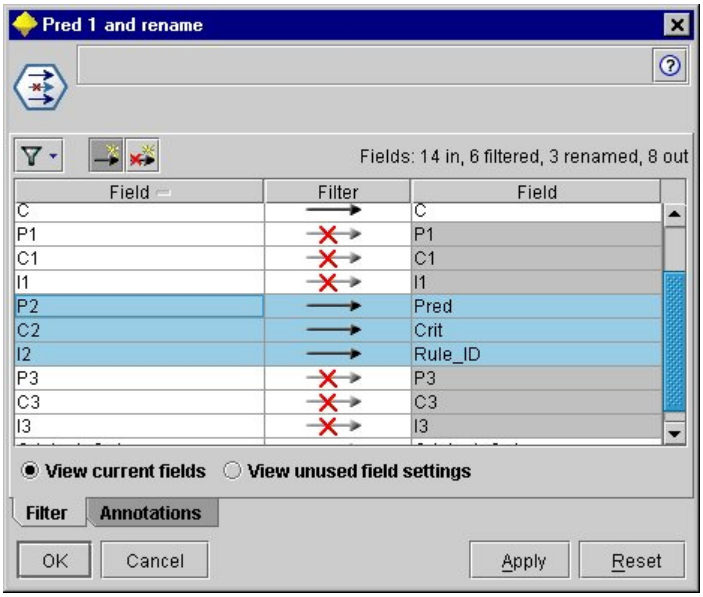
Oceny tabelaryczne można w programie IBM SPSS Modeler przetransponować z kolumn na wiersze, wykonując procedurę opisaną poniżej.



Rysunek 49. Przykładowy strumień służący do transponowania danych tabelarycznych do formatu kasowego

1. Użyj funkcji @INDEX w węźle wyliczenia, aby określić bieżącą kolejność predykcji i zapisz ten wskaźnik w nowej zmiennej, na przykład *Original_order*.
2. Dodaj węzeł Typ, aby zagwarantować, że wszystkie zmienne zostały zrealizowane jako instancje.

3. Użyj węzła filtrowania, aby zmienić nazwy zmiennych domyślnej predykcji, ufności i identyfikatora (*P1, C1, I1*) na bardziej czytelne, takie jak *Pred*, *Crit* i *Rule_ID*, które będą później używane do dodawania rekordów. Potrzebny będzie jeden węzeł filtrowania na każdą wygenerowaną predykcję.



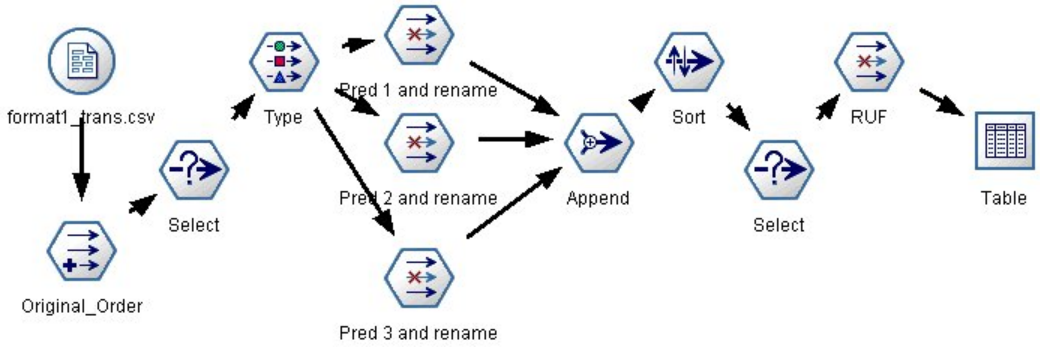
Rysunek 50. Filtrowanie zmiennych pod kątem predykcji 1 i 3 wraz ze zmianą nazw zmiennych dla predykcji 2.

4. Za pomocą węzła Dołączanie dołącz wartości dla wspólnych zmiennych *Pred*, *Crit* i *Rule_ID*.
5. Dołącz węzeł Sortowanie, aby posortować rekordy w porządku rosnącym dla zmiennej *Original_order* i w porządku malejącym dla zmiennej *Crit*, która jest zmienną używaną do sortowania predykcji wg takich kryteriów, jak ufność, wzrost i pokrycie.
6. Za pomocą kolejnego węzła filtrującego odfiltruj z wyników zmienną *Original_order*.

Na tym etapie dane są gotowe do wdrożenia.

Transpozycja ocen transakcyjnych

Proces transpozycji ocen transakcyjnych jest podobny. Na przykład strumień przedstawiony poniżej dokonuje transpozycji ocen do formatu z jedną predykcją w każdym wierszu.



Rysunek 51. Przykładowy strumień służący do transponowania danych transakcyjnych do formatu kasowego

Proces jest prawie identyczny z opisanym powyżej — jedyna różnica polega na dodaniu dwóch węzłów selekcji.

- Pierwszy węzeł selekcji służy do porównywania identyfikatorów reguł w sąsiednich rekordach i uwzględniania tylko rekordów unikalnych lub niezdefiniowanych. W tym węźle selekcji do wyboru rekordów używane jest wyrażenie języka CLEM: ID /= @OFFSET(ID,-1) or @OFFSET(ID,-1) = undef.
- Drugi węzeł selekcji służy od odrzucania dodatkowych reguł lub reguł, w których Rule_ID ma wartość null. W tym węźle selekcji do odrzucania rekordów używane jest następujące wyrażenie języka CLEM: not(@NULL(Rule_ID)).

Aby uzyskać więcej informacji o transpozycji ocen przed wdrożeniem, należy skontaktować się z działem wsparcia technicznego.

---

węzeł Sekwencje

Węzeł Sekwencje wykrywa wzorce w sekwencyjnych lub zorientowanych czasowo danych, w następującej postaci bread -> cheese. Elementy w kolejności to **zbiory elementów**, które tworzą pojedynczą transakcję. Na przykład, jeśli dana osoba idzie do sklepu i kupuje chleb i mleko, a następnie po kilku dniach wraca do sklepu i kupuje ser, aktywność zakupową tej osoby mogą reprezentować dwa zbiory elementów. Pierwszy zbiór elementów zawiera chleb i mleko, a drugi zawiera ser. **Sekwencja** to lista zbiorów elementów z tendencją do występowania w przewidywalnej kolejności. Węzeł Sekwencje wykrywa powtarzające się sekwencje i tworzy wygenerowany węzeł modelu, którego można użyć do predykcji.

Wymagania. W celu utworzenia zestawu reguł Sekwencje konieczne jest określenie zmiennej identyfikacyjnej, opcjonalnej zmiennej czasowej i jednej lub więcej zmiennych zawartości. Należy zwrócić uwagę, że te ustawienia należy wprowadzić na karcie Zmienne węzła modelowania; nie można ich odczytać z poprzedzającego węzła Typ. Zmienną identyfikacyjną może charakteryzować dowolna rola i dowolny poziom pomiaru. W przypadku określenia zmiennej czasu może mieć ona dowolną rolę, lecz jej składowaniem musi być wartość numeryczna, data, godzina lub znacznik czasu. W przypadku nieokreślenia zmiennej czasu węzeł Sekwencje będzie używał implikowanego znacznika czasowego, co oznacza używanie numerów wierszy jako wartości czasu. Zmienne zawartości mogą mieć dowolny poziom pomiaru i rolę, lecz wszystkie zmienne treści muszą być tego samego typu. Jeśli są one numeryczne, muszą być przedziałami całkowitoliczbowymi (nie przedziałami liczb rzeczywistych).

Mocne strony. Węzeł Sekwencje bazuje na algorytmie reguł asocjacyjnych CARMA, który korzysta z efektywnej metody dwu przejść do znajdowania sekwencji. Ponadto generowany węzeł modelu tworzony przez węzeł Sekwencje może zostać wstawiony do strumienia danych, co pozwoli na utworzenie predykcji. Generowany węzeł modelu umożliwia także generowanie superwęzłów umożliwiających wykrywanie i zliczanie określonych sekwencji oraz tworzenie predykcji w oparciu o określone sekwencje.

Opcje zmiennych węzła Sekwencje

Przed wykonaniem węzła Sekwencje konieczne jest określenie zmiennej identyfikacyjnej i zmiennej zawartości na karcie Zmienne węzła Sekwencje. W celu użycia zmiennej czasu musi ona również zostać tutaj zdefiniowana.

Zmienna identyfikacyjna. Należy wybrać zmienną identyfikacyjną z listy. Jako zmienna identyfikacyjna mogą być używane zmienne numeryczne lub symboliczne. Każda unikalna wartość tej zmiennej powinna wskazywać na określoną jednostkę analizy. Na przykład w aplikacji do obsługi koszyka zakupów każdy identyfikator może reprezentować jednego klienta. W przypadku aplikacji do analizy dzienników sieciowych każdy identyfikator może reprezentować komputer (wg adresu IP) lub użytkownika (wg danych logowania).

- **Wartości identyfikatorów są posortowane.** Jeśli dane zostały wstępnie posortowane tak, że wszystkie rekordy o tym samym identyfikatorze są zgrupowane w strumieniu danych, należy wybrać tę opcję w celu przyspieszenia przetwarzania. Jeśli dane nie zostały wstępnie posortowane (lub nie ma co do tego pewności), należy pozostawić tę opcję niezaznaczoną. Węzeł Sekwencje posortuje dane automatycznie.

Uwaga: Jeśli dane nie są posortowane, a użytkownik wybierze tę opcję, model Sekwencje może zwrócić niepoprawne wyniki.

Zmienna czasu. W celu użycia zmiennej w danych w celu określenia czasu zdarzeń należy zaznaczyć opcję **Użyj zmiennej czasu** i określić zmienną, która ma być używana. Zmienna czasu musi być wartością numeryczną, wartością

typu data, czas lub znacznik czasu. Jeśli zmienna czasu nie została określona, zakłada się, że rekordy są pobierane ze źródła danych kolejno, zaś jako wartości czasu używane są numery rekordów (pierwszy rekord ma miejsce o czasie „1”; drugi o czasie „2” itd).

Zmienne zawartości. Należy określić zmienne zawartości dla modelu. Zmienne te zawierają interesujące zdarzenia w procesie modelowania sekwencji.

Węzeł Sekwencje umożliwia obsługę danych w formacie tabelarycznym lub jako dane transakcyjne. W przypadku użycia wielu zmiennych z danymi transakcyjnymi zakłada się, że elementy określone w tych zmiennych dla konkretnego rekordu reprezentują elementy znajdujące się w pojedynczej transakcji i opatrzone pojedynczym znacznikiem czasu. Więcej informacji można znaleźć w temacie “Dane tabelaryczne a dane transakcyjne” na stronie 258.

Podział. To pole umożliwia określenie zmiennej używanej do podziału danych na osobne próby do uczenia, testowania i walidacji podczas budowania modelu. Korzystając z jednej próby do generowania modelu oraz innej do testowania go, można uzyskać wiarygodne informacje o tym, jak dobrze model pozwala uogólnić większe zbiory danych, podobne do danych bieżących. Jeśli korzystając z węzłów Typ lub Partycja, zdefiniowano wiele zmiennych dzielących na podzbiory, na karcie Zmienne każdego węzła modelowania korzystającego z tego podziału na podzbiory należy wybrać jedną zmienną dzielącą na podzbiory. (Jeśli obecna jest tylko jedna zmienna dzieląca na podzbiory, jest ona automatycznie używana po aktywowaniu dzielenia). Należy również pamiętać, że aby zastosować wybrany podział w analizie, dzielenie musi być również włączone na karcie Opcje modelu danego węzła. (Usunięcie zaznaczenia tej opcji umożliwia wyłączenie podziału bez zmiany ustawień zmiennych).

Opcje modelu węzła Sekwencje

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Minimalne pokrycie reguł (%) Istnieje możliwość określenia kryterium pokrycia. *Pokrycie reguł* odnosi się do proporcji identyfikatorów w danych uczących, które zawierają całą sekwencję. W celu skoncentrowania się na bardziej typowych sekwencjach należy zwiększyć wartość tego ustawienia.

Minimalna ufność reguły (%) Można określić kryterium ufności umożliwiające zachowanie sekwencji w zestawie sekwencji. *Ufność* odnosi się do wartości procentowej identyfikatorów, dla których tworzona jest poprawna predykcja, względem wszystkich identyfikatorów, dla których ta reguła powoduje utworzenie predykcji. Jest ona obliczana jako iloraz liczby identyfikatorów, dla których znajduje się cała sekwencja, i liczby identyfikatorów, dla których w oparciu o dane uczące znajdują się poprzedniki. Sekwencje o ufności niższej od zadanego kryterium są odrzucane. W przypadku uzyskiwania zbyt wielu sekwencji lub sekwencji mało interesujących, należy spróbować zwiększyć wartość tego ustawienia. W przypadku uzyskiwania zbyt niewielu sekwencji należy sprobować obniżyć wartość tego ustawienia.

Uwaga: W razie potrzeby można zaznaczyć wartość i zamiast niej wpisać własną wartość. Należy pamiętać, że zmniejszenie ufności poniżej 1,0 nie tylko spowoduje znaczne zwiększenie zapotrzebowania na pamięć, ale także potencjalne znaczące wydłużenie czasu budowania reguł.

Maksymalna wielkość sekwencji Użytkownik może ustawić maksymalną liczbę różnych elementów w sekwencji. Jeśli sekwencje są relatywnie krótkie, można zmniejszyć wartość tego ustawienia w celu przyspieszenia budowy zestawu sekwencji.

Predykcje do dodania do strumienia Umożliwia określenie liczby predykcji do dodania do strumienia przez wynikowy wygenerowany węzeł Model. Aby uzyskać więcej informacji, zobacz “Modele użytkowe sekwencji” na stronie 276.

Opcje zaawansowane węzła Sekwencje

Osobom ze szczegółową wiedzą na temat działania węzła Sekwencje poniższe opcje zaawansowane umożliwiają precyzyjne dostosowanie procesu budowy modelu. W celu uzyskania dostępu do opcji zaawansowanych należy ustawić opcję Tryb na wartość **Zaawansowany** na karcie Zaawansowany.

Maksymalny czas trwania. Zaznaczenie tej opcji powoduje, że sekwencje zostaną ograniczone do tych o czasie trwania (czasie między pierwszą a ostatnią składową) mniejszym lub równym zadanej wartości. Jeśli nie określono zmiennej czasu, czas trwania będzie wyrażony liczbą wierszy (rekordów) w danych nieprzetworzonych. Jeśli zmienna czasu ma format czasu, daty lub znacznika czasu, czas trwania jest wyrażony w sekundach. W przypadku zmiennych numerycznych czas trwania jest wyrażony w tych samych jednostkach, co sama zmienna.

Wartość przycinania. Algorytm CARMA używany w węźle Sekwencje okresowo usuwa (**przycina**) podczas przetwarzania rzadko występujące składowe z listy potencjalnych składowych, oszczędzając w ten sposób miejsce w pamięci. Tę opcję należy zaznaczyć, aby dostosować częstość przycinania. Podana liczba będzie określać częstość przycinania. Wprowadzenie mniejszej wartości powoduje obniżenie wymagań algorytmu co do ilości pamięci (lecz może także wydłużyć niezbędny czas uczenia), natomiast wprowadzenie większej wartości przyspieszy proces uczenia (lecz jednocześnie może spowodować zwiększenie ilości potrzebnej pamięci).

Maksymalna liczba sekwencji w pamięci. Po wybraniu tej opcji algorytm CARMA ograniczy wielkość pamięci przeznaczonej do przechowywania sekwencji kandydackich w trakcie budowania modelu do wskazanej liczby sekwencji. Należy wybrać tę opcję, jeśli program IBM SPSS Modeler wykorzystuje zbyt dużo pamięci podczas budowy modeli Sekwencje. Należy zwrócić uwagę, że maksymalna określona tutaj wartość sekwencji to liczba sekwencji kandydackich monitorowana wewnętrznie w miarę budowania modelu. Liczba ta powinna być większa od liczby sekwencji oczekiwanej w modelu finalnym.

Ogranicz odstępy między składowymi. Ta opcja umożliwia nałożenie ograniczeń co do odstępów w czasie oddzielających poszczególne składowe. Zaznaczenie tej opcji spowoduje, że składowe o odstępach mniejszych niż określony przez użytkownika **minimalny odstęp między składowymi** lub większych niż określony przez użytkownika **maksymalny odstęp między składowymi** nie będą uwzględniane podczas tworzenia fragmentu sekwencji. Użycie tej opcji pozwala uniknąć sekwencji zliczania obejmujących długie przedziały czasowe lub tych mających miejsce w bardzo krótkim przedziale czasowym.

Uwaga: Jeśli zmienna czasu ma format czasu, daty lub znacznika czasu, czas trwania jest wyrażony w sekundach. W przypadku zmiennych numerycznych odstęp jest wyrażony w tych samych jednostkach, co sama zmienna czasu.

Należy na przykład rozważyć następującą listę transakcji.

Tabela 20. Przykładowa lista transakcji

Id.	Czas	Zawartość
1001	1	apples
1001	2	bread
1001	5	cheese
1001	6	dressing

W przypadku zbudowania na podstawie tych danych modelu o minimalnym odstępie wynoszącym 2 otrzymamy następujące sekwencje:

```
apples -> cheese

apples -> dressing

bread -> cheese

bread -> dressing
```

Nie zostaną wyświetlone sekwencje takie jak apples -> bread, ponieważ odstęp między składowymi apples i bread jest mniejszy niż odstęp minimalny. W podobny sposób można przeanalizować poniższy, alternatywny przykład.

Tabela 21. Przykładowa lista transakcji

Id.	Czas	Zawartość
1001	1	apples
1001	2	bread
1001	5	cheese
1001	20	dressing

Jeśli jako odstęp maksymalny ustawiono 10, nie zostanie wyświetlona żadna sekwencja zawierająca termin dressing, ponieważ odstęp między składowymi cheese i dressing będzie zbyt duży, aby mogły one zostać uwzględnione jako fragment tej samej sekwencji.

Modele użytkowe sekwencji

Modele użytkowe sekwencji reprezentują sekwencje znajdowane dla określonej zmiennej wyjściowej przez węzeł Sekwencje i mogą być dodawane do strumieni z myślą o generowaniu predykcji.

Po uruchomieniu strumienia zawierającego węzeł Sekwencje węzeł dodaje parę zmiennych zawierających predykcje i powiązane wartości ufności dla każdej predykcji z modelu sekwencji do danych. Domyślnie dodawane są trzy pary zmiennych zawierających trzy najlepsze predykcje (oraz powiązane z nimi wartości ufności). Liczbę generowanych predykcji można zmienić podczas budowy modelu, ustawiając opcje modelu węzła Sekwencje w czasie budowy, a także na karcie Ustawienia, po dodaniu modelu użytkowego do strumienia. Więcej informacji można znaleźć w temacie “Ustawienia modelu użytkowego sekwencji” na stronie 279.

Nazwy nowych zmiennych stanowią pochodne nazwy modeli. Nazwy zmiennych mają postać *$S-sequence-n* w przypadku zmiennej predykcyjnej (gdzie *n* oznacza *n*-tą predykcję) oraz *$SC-sequence-n* w przypadku zmiennej ufności. W strumieniu z wieloma węzłami reguł sekwencji nowe nazwy zmiennych będą opatrzone przedrostkami liczbowymi odróżniającym je od siebie. W pierwszym węźle w strumieniu będą używane zwykłe nazwy, w drugim nazwy rozpoczynające się od *$S1-* i *$SC1-*, zaś w trzecim nazwy rozpoczynające się od *$S2-* i *$SC2-* itd. Predykcje są wyświetlane w kolejności wg ufności, tak że *$S-sequence-1* zawiera predykcję o najwyższej ufności, *$S-sequence-2* zawiera predykcję o kolejnej najwyższej ufności itd. W przypadku rekordów, w których liczba dostępnych predykcji jest mniejsza od liczby predykcji wymaganych, pozostałe predykcje zawierają wartość $null$. Na przykład, jeśli dla konkretnego rekordu można utworzyć tylko dwie predykcje, wartości *$S-sequence-3* i *$SC-sequence-3* będą równe $null$.

Dla każdego rekordu reguły w modelu są porównywane z zestawem transakcji przetworzonych dotąd dla bieżącego identyfikatora, w tym dla bieżącego rekordu oraz wszelkich poprzednich rekordów o tym samym identyfikatorze i wcześniejszym znaczniku czasu. Reguły *k* o wartościach najwyższej ufności, mające zastosowanie do tego zestawu transakcji, są używane do generowania predykcji *k* dla tego rekordu, przy czym *k* jest liczbą predykcji określonych na karcie Ustawienia po dodaniu modelu do strumienia. (Jeśli wiele reguł pozwala przewidzieć ten sam wynik dla zestawu transakcji, używana jest tylko reguła o najwyższej ufności). Więcej informacji można znaleźć w temacie “Ustawienia modelu użytkowego sekwencji” na stronie 279.

Podobnie jak w przypadku pozostałych typów modeli reguł asocjacyjnych, format danych musi odpowiadać formatowi używanemu do budowy modelu sekwencji. Na przykład modele budowane z użyciem danych tabelarycznych mogą być używane do oceniania tylko danych tabelarycznych. Więcej informacji można znaleźć w temacie “Ocenianie reguł asocjacyjnych” na stronie 270.

Uwaga: Podczas oceniania danych z użyciem wygenerowanego węzła zestawu sekwencji w strumieniu wszelkie ustawienia tolerancji czy odstępów wybrane podczas budowania modelu są ignorowane.

Predykcje na podstawie reguł sekwencji

Węzeł obsługuje rekordy w sposób zależny od czasu (lub w sposób zależny od kolejności, jeśli do budowy modelu nie użyto znacznika czasu). Rekordy należy sortować wg zmiennej identyfikacyjnej i zmiennej znacznika czasu (jeśli istnieje). Predykcje nie są jednak powiązane z rekordem, do którego są dodawane. Odnoszą się one po prostu do elementów charakteryzujących się najwyższym prawdopodobieństwem wystąpienia *w pewnym momencie w przyszłości*, biorąc pod uwagę historię transakcji dla bieżącego identyfikatora aż do bieżącego rekordu.

Należy zwrócić uwagę, że predykcje dla każdego rekordu niekoniecznie zależą od transakcji tego rekordu. Jeśli transakcje dla bieżącego rekordu nie wyzwalają określonej reguły, reguły zostaną wybrane w oparciu o poprzednie transakcje dla bieżącego identyfikatora. Innymi słowy, jeśli bieżący rekord nie uzupełnia sekwencji o żadne użyteczne informacje predykcyjne, predykcja od ostatniej użytecznej transakcji dla tego identyfikatora ID jest przenoszona do bieżącego rekordu.

Załóżmy na przykład, że mamy model Sekwencje z jedną regułą

```
Jam -> Bread (0.66)
```

i przekazujemy do niego następujące rekordy.

Tabela 22. Rekordy przykładowe

Id.	Zakup	Predykcja
001	dżem	bread
001	mleko	bread

Zwróćmy uwagę, że pierwszy rekord generuje predykcję *bread*, tak jak można tego oczekiwać. Drugi rekord również zawiera predykcję *bread*, ponieważ brak jest reguły dla *jam*, po którym następuje *milk*; stąd transakcja *milk* nie dodaje żadnych użytecznych informacji, a reguła Jam -> Bread nadal obowiązuje.

Generowanie nowych węzłów

Menu Utwórz umożliwia tworzenie nowych superwęzłów w oparciu o model sekwencji.

- **Superwęzeł reguły.** Tworzy superwęzeł umożliwiający wykrywanie i zliczanie wystąpień sekwencji w ocenianych danych. Ta opcja jest wyłączona, jeśli nie wybrano reguły. Więcej informacji można znaleźć w temacie “Generowanie superwęzła reguł z modelu użytkowego sekwencji” na stronie 279.
- **Model do palety.** Zwraca model na paletę Modele. Opcja ta jest przydatna w sytuacjach, gdy kolega/koleżanka wysłał(a) użytkownikowi strumień zawierający model, a nie sam model.

Szczegóły modelu użytkowego sekwencji

Karta Model dla modelu Sekwencje prezentuje reguły wyodrębnione przez algorytm. Każdy wiersz w tabeli reprezentuje regułę, przy czym poprzednik (część „if” reguły) znajduje się w pierwszej kolumnie, oraz następnik (część „then” reguły) w drugiej kolumnie.

Każda reguła jest wyświetlana w następującym formacie.

Tabela 23. Format reguły

Poprzednik	Następnik
beer i cannedveg	beer
fish fish	fish

Reguła z pierwszego przykładu jest interpretowana następująco: *dla identyfikatorów zawierających „beer” i „cannedveg” w tej samej transakcji istnieje wysokie prawdopodobieństwo, że następnym wystąpieniem będzie „beer”.* Reguła z drugiego przykładu może być interpretowana następująco: *dla identyfikatorów zawierających „fish” w jednej transakcji oraz „fish” w drugiej istnieje wysokie prawdopodobieństwo wystąpienia po raz kolejny „fish”.* Należy

zwrócić uwagę, że w przypadku pierwszej reguły *beer* i *cannedveg* są nabywane jednocześnie; w przypadku drugiej reguły zakup *fish* następuje dwukrotnie, w dwu osobnych transakcjach.

Menu Sortowanie. Przycisk menu Sortowanie na pasku narzędzi steruje regułami sortowania. Kierunek sortowania (rosnąco/malejąco) można zmienić za pomocą przycisku kierunku sortowania (strzałka w górę lub w dół).

Reguły można sortować wg następujących pól:

- Pokrycie %
- Ufność %
- Pokrycie reguł %
- Następnik
- Pierwszy poprzednik
- Ostatni poprzednik
- Liczba elementów (poprzedników)

Na przykład poniższa tabela jest sortowana w kolejności malejącej według liczby elementów. Reguły z wieloma elementami w zestawie poprzedników poprzedzają te o mniejszej liczbie elementów.

Tabela 24. Reguły posortowane wg liczby elementów

Poprzednik	Następnik
beer i cannedveg i frozenmeal	frozenmeal
beer i cannedveg	beer
fish fish	fish
softdrink	softdrink

Pokaż/Ukryj menu kryteriów. Przycisk Pokaż/Ukryj menu kryteriów (ikona siatki) steruje opcjami wyświetlania reguł. Dostępne są następujące opcje wyświetlania:

- **Instancje** to opcja prezentująca informacje na temat liczby unikalnych identyfikatorów, dla których występuje *pełna sekwencja* — zarówno poprzedniki, jak i następnik. (Należy zwrócić uwagę, że ta opcja różni się względem modeli asocjacyjnych, w przypadku których liczba instancji odnosi się do liczby identyfikatorów, w przypadku których mają zastosowanie *tylko* poprzedniki). Na przykład w przypadku reguły bread -> cheese liczba identyfikatorów w danych uczących, obejmujących zarówno element *bread*, jak i*cheese*, jest określana jako **instancje**.
- **Pokrycie** wyświetla proporcję identyfikatorów w danych uczących, dla których poprzedniki mają wartość prawda. Na przykład, jeśli 50% danych uczących obejmuje poprzednik *bread*, wówczas pokrycie dla reguły bread -> cheese będzie wynosić 50%. (Jak już wcześniej wspomniano, inaczej niż w przypadku modeli asocjacyjnych, pokrycie *nie* bazuje na liczbie instancji).
- **Ufność** odnosi się do wartości procentowej identyfikatorów, dla których tworzona jest poprawna predykcja, względem wszystkich identyfikatorów, dla których ta reguła powoduje utworzenie predykcji. Jest ona obliczana jako iloraz liczby identyfikatorów, dla których znajduje się cała sekwencja, i liczby identyfikatorów, dla których w oparciu o dane uczące znajdują się poprzedniki. Na przykład, jeśli 50% danych uczących zawiera element cannedveg (co oznacza pokrycie poprzednika), lecz tylko 20% zawiera zarówno element cannedveg, jak ifrozenmeal, wówczas ufność dla reguły cannedveg -> frozenmeal będzie równa ilorazowi: Pokrycie reguł / Pokrycie poprzedników lub, w tym przypadku, 40%.
- **Pokrycie reguł** dla modeli Sekwencje bazuje na instancjach i powoduje wyświetlanie proporcji rekordów uczących, dla których cała reguła — poprzednik(i) i następnik(i) — jest prawdziwa. Na przykład, jeśli 20% danych uczących zawiera zarówno element *bread*, jak i *cheese*, wówczas pokrycie reguły dla reguły bread -> cheese wynosi 20%.

Należy zwrócić uwagę, że proporcje bazują na poprawnych transakcjach (transakcjach z co najmniej jednym obserwowanym elementem lub wartością prawda), nie zaś na łącznej liczbie transakcji. Transakcje niepoprawne — te bez elementów i wartości prawda — są w przypadku tych obliczeń odrzucane.

Przycisk Filtr. Przycisk Filtr (ikona lejka) w menu umożliwia rozwinięcie dolnej części okna dialogowego, zawierającej panel, w którym wyświetlane są filtry aktywnych reguł. Filtry służą do zawężenia liczby reguł wyświetlanych na karcie Modele.

Rysunek 52. Przycisk Filtr

Aby utworzyć filtr, kliknij ikonę Filtr po prawej stronie rozwiniętego panelu. Spowoduje to otwarcie osobnego okna dialogowego, w którym można określić ograniczenia co do wyświetlania reguł. Należy zwrócić uwagę, że przycisk Filtr jest często używany w połączeniu z menu Utwórz, w pierwszej kolejności do filtrowania reguł, a następnie do generowania modelu zawierającego ten podzbiór reguł. Aby uzyskać więcej informacji, patrz “Określanie filtrów dla reguł” na stronie 266 poniżej.

Ustawienia modelu użytkowego sekwencji

Karta Ustawienia dla modelu użytkowego Sekwencje zawiera opcje oceniania dla modelu. Ta karta jest dostępna tylko po dodaniu modelu do obszaru roboczego strumienia na potrzeby oceniania.

Maksymalna liczba predykcji. Określa maksymalną liczbę predykcji w każdym zestawie elementów w koszyku. Reguły o najwyższych wartościach ufności mające zastosowanie do tego zestawu transakcji są używane do generowania predykcji dla tego rekordu, aż do zadanego limitu.

Podsumowanie modelu użytkowego sekwencji

Karta Podsumowanie dla modelu użytkowego reguł sekwencji przedstawia liczbę wykrytych reguł oraz wartości minimum i maksimum dla pokrycia i ufności dla reguł. Jeśli wykonano węzeł analizy dołączony do tego węzła modelowania, informacje z tej analizy również będą wyświetlane w tej sekcji.

Więcej informacji można znaleźć w temacie “Przeglądanie modeli użytkowych” na stronie 42.

Generowanie superwęzła reguł z modelu użytkowego sekwencji

Aby wygenerować superwęzeł reguł na podstawie reguł sekwencyjnych:

1. Na karcie Model dla modelu użytkowego reguł sekwencyjnych kliknij wiersz w tabeli, aby wybrać żądaną regułę.
2. Z menu przeglądarki reguł wybierz opcje:

   Utwórz > Superwęzeł reguł

Ważne: Aby użyć wygenerowanego superwęzła, należy posortować dane według zmiennej ID (i zmiennej czasu, jeśli taka istnieje) przed przekazaniem ich do superwęzła. Superwęzeł nie wykryje poprawnej sekwencji w danych, które nie są posortowane.

Dla generowania superwęzła reguł można określić następujące opcje:

Wykryj. Określa, jak definiowane są dopasowania dla danych przekazanych do superwęzła.

- **Wyłącznie poprzedniki.** Superwęzeł będzie określał dopasowanie za każdym razem, kiedy znajdzie poprzedniki dla wybranej reguły w poprawnej kolejności w zbiorze rekordów z takim samym identyfikatorem, niezależnie od tego, czy znaleziony również zostanie następnik. Należy pamiętać, że nie jest tu uwzględniana tolerancja znacznika czasu ani ustawienia ograniczeń odstępu między składowymi z oryginalnego węzła modelowania sekwencji. Jeśli w strumieniu wykryty zostanie zestaw ostatniego poprzednika (a wszystkie pozostałe poprzedniki zostały już znalezione we właściwej kolejności), wszystkie kolejne rekordy z bieżącym identyfikatorem będą zawierały wybrane poniżej podsumowanie.
- **Cała sekwencja.** Superwęzeł będzie określał dopasowanie za każdym razem, kiedy znajdzie poprzedniki i następnika dla wybranej reguły we właściwej kolejności w zbiorze rekordów o takim samym identyfikatorze. Nie jest tu uwzględniana tolerancja znacznika czasu ani ustawienia ograniczeń odstępu między składowymi z oryginalnego węzła modelowania sekwencji. Jeśli w strumieniu zostanie wykryty następnik (a wszystkie

poprzedniki zostały również znalezione we właściwej kolejności), bieżący rekord i wszystkie kolejne rekordy z bieżącym identyfikatorem będą zawierały wybrane poniżej podsumowanie.

Pokaż. Decyduje, w jaki sposób podsumowania dopasowań są dodawane do wyniku superwęzła reguły.

- **Wartość następnika dla pierwszego wystąpienia.** Wartość dodawana do danych jest następną wartością predykcji ustalaną w oparciu o pierwsze wystąpienie dopasowania. Wartości są dodawane jako nowa zmienna o nazwie *rule_n_consequent*, gdzie *n* oznacza numer reguły (ustalany w oparciu o kolejność tworzenia superwęzłów reguły w strumieniu).
- **Wartość prawdy dla pierwszego wystąpienia.** Wartość dodawana do danych jest prawdziwa, jeśli istnieje co najmniej jedno dopasowanie dla identyfikatora, a fałszywa, jeśli nie ma żadnych dopasowań. Wartości są dodawane jako nowa zmienna o nazwie *rule_n_flag*.
- **Liczebność wystąpień.** Wartość dodawana do danych jest liczbą dopasowań dla identyfikatora. Wartości są dodawane jako nowa zmienna o nazwie *rule_n_count*.
- **Numer reguły.** Dodawana wartość jest numerem dla wybranej reguły. **Numery reguł** są przypisywane w oparciu o kolejność, w jakiej superwęzeł był dodawany do strumienia. Na przykład, pierwszy superwęzeł reguły jest nazywany *rule 1*, drugi superwęzeł reguły jest nazywany *rule 2* itd. Ta opcja jest najbardziej przydatna, jeśli w strumieniu uwzględnianych będzie wiele superwęzłów reguł. Wartości są dodawane jako nowa zmienna o nazwie *rule_n_number*.
- **Dołącz wartości ufności.** Jeśli ta opcja zostanie zaznaczona, do strumienia danych będzie dodawana ufność reguły oraz wybrane podsumowanie. Wartości są dodawane jako nowa zmienna o nazwie *rule_n_confidence*.

---

reguły asocjacyjne, węzeł

Reguły asocjacyjne są instrukcjami w następującej postaci.

Na przykład „Jeśli klient kupi maszynkę i płyn po goleniu, wówczas ten klient kupi krem do golenia z 80% pewnością”. Węzeł Reguły asocjacyjne pozwala wyodrębnić zestaw reguł na podstawie danych, pobierając reguły o najwyższej możliwej zawartości informacji. Węzeł Reguły asocjacyjne jest bardzo podobny do węzła Apriori, jednak istnieją pewne istotne różnice:

- Węzeł Reguły asocjacyjne nie może przetwarzać danych transakcyjnych.
- Węzeł Reguły asocjacyjne nie może przetwarzać danych typu Składowanie listy i na poziomie pomiaru zbioru.
- Węzeł Reguły asocjacyjne może być używany z produktem IBM SPSS Analytic Server. Dzięki temu możliwa jest skalowalność i oznacza to, że możliwe jest przetwarzanie wielkich zbiorów danych oraz korzystanie z szybszego przetwarzania równoległego.
- Węzeł Reguły asocjacyjne udostępnia dodatkowe ustawienia, takie jak możliwość ograniczenia liczby generowanych reguł, co prowadzi do zwiększenia szybkości przetwarzania.
- Wynik z modelu użytkowego jest przedstawiony w przeglądarce wyników.

Uwaga: Węzeł Reguły asocjacyjne nie obsługuje kroków Ocena modelu ani Champion Challenger w produkcie IBM SPSS Collaboration and Deployment Services.

Uwaga: Węzeł Reguły asocjacyjne ignoruje puste rekordy podczas budowania modelu, jeśli zmienna jest typu flaga. Puste rekordy to rekordy, w których wszystkie zmienne używane w budowie modelu mają wartość fałszywą.

Strumień przedstawiający roboczy przykład używania reguł asocjacyjnych, posiadający nazwę geospatial_association.str i odwołujący się do plików danych InsuranceData.sav, CountyData.sav i ChicagoAreaCounties.shp, jest dostępny w katalogu Demos instalacji produktu IBM SPSS Modeler. Dostęp do katalogu Demos można uzyskać z grupy programu IBM SPSS Modeler w menu Start systemu Windows. Plik geospatial_association.str znajduje się w katalogu streams.

Reguły asocjacyjne — opcje zmiennych

Na karcie **Zmienne** użytkownik może wybrać, czy używane będą ustawienia roli zmiennej zdefiniowane we wcześniejszych węzłach, np. we wcześniejszym węźle Typ, czy przypisania zmiennych zostaną utworzone ręcznie.

Użyj wstępnie zdefiniowanych ról
W przypadku tej opcji używane są ustawienia roli (np. zmienne przewidywane lub predyktory) z poprzedzającego węzła Typ (albo z karty Typy poprzedzającego węzła źródłowego). Zmienne o roli zmiennych wejściowych są traktowane jako Warunki, a zmienne o roli zmiennych przewidywanych są traktowane jako Predykcje. Natomiast zmienne używane jako wejściowe i przewidywane są traktowane jak zmienne pełniące obie te role.

Użyj niestandardowych przypisań
Opcję tę należy wybrać w celu ręcznego przypisania zmiennych przewidywanych, predyktorów oraz innych ról na tym ekranie.

Zmienne
Jeśli wybierzesz opcję **Użyj niestandardowych przypisań**, wówczas użyj przycisków ze strzałkami, aby ręcznie przypisać elementy z tej listy do pól po prawej stronie ekranu. Ikony wskazują prawidłowe poziomy pomiaru dla każdej zmiennej.

Oba (Warunek i Predykcja)
Zmienne dodane do tej listy mogą przyjmować rolę warunku lub predykcji w regułach generowanych przez model. Ten proces odbywa się osobno dla poszczególnych reguł, dlatego zmienna może być warunkiem w jednej roli i predykcją w innej.

Tylko predykcja
Zmienne dodane do tej listy mogą być tylko predykcjami (co jest również określane jako „następnik") w regule. Obecność zmiennej na tej liście nie oznacza, że zmienna jest używana w jakichkolwiek regułach — oznacza tylko, że jeśli zmienna jest używana, wówczas może być tylko predykcją.

Tylko warunek
Zmienne dodane do tej listy mogą być tylko warunkami (co jest również określane jako „poprzednik") w regule. Obecność zmiennej na tej liście nie oznacza, że zmienna jest używana w jakichkolwiek regułach — oznacza tylko, że jeśli zmienna jest używana, wówczas może być tylko warunkiem.

Reguły asocjacyjne — budowanie reguły

Elementy dla reguł

Te opcje umożliwiają określenie liczby elementów lub wartości, które mogą być używane w każdej regule.

Uwaga: Suma wartości tych dwóch zmiennych nie może przekroczyć 10.

Maksymalna liczba warunków
Wybierz maksymalną liczbę warunków, które mogą być uwzględnione w jednej regule.

Maksymalna liczba predykcji
Wybierz maksymalną liczbę predykcji, które mogą być uwzględnione w jednej regule.

Budowanie reguły

Te opcje umożliwiają określanie liczby i typu reguł do zbudowania.

Maksymalna liczba reguł
Określ maksymalną liczbę reguł, które mogą być uwzględnione do użycia podczas budowania reguł dla Twojego modelu.

Kryterium reguły dla N najważniejszych
Wybierz kryterium używane w celu ustalenia najważniejszych N reguł, gdzie N jest wartością wprowadzaną do zmiennej **Maksymalna liczba reguł**. Do wyboru dostępne są następujące kryteria.

- **Ufność**
- **Pokrycie reguł**
- **Pokrycie warunków**
- **Wzrost**

- **Wdrażalność**

Tylko wartości true dla flag

Gdy dane są dostępne w formacie tabelarycznym, wybierz tę opcję, aby dla zmiennych flagi uwzględniać tylko wartości true w regułach wynikowych. Wybranie wartości true ułatwia zrozumienie reguł. Ta opcja nie ma zastosowania do danych w formacie transakcyjnym. Aby uzyskać więcej informacji, zobacz “Dane tabelaryczne a dane transakcyjne” na stronie 258.

Kryterium reguły

Jeśli wybierzesz opcję **Włącz kryterium reguły**, możesz użyć tych opcji w celu wyboru minimalnej siły, jaką muszą mieć reguły uwzględniane w konkretnym modelu.

- **Ufność** Określ minimalną wartość procentową poziomu ufności dla reguły generowanej przez model. Jeśli ten model generuje regułę o poziomie niższym niż ta wartość, reguła jest odrzucana.
- **Pokrycie reguł** Określ minimalną wartość procentową poziomu pokrycia reguł dla reguły generowanej przez model. Jeśli ten model generuje regułę o poziomie niższym niż ta wartość, reguła jest odrzucana.
- **Pokrycie warunków** Określ minimalną wartość procentową poziomu pokrycia warunku dla reguły generowanej przez model. Jeśli ten model generuje regułę o poziomie niższym niż ta wartość, reguła jest odrzucana.
- **Wzrost** Określ minimalną wartość wzrostu, jaka jest dozwolona dla reguły generowanej przez model. Jeśli ten model generuje regułę o wartości niższej niż ta wartość, reguła jest odrzucana.

Wyklucz reguły

W niektórych przypadkach związek między dwiema zmiennymi lub większą liczbą zmiennych jest znany lub łatwo zauważalny — w takim przypadku można wykluczyć reguły, w których zmienne przewidują się wzajemnie. Wykluczenie reguł zawierających obie wartości powoduje ograniczenie ilości zbędnych danych wejściowych i zwiększa prawdopodobieństwo znalezienia użytecznych wyników.

Zmienne

Wybierz skojarzone zmienne, których nie zamierzasz używać razem podczas budowania reguły. Na przykład skojarzone zmienne mogą być następujące: Producent samochodów i Model samochodu albo Rok szkolny i Wiek ucznia. Gdy model tworzy regułę, a reguła zawiera co najmniej jedną zmienną wybraną po dowolnej stronie reguły (warunek lub predykcja), reguła jest odrzucana.

Reguły asocjacyjne — transformacje

Grupowanie w przedziałach

Te opcje umożliwiają określenie sposobu grupowania w przedziałach zmiennych ilościowych (przedziałów liczb).

Liczba przedziałów

Dowolne zmienne ilościowe ustawione jako automatycznie grupowane w przedziałach są dzielone na pewną liczbę równo rozmieszczonych kategorii, które określa użytkownik. Można wybrać dowolną liczbę z przedziału od 2 do 10.

Zmienne listy

Maksymalna długość listy

Aby ograniczyć liczbę elementów do uwzględnienia w modelu, gdy długość zmiennych listy jest nieznana, należy wprowadzić maksymalną długość listy. Można wybrać dowolną liczbę z przedziału od 1 do 100. Jeśli długość listy jest większa niż wprowadzona liczba, model użyje zmiennej, ale uwzględni tylko wartości do tej liczby; wszelkie wartości dodatkowe w zmiennej będą ignorowane.

Reguły asocjacyjne — wyniki

Opcje dostępne w tym panelu umożliwiają kontrolowanie tego, jakie dane wynikowe są generowane podczas budowania modelu.

Tabele reguł

Poniższe opcje umożliwiają tworzenie tabel co najmniej jednego typu, które przedstawiają najlepszą liczbę reguł (na podstawie określonej liczby) dla każdego wybranego kryterium.

Ufność Ufność to stosunek pokrycia reguły do pokrycia warunku. Spośród elementów z podanymi wartościami warunku jest to procent elementów, które zawierają przewidywane wartości następnika. Tworzy tabelę zawierającą N najlepszych reguł asocjacyjnych, które są oparte na ufności uwzględnianej w wynikach (gdzie N jest wartością **Reguły do wyświetlenia**).

Pokrycie reguł
Odsetek elementów, dla których spełniona jest cała reguła, spełnione są warunki oraz predykcje. W przypadku wszystkich wartości w zbiorze danych jest to procent, który jest poprawnie uwzględniany i przewidywany przez regułę. Ta miara przedstawia ogólną ważność reguły. Tworzy tabelę zawierającą N najlepszych reguł asocjacyjnych, które są oparte na pokryciu reguł uwzględnianym w wynikach (gdzie N jest wartością **Reguły do wyświetlenia**).

Wzrost
Stosunek ufności reguły i wstępnego prawdopodobieństwa posiadania predykcji. Stosunek wartości ufności dla reguły w porównaniu do procentu występowania wartości następnika w całej populacji. Ten stosunek przedstawia miarę tego, w jakim stopniu reguła ulega poprawie w zależności od prawdopodobieństwa. Tworzy tabelę zawierającą N najlepszych reguł asocjacyjnych, które są oparte na wzroście uwzględnianym w wynikach (gdzie N jest wartością **Reguły do wyświetlenia**).

Pokrycie warunków
Odsetek elementów, dla których warunki są spełnione. Tworzy tabelę zawierającą N najlepszych reguł asocjacyjnych, które są oparte na pokryciu poprzedników uwzględnianym w wynikach (gdzie N jest wartością **Reguły do wyświetlenia**).

Wdrażalność
Miara tego, jaki procent danych uczących spełnia warunek, ale nie spełnia predykcji. Ta miara pokazuje, jak często konkretna reguła nie jest spełniona. Jest to w rzeczywistości przeciwieństwo ufności. Tworzy tabelę zawierającą N najlepszych reguł asocjacyjnych, które są oparte na wdrażalności uwzględnianej w wynikach (N jest wartością **Reguły do wyświetlenia**).

Reguły do wyświetlenia
Umożliwia ustawienie maksymalnej liczby reguł do wyświetlenia w tabelach.

Tabele informacji o modelu

W celu wybrania tabel modelu do uwzględnienia w wynikach należy użyć co najmniej jednej z tych opcji.

- **Transformacje zmiennych**
- **Podsumowanie rekordów**
- **Statystyki reguł**
- **Najczęstsze wartości**
- **Najczęstsze zmienne**

Sortowalna chmura słów reguły

Użyj tych opcji w celu utworzenia chmury słów, która przedstawia wyniki reguły. Słowa są wyświetlane ze zwiększającym się rozmiarem czcionki w celu wskazania ich ważności.

Utwórz sortowalną chmurę słów.
Zaznacz to pole, aby w wynikach utworzyć sortowalną chmurę słów.

Domyślny porządek sortowania
Wybierz typ sortowania, który będzie stosowany po zainicjowaniu tworzenia chmury słów. Chmura słów jest interaktywna, a w celu wyświetlenia innych reguł i sortowań można zmienić kryterium w przeglądarce modelu. Do wyboru dostępne są następujące opcje sortowania:

- Ufność
- Pokrycie reguł
- Wzrost
- Pokrycie warunków
- Wdrażalność

Maksimum reguł do wyświetlenia
Ustaw liczbę reguł do wyświetlenia w chmurze słów; maksymalnie można wybrać 20.

Reguły asocjacyjne — opcje modelu

Za pomocą opcji dostępnych na tej karcie można określić opcje oceny dla modeli reguł asocjacyjnych.

Nazwa modelu Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Maksymalna liczba predykcji Określ maksymalną liczbę predykcji, które zostaną uwzględnione w wyniku oceny. Ta opcja jest używana z pozycjami **Kryterium reguły** w celu wygenerowania „najważniejszych" predykcji, w których „najważniejsza" oznacza najwyższy poziom ufności, pokrycia, wzrostu itp.

Kryterium reguły Wybierz miarę, która będzie używana w celu określenia siły reguł. Reguły są sortowane wg siły kryteriów wybranych w tej opcji w celu uzyskania najważniejszych predykcji dla zbioru elementów. Do wyboru dostępnych jest 5 różnych kryteriów.

- **Ufność** Ufność to stosunek pokrycia reguły do pokrycia warunku. Spośród elementów z podanymi wartościami warunku jest to procent elementów, które zawierają przewidywane wartości następnika.
- **Pokrycie warunków** Odsetek elementów, dla których warunki są spełnione.
- **Pokrycie reguł** Odsetek elementów, dla których spełniona jest cała reguła, spełnione są warunki oraz predykcje. W celu obliczenia wartość **Pokrycie warunków** jest mnożona przez wartość **Ufność**.
- **Wzrost** Stosunek ufności reguły i wstępnego prawdopodobieństwa posiadania predykcji.
- **Wdrażalność** Miara tego, jaki procent danych uczących spełnia warunek, ale nie spełnia predykcji.

Zezwalaj na powtórzone predykcje Zaznacz to pole wyboru, aby podczas oceniania uwzględnić wiele reguł posiadających tę samą predykcję. Zaznaczenie tego pola umożliwi ocenę na przykład następujących reguł.

```
bread & cheese -> wine
cheese & fruit -> wine
```

Uwaga: Reguły z wieloma predykcjami (bread & cheese & fruit -> wine & pate) są traktowane jako powtórzone predykcje, pod warunkiem że wszystkie predykcje (wine & pate) zostały wcześniej przewidziane.

Oceniaj reguły tylko wówczas, gdy zmienne wejściowe nie zawierają predykcji Wybierz tę opcję, aby upewnić się, że predykcje nie istnieją w zmiennych wejściowych. Na przykład, jeśli celem oceny jest uzyskanie rekomendacji na temat mebli domowych, wówczas mało prawdopodobne jest to, że zmienna wejściowa, która zawiera już stół do jadalni, kupi kolejny taki stół. W takim przypadku wybierz tę opcję. Jednak w przypadku produktów psujących się lub jednorazowych (takich jak ser, mleko dla dzieci lub chusteczka higieniczna) wartościowe mogą być reguły, w których następnik istnieje już w zmiennych wejściowych. W tym drugim przypadku najbardziej użyteczną opcją może być **Oceń wszystkie reguły**.

Oceniaj reguły tylko wówczas, gdy zmienne wejściowe zawierają predykcje Wybierz tę opcję, aby upewnić się, że predykcje istnieją w zmiennych wejściowych. Takie podejście jest stosowane przy próbie uzyskania wglądu w istniejących klientów lub istniejące transakcje. Na przykład może pojawić się potrzeba znalezienia reguł o największym wzroście, a następnie sprawdzenia, którzy klienci spełniają te reguły.

Oceń wszystkie reguły Zaznacz tę opcję, aby podczas oceny uwzględnić wszystkie reguły bez względu na obecność lub brak predykcji.

Model użytkowy Reguły asocjacyjne

Model użytkowy zawiera informacje o regułach wyodrębnionych z posiadanych danych podczas budowy modelu.

Wyświetlanie wyników

W celu przeglądania reguł wygenerowanych przez modele reguł asocjacyjnych można korzystać z karty Model w oknie dialogowym. Przeglądanie modelu użytkowego ujawnia informacje o regułach przed wygenerowaniem nowych węzłów i przed oceną modelu.

Ocenianie modelu

Udoskonalone modele użytkowe mogą być dodawane do strumienia, a następnie oceniane. Więcej informacji można znaleźć w temacie “Używanie modeli użytkowych w strumieniach” na stronie 48. Okna dialogowe wartościowych informacji z modelu, które są używane podczas oceniania, zawierają dodatkową kartę Ustawienia. Więcej informacji można znaleźć w temacie “Ustawienia modelu użytkowego reguł asocjacyjnych”.

Szczegóły modelu użytkowego reguł asocjacyjnych

Model użytkowy modelu reguł asocjacyjnych przedstawia szczegóły dotyczące modelu na karcie Model w przeglądarce wyników. Więcej informacji na temat korzystania z okna raportów wynikowych można znaleźć w temacie „Praca z oknem wyników" w Podręczniku użytkownika programu Modeler (ModelerUsersGuide.pdf).

Operacja modelowania GSAR tworzy pewną liczbę nowych zmiennych z przedrostkiem $A, co przedstawiono w tabeli poniżej.

Tabela 25. Nowe zmienne tworzone przez operację modelowania reguł asocjacyjnych

Nazwa zmiennej	Opis
$A-<predykcja>#	Ta zmienna zawiera predykcję z modelu dla rekordów podlegających ocenie. <predykcja> to nazwa zmiennej uwzględnionej w roli Predykcje w modelu, a # jest sekwencją numerów reguł danych wyjściowych (na przykład, jeśli ocena jest ustawiona w taki sposób, aby zawierała 3 reguły, wówczas sekwencja będzie od 1 do 3).
$AC-<predykcja>#	Ta zmienna zawiera ufność w predykcji. <predykcja> to nazwa zmiennej uwzględnionej w roli Predykcje w modelu, a # jest sekwencją numerów reguł danych wyjściowych (na przykład, jeśli ocena jest ustawiona w taki sposób, aby zawierała 3 reguły, wówczas sekwencja będzie od 1 do 3).
$A-ID_reguły#	Ta kolumna zawiera identyfikator reguły przewidywanej dla każdego rekordu w ocenianym zbiorze danych. # jest sekwencją liczb dla reguł wyjściowych (na przykład, jeśli ocena jest ustawiona w taki sposób, aby zawierała 3 reguły, wówczas sekwencja będzie od 1 do 3).

Ustawienia modelu użytkowego reguł asocjacyjnych

Karta Ustawienia dla modelu użytkowego modelu reguł asocjacyjnych przedstawia opcje oceny modelu. Ta karta jest dostępna tylko po dodaniu modelu do obszaru roboczego strumienia na potrzeby oceny.

Maksymalna liczba predykcji Określ maksymalną liczbę predykcji, które będą uwzględnione dla każdego zbioru elementów. Reguły o najwyższych wartościach ufności mające zastosowanie do tego zestawu transakcji są używane do generowania predykcji dla tego rekordu, aż do zadanego limitu. Użyj tej opcji z pozycjami **Kryterium reguły** w celu wygenerowania „najważniejszych” predykcji, w których *najważniejsza* oznacza najwyższy poziom ufności, pokrycia, wzrostu itp.

Kryterium reguły Wybierz miarę, która będzie używana w celu określenia siły reguł. Reguły są sortowane wg siły kryteriów wybranych w tej opcji w celu uzyskania najważniejszych predykcji dla zbioru elementów. Do wyboru dostępne są następujące kryteria.

- **Ufność**
- **Pokrycie reguł**
- **Wzrost**
- **Pokrycie warunków**
- **Wdrażalność**

Zezwalaj na powtórzone predykcje Zaznacz to pole wyboru, aby podczas oceniania uwzględnić wiele reguł posiadających ten sam następnik. Na przykład zaznaczenie tej opcji oznacza, że może być oceniana następująca reguła:

```
bread & cheese -> wine
cheese & fruit -> wine
```

Aby wykluczyć powtórzone predykcje podczas oceniania, usuń zaznaczenie tego pola wyboru.

Uwaga: Reguły z wieloma następnikami (bread & cheese & fruit -> wine & pate) są traktowane jako powtórzone predykcje, pod warunkiem że wszystkie następniki (wine & pate) zostały wcześniej przewidziane.

Oceniaj reguły tylko wówczas, gdy zmienne wejściowe nie zawierają predykcji Wybierz tę opcję, aby upewnić się, że następniki nie istnieją w zmiennych wejściowych. Na przykład, jeśli celem oceny jest uzyskanie rekomendacji na temat mebli domowych, wówczas mało prawdopodobne jest to, że zmienna wejściowa, która zawiera już stół do jadalni, kupi kolejny taki stół. W takim przypadku wybierz tę opcję. Natomiast w przypadku produktów psujących się lub jednorazowych (takich jak ser, mleko dla dzieci lub chusteczka higieniczna) wartościowe mogą być reguły, w których następnik istnieje już w zmiennych wejściowych. W tym drugim przypadku najbardziej użyteczną opcją może być **Oceń wszystkie reguły**.

Oceniaj reguły tylko wówczas, gdy zmienne wejściowe zawierają predykcje Wybierz tę opcję, aby upewnić się, że następniki istnieją w zmiennych wejściowych. Takie podejście jest stosowane przy próbie uzyskania wglądu w istniejących klientów lub istniejące transakcje. Na przykład może pojawić się potrzeba znalezienia reguł o największym wzroście, a następnie sprawdzenia, którzy klienci spełniają te reguły.

Oceń wszystkie reguły Zaznacz tę opcję, aby podczas oceny uwzględnić wszystkie reguły bez względu na obecność lub brak następników w zmiennych wejściowych.

Rozdział 13. Modele szeregów czasowych

Dlaczego prognoza?

Prognozowanie oznacza predykcję wartości jednego lub większej liczby szeregów w czasie. Na przykład może okazać się przydatna predykcja oczekiwanego popytu na linię produktów lub usług, pozwalająca na odpowiednią alokację zasobów do działań produkcyjnych lub dystrybucyjnych. Ponieważ implementacja decyzji wynikających z planowania jest czasochłonna, prognozy są ważnym narzędziem w wielu procesach planowania.

W metodach modelowania szeregów czasowych zakłada się, że historia lubi się powtarzać — jeśli nie dokładnie, to wystarczająco podobnie, aby na podstawie badań przeszłości można było poprawić jakość przyszłych decyzji. Na przykład predykcja sprzedaży na następny rok zwykle zaczyna się od przejrzenia sprzedaży tegorocznej oraz jej porównania z latami przeszłymi pod kątem wyłaniających się trendów lub wzorców. Pomiar wzorców może jednak nastręczać trudności. Czy wzrost sprzedaży na przestrzeni kilku kolejnych tygodni jest na przykład elementem cyklu sezonowego, czy początkiem długoterminowego trendu?

Korzystając z technik modelowania statystycznego, można przeanalizować wzorce w danych przeszłych i rzutować te wzorce w celu określenia przedziału, w którym z dużym prawdopodobieństwem zawierać się będą przyszłe wartości szeregów. W efekcie uzyskujemy dokładniejsze prognozy stanowiące podstawę podejmowanych decyzji.

Dane szeregu czasowego

Szereg czasowy to uporządkowany zbiór pomiarów dokonywanych w regularnych odstępach czasu, na przykład codzienne notowania giełdowe lub cotygodniowe dane sprzedaży. Pomiary mogą dotyczyć dowolnego interesującego użytkownika zagadnienia, zaś każdy szereg można sklasyfikować jako jeden z następujących:

- **Zależny.** Szereg, dla którego chcesz utworzyć prognozę.
- **Predyktor.** Szereg, który może okazać się pomocny w wyjaśnieniu wartości przewidywanej — na przykład budżet reklamowy przy predykcji sprzedaży. Predyktory mogą być używane wyłącznie z modelami ARIMA.
- **Zdarzenie.** Specjalny szereg predykcyjny służący do rejestracji przewidywalnych incydentów rekurencyjnych — na przykład promocji sprzedaży.
- **Interwencja.** Specjalny szereg predykcyjny służący do rejestracji jednorazowych incydentów w przeszłości — na przykład przerwy w zasilaniu czy strajku pracowniczego.

Przedziały mogą reprezentować dowolną jednostkę czasu, lecz przedział musi być taki sam w przypadku wszystkich pomiarów. Co więcej, dla każdego przedziału, dla którego brak pomiarów, należy ustawić brak danych. W efekcie liczba przedziałów z pomiarami (w tym z brakami danych) definiuje długość czasu rozpiętości historycznej danych.

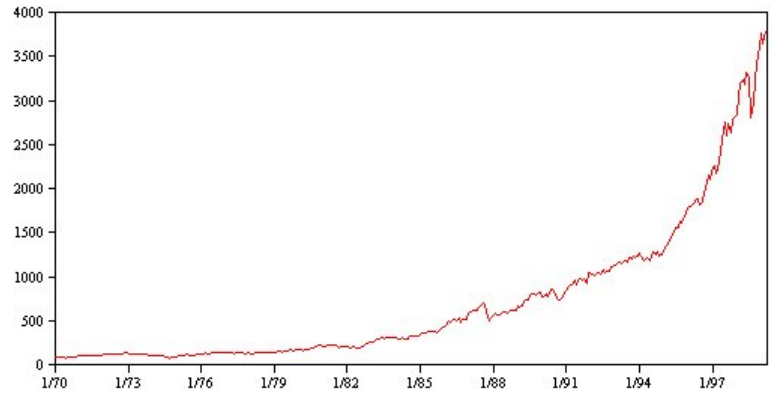
Cechy szeregu czasowego

Analiza przebiegu szeregu w przeszłości pomaga w identyfikacji wzorców i poprawia dokładność prognoz. Wykreślenie szeregów czasowych ujawnia w przypadku wielu z nich jedną lub więcej z poniższych właściwości:

- Trendy
- Cykle sezonowe i niesezonowe
- Pulsy i kroki
- Wartości odstające

Trendy

Trend oznacza stopniowe przesuwanie się poziomu szeregu w górę lub w dół albo tendencję wzrostową lub spadkową wartości szeregu w czasie.



Rysunek 53. Trend

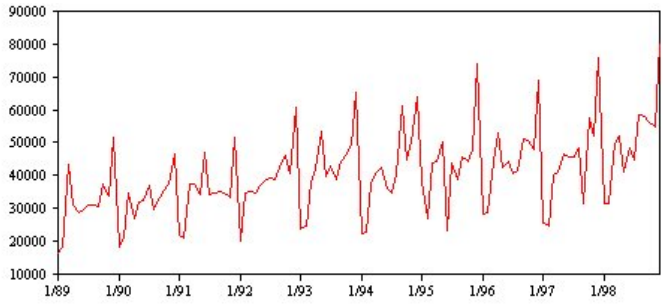
Trendy mogą mieć zasięg **lokalny** lub **globalny**, przy czym w jednym szeregu mogą występować oba ich rodzaje. Historycznie rzecz biorąc, wykresy szeregów indeksów giełdowych wykazują globalny trend rosnący. Lokalne trendy malejące pojawiały się w czasach recesji, zaś lokalne trendy rosnące w okresach prosperity.

Trendy mogą być również określane jako **liniowe** lub **nieliniowe**. Trendy liniowe to dodatnie lub ujemne przyrosty poziomu szeregu, porównywalne co do efektów z prostymi odsetkami od kapitału głównego. Trendy nieliniowe są często multiplikatywne, przy czym ich przyrosty są proporcjonalne do poprzednich wartości szeregu.

Globalne trendy liniowe poddają się dobrze dopasowaniom i prognozom zarówno za pomocą modelu wygładzania wykładniczego, jak i modelu ARIMA. Podczas budowy modeli ARIMA szeregi wykazujące trendy są zwykle różnicowane w celu wyeliminowania wpływu trendów.

Cykle sezonowe

Cykl sezonowy to powtarzalny, przewidywalny wzorzec kształtowania się wartości szeregu.



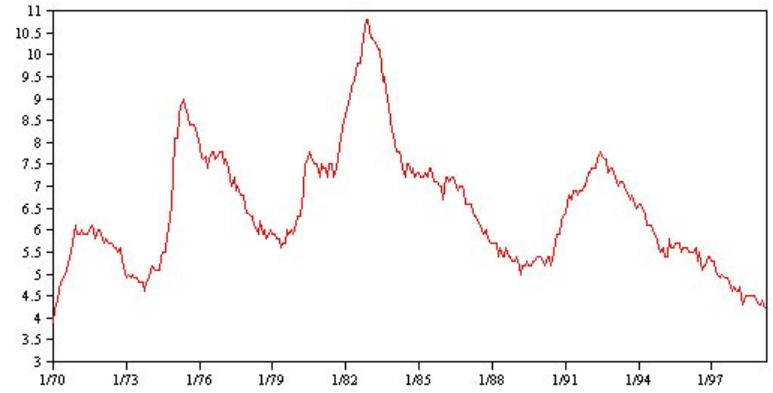
Rysunek 54. Cykl sezonowy

Cykle sezonowe są powiązane z przedziałem szeregu. Na przykład dane miesięczne zwykle powtarzają się cyklicznie w kolejnych kwartałach i latach. Dane szeregów miesięcznych mogą wskazywać na znaczącą cykliczność kwartalną, z niskimi wartościami w pierwszym kwartale lub na cykliczność roczną ze szczytem przypadającym na grudzień. Szereg przedstawiający cykl sezonowy jest określany jako wykazujący **sezonowość**.

Wzorce sezonowe pozwalają uzyskiwać dobre dopasowania i prognozy. Sezonowość wychwytują model wygładzania wykładniczego oraz model ARIMA.

Cykle niesezonowe

Cykl niesezonowy to powtarzalny, najprawdopodobniej nieprzewidywalny wzorzec kształtowania się wartości szeregu.



Rysunek 55. Cykl niesezonowy

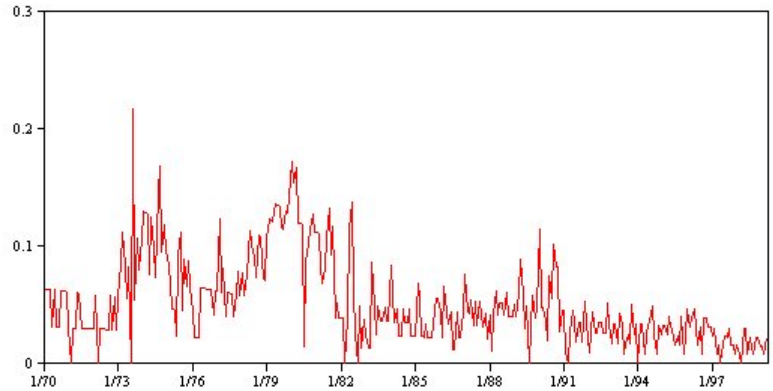
Niektóre szeregi, takie jak wskaźnik bezrobocia, wykazują wyraźną cykliczność; jednak okresowość cyklu zmienia się w czasie, co utrudnia predykcję wystąpienia kolejnego spadku lub wzrostu wskaźnika. Z kolei w przypadku innych szeregów cykle mogą być przewidywalne, lecz nie wpasowują się łatwo w kalendarz gregoriański lub są dłuższe niż rok. Na przykład pływy występują zgodnie z kalendarzem księżycowym, podróże międzynarodowe i transakcje handlowe związane z igrzyskami olimpijskimi co cztery lata, zaś w przypadku wielu świąt religijnych daty wg kalendarza gregoriańskiego zmieniają się z roku na rok.

Niesezonowe wzorce cykliczne sprawiają trudności w modelowaniu i ogólnie zwiększają niepewność prognozowania. Na przykład w przypadku rynku giełdowego można wskazać szereg instancji szeregów, które wymykają się prognozom. Jednocześnie konieczna jest rejestracja wzorców niesezonowych, o ile tylko istnieją. W wielu przypadkach użytkownik może nadal identyfikować model o względnie dobrym dopasowaniu do danych historycznych, co daje najlepsze szanse na minimalizację niepewności prognozowania.

Pulsy i kroki

W przypadku wielu szeregów występują gwałtowne zmiany poziomu. Zwykle są one jednego z dwu typów:

- Nagłe, *czasowe* przesunięcie lub **puls** w poziomie szeregu
- Nagłe, *trwałe* przesunięcie lub **krok** w poziomie szeregu



Rysunek 56. Szereg z pulsem

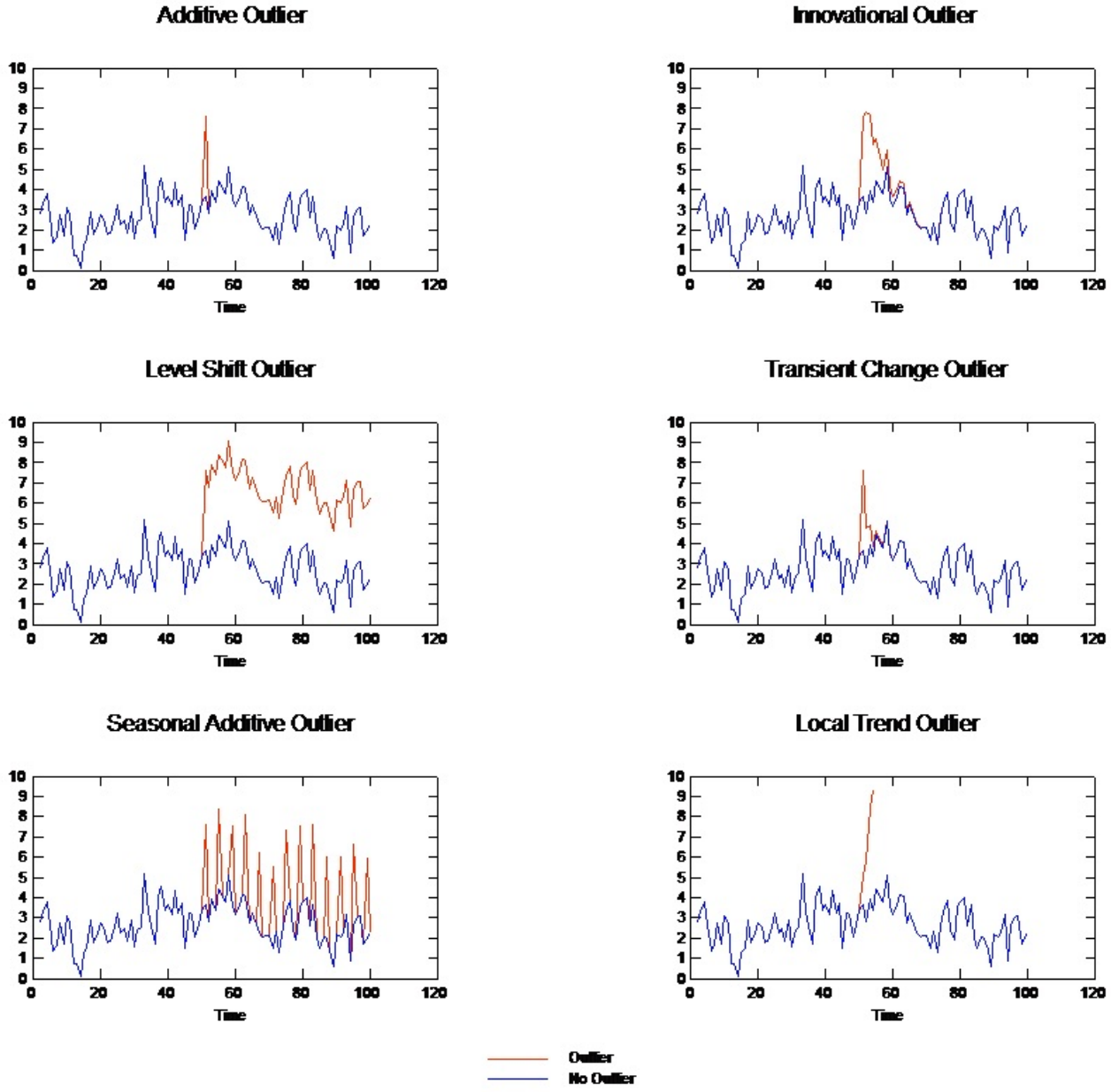
W przypadku zaobserwowania kroków lub pulsów ważne jest znalezienie prawdopodobnego wyjaśnienia ich pojawienia się. Modele szeregów czasowych opracowano z myślą o rejestrowaniu zmian stopniowych, nie zaś gwałtownych. W efekcie wykazują one tendencje do niedoszacowywania pulsów i gubienia się w przypadku kroków, to zaś prowadzi do słabego dopasowania modelu i niepewnych prognoz. (Niektóre instancje sezonowości mogą wykazywać nagłe zmiany w poziomie, przy czym poziom ten może utrzymywać się na stałym poziomie przy porównaniu jednego okresu sezonowego z drugim).

Jeśli zakłócenie można wyjaśnić, można je zamodelować za pomocą **interwencji** lub **zdarzenia**. Na przykład w sierpniu 1973 embargo na ropę naftową nałożone przez Organizację Państw Eksportujących Ropę Naftową (ang. Organization of Petroleum Exporting Countries (OPEC)) spowodowało drastyczną zmianę wskaźnika inflacji, który następnie, w kolejnych miesiącach, wrócił do normalnych poziomów. Opisując miesiąc obowiązywania embarga jako **interwencję punktową**, można poprawić dopasowanie modelu, pośrednio zwiększając tym samym dokładność prognoz. Na przykład sklep detaliczny może stwierdzić, że sprzedaż w dniu, w którym wszystkie towary przeceniono o 50%, była znacznie wyższa niż zwykle. Określając 50-procentową promocję jako rekurencyjne **zdarzenie**, można poprawić dopasowanie modelu i oszacować efekt powtórzenia promocji na wartości sprzedaży dla dni w przyszłości.

Wartości odstające

Przesunięcia w poziomie szeregu czasowego, których nie daje się wyjaśnić, są zwane **wartościami odstającymi**. Obserwacje te są niespójne z pozostałymi danymi szeregu i mogą drastycznie wpłynąć na analizę, a w efekcie pogorszyć możliwości prognozowania modelu szeregów czasowych.

Na poniższym rysunku przedstawiono kilka typów wartości odstających zwykle występujących w szeregach czasowych. Niebieska linia reprezentuje szereg bez wartości odstających. Czerwona linia sugeruje możliwość występowania wzorca, jeśli szereg zawierał wartości odstające. Te wartości odstające są klasyfikowane jako **deterministyczne**, ponieważ wpływają tylko na średni poziom szeregu.



Rysunek 57. Typy wartości odstających

- **Addytywna wartość odstająca.** Addytywna wartość odstająca pojawia się jako zaskakująco duża lub zaskakująco mała wartość występująca dla pojedynczej obserwacji. Addytywna wartość odstająca nie ma wpływu na kolejne obserwacje. Kolejne addytywne wartości odstające są zwykle zwane **addytywnymi wiązkami wartości odstających**.
- **Innowacyjna wartość odstająca.** Innowacyjna wartość odstająca charakteryzuje się wpływem początkowym z efektami opóźnionymi i rozciągniętymi na kolejne obserwacje. Wpływ wartości odstających może zwiększać się w miarę upływu czasu.
- **Wartość odstająca przesunięcia poziomu.** W przypadku przesunięcia poziomu wszystkie obserwacje za wartością odstającą przesuwają się na nowy poziom. Inaczej niż w przypadku addytywnych wartości odstających, wartość odstająca przesunięcia poziomu wpływa na wiele obserwacji, a wpływ ten jest trwały.

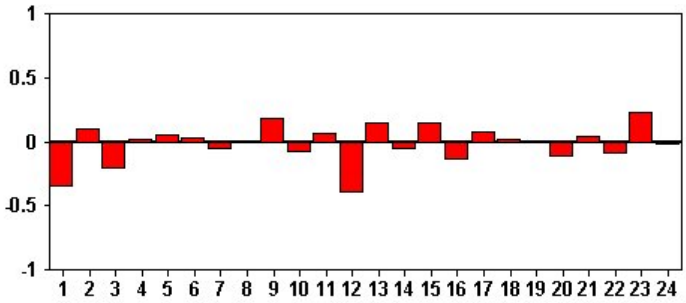
- **Wartość odstająca zmiany przemijającej.** Wartości odstające zmiany przemijającej przypominają wartości odstające przesunięcia poziomu, przy czym wpływ wartości odstającej wygasza się wykładniczo w miarę kolejnych obserwacji. Ostatecznie szereg powraca do normalnego poziomu.
- **Wartość odstająca sezonowo addytywna.** Wartość odstająca sezonowo addytywna pojawia się jako zaskakująco duża lub zaskakująco mała wartość występująca okresowo, w regularnych odstępach.
- **Wartość odstająca trendu lokalnego.** Wartość odstająca trendu lokalnego powoduje ogólne przesunięcie w szeregu spowodowane przez wzorzec powstały w wartościach odstających po wystąpieniu początkowej wartości odstającej.

Wykrywanie wartości odstających w szeregach czasowych wymaga określenia lokalizacji, typu i wartości bezwzględnej dla każdej występującej wartości odstającej. Tsay (1988) zaproponował procedurę iteracyjną wykrywania zmiany poziomu średniej z myślą o identyfikacji deterministycznych wartości odstających. Proces ten wymaga porównania modelu szeregów czasowych uznawanego za niezawierający wartości odstających z drugim modelem, w którym wartości odstające występują. Różnice między modelami stanowią oszacowanie wpływu traktowania danego punktu jako wartości odstającej.

Funkcje autokorelacji i autokorelacji cząstkowej

Autokorelacja i autokorelacja cząstkowa to miary związków między bieżącymi i przeszłymi wartościami szeregów określające, które przeszłe wartości szeregów są najbardziej użyteczne przy przewidywaniu przyszłych wartości. Dzięki tej wiedzy można określić kolejność procesów w modelu ARIMA. Dokładniej rzecz ujmując:

- **Funkcja autokorelacji (ACF).** Przy przesunięciu *k* jest to korelacja między wartościami szeregu oddalonymi o *k* przedziałów od siebie.
- **Funkcja autokorelacji cząstkowej (PACF).** Przy przesunięciu *k* jest to korelacja między wartościami szeregu oddalonymi o *k* przedziałów od siebie, z jednoczesną rejestracją wartości z przedziałów znajdujących się pomiędzy.



Rysunek 58. Wykres ACF szeregu

Oś *x* wykresu ACF wskazuje przesunięcie, dla jakiego obliczana jest autokorelacja; oś *y* wskazuje wartość korelacji (między −1 a 1). Na przykład linia rzutowania przy przesunięciu wynoszącym 1 na wykresie ACF oznacza silną korelację między każdą wartością szeregu a wartością poprzedzającą, zaś linia rzutowania przy przesunięciu wynoszącym 2 oznacza silną korelację między każdą wartością a wartością występującą dwa punkty wcześniej itd.

- Korelacja dodatnia oznacza, że duże wartości bieżące odpowiadają dużym wartościom dla danego przesunięcia; ujemna korelacja oznacza, że duże wartości bieżące odpowiadają niewielkim wartościom dla danego przesunięcia.
- Wartość bezwzględna korelacji stanowi miarę siły powiązania, przy czym większe wartości bezwzględne oznaczają silniejsze relacje.

Transformacje szeregów

Transformacje są często przydatne do stabilizowania szeregów przed oszacowaniem modeli. Jest to szczególnie ważne w przypadku modeli ARIMA, które wymagają, aby estymowany model był **stacjonarny**. Szereg jest stacjonarny, jeśli poziom globalny (średnia) oraz odchylenie średnie od poziomu (wariancja) są stałe na przestrzeni szeregu.

Choć najbardziej interesujące szeregi nie są stacjonarne, model ARIMA jest przydatny, o ile tylko szeregi te można przekształcić w stacjonarne, stosując transformacje, takie jak logarytm naturalny, różnicowanie czy różnicowanie sezonowe.

Transformacje stabilizujące wariancje. Szereg, w ramach którego zmiany wariancji w czasie mogą być często stabilizowane przez transformacje, takie jak logarytm naturalny czy pierwiastek kwadratowy. Są one również zwane transformacjami funkcjonalnymi.

- **Logarytm naturalny.** Do wartości szeregu stosowany jest logarytm naturalny.
- **Pierwiastek kwadratowy.** Do wartości szeregu stosowany jest pierwiastek kwadratowy.

Transformacje logarytmem naturalnym i pierwiastkiem kwadratowym nie mogą być używane w przypadku szeregów zawierających wartości ujemne.

Transformacje stabilizujące poziom. Niewielki spadek wartości ACF oznacza, że każda wartość szeregu jest silnie skorelowana z poprzednią. Analizując zmianę w wartościach szeregu, uzyskuje się stabilny poziom.

- **Różnicowanie proste.** Polega na obliczaniu różnic między każdą wartością a poprzednią wartością w szeregu, z pominięciem najstarszej wartości w szeregu. Oznacza to, że różnicowany szereg będzie miał o jedną wartość mniej niż szereg oryginalny.
- **Różnicowanie sezonowe.** Przebiega podobnie do różnicowania prostego, z tą różnicą, że obliczane są różnice między każdą wartością a poprzednią wartością sezonową.

Jeśli różnicowanie proste lub różnicowanie sezonowe jest stosowane jednocześnie z transformacją logarytmiczną lub z transformacją pierwiastkiem kwadratowym, transformacja stabilizująca wariancję jest zawsze przeprowadzana jako pierwsza. Jeśli używane są jednocześnie różnicowanie proste i różnicowanie sezonowe, wartości wynikowego szeregu są takie same, niezależnie od tego, która transformacja zostanie przeprowadzona jako pierwsza.

Szereg predykcyjny

Szereg predykcyjny obejmuje powiązane dane, które mogą pomóc w wyjaśnieniu zachowań szeregu, dla którego ma zostać utworzona prognoza. Na przykład sklep internetowy lub sprzedaży wysyłkowej może prognozować sprzedaż w oparciu o liczbę rozesłanych katalogów, liczbę czynnych linii telefonicznych czy liczbę odwiedzin strony WWW firmy.

Jako predyktor może zostać użyty każdy szereg, pod warunkiem że zostanie on rozciągnięty w przyszłość odpowiednio do oczekiwanego okresu prognozy, i jest kompletny, to znaczy nie zawiera braków danych.

Dodając predyktory do modelu, należy zachować ostrożność. Dodanie dużej liczby predyktorów wydłuży czas potrzebny do oszacowania modelu. Podczas gdy dodanie predyktorów może poprawić zdolność modelu do dopasowania do danych historycznych, nie oznacza ono koniecznie, że jakość prognozowania modelu ulegnie poprawie, zatem zwiększenie złożoności modelu niekoniecznie się opłaci. W idealnym przypadku celem powinno być określenie najprostszego dobrze sprawdzającego się w prognozowaniu modelu.

Jako ogólną zasadę zaleca się stosowanie liczby predyktorów mniejszej od wielkości próby podzielonej przez 15 (to jest, aby na 15 obserwacji przypadał maksymalnie jeden predyktor).

Predyktory z brakami danych. Predyktory z niekompletnymi danymi lub z brakami danych nie mogą być używane podczas prognozowania. Dotyczy to zarówno danych historycznych, jak i przyszłych wartości. W niektórych przypadkach można uniknąć tego ograniczenia, ustawiając rozpiętość oszacowania modelu tak, aby wykluczyć z estymacji modeli najstarsze dane.

Węzeł modelowania Predykcja przestrzenno-czasowa (STP)

Predykcja przestrzenno-czasowa (STP) ma wiele potencjalnych zastosowań, takich jak zarządzanie energią w budynkach i zakładach produkcyjnych, analiza wydajności i prognozowanie obciążenia inżynierów serwisu czy planowanie transportu publicznego. W tych zastosowaniach pomiary (np. energii) często prowadzone są w różnych miejscach i w różnym czasie. Pytania związane z rejestrowaniem tych pomiarów to między innymi pytania o czynniki

wpływające na przyszłe obserwacje, pytania o to, co można zrobić, aby uzyskać pożądaną zmianę lub aby lepiej zarządzać systemem. Odpowiedzi na te pytania można uzyskać, korzystając z technik statystycznych umożliwiających prognozowanie przyszłych wartości w różnych lokalizacjach, a także jawne modelowanie współczynników w celu przeprowadzenia analizy co-jeśli.

W predykcji przestrzenno-czasowej używa się danych zawierających informacje o lokalizacji, zmiennych wejściowych predykcji (predyktorów), zmiennej czasu i zmiennej przewidywanej. W danych z każdą lokalizacją powiązany jest szereg wierszy, które odzwierciedlają wartości predyktorów w różnych punktach w czasie. Po przeanalizowaniu danych mogą być one używane do przewidywania wartości w dowolnej lokalizacji w danych kształtu używanych w analizie. Pozwalają one także tworzyć prognozy, o ile znane są dane wejściowe dla przyszłych punktów w czasie.

Uwaga: Węzeł STP nie obsługuje etapów Ocena modelu ani Champion Challenger programu IBM SPSS Collaboration and Deployment Services.

Strumień przedstawiający sprawdzony przykład użycia predykcji STP, o nazwie stp_server_demo.str, odwołujący się do plików danych room_data.csv i score_data.csv, jest dostępny w katalogu Demos pakietu instalacyjnego programu IBM SPSS Modeler. Dostęp do katalogu Demos można uzyskać z grupy programu IBM SPSS Modeler w menu Start systemu Windows. Plik stp_server_demo.str znajduje się w katalogu streams.

Predykcja przestrzenno-czasowa — opcje zmiennych

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról
Ta opcja wykorzystuje ustawienia ról (tylko wartości przewidywane i predyktory) z poprzedzającego węzła Typ (lub z karty Typ poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań
Zaznacz tę opcję, aby ręcznie przypisać na tym ekranie wartości przewidywane, predyktory i inne role.

Zmienne
Wyświetla wszystkie zmienne w danych, które można wybrać. Za pomocą przycisków strzałek przypisz elementy z listy ręcznie do poszczególnych pól po prawej stronie ekranu. Ikony wskazują prawidłowe poziomy pomiaru dla każdej zmiennej.

Uwaga: Prawidłowe funkcjonowanie STP wymaga 1 rekordu na lokalizację na przedział czasu; dlatego pola te są wymagane.

W dolnej części panelu **Zmienne** kliknij przycisk **Wszystko**, aby wybrać wszystkie zmienne, niezależnie od poziomu pomiaru, lub kliknij przycisk indywidualnego poziomu pomiaru, aby wybrać wszystkie zmienne o tym poziomie pomiaru.

Zmienna przewidywana
Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji.

Uwaga: Można wybrać wyłącznie spośród zmiennych, których poziom pomiaru to Ilościowy.

Lokalizacja
Wybierz typ lokalizacji, który ma być używany w modelu.

Uwaga: Można wybrać wyłącznie spośród zmiennych, których poziom pomiaru to geoprzestrzenny.

Etykieta lokalizacji
Dane kształtu często obejmują zmienną przedstawiającą nazwy funkcji w warstwie, na przykład nazwy województw lub regionów. Użyj tej zmiennej w celu powiązania nazwy lub etykiety z lokalizacją, wybierając zmienną jakościową w celu nadania etykiety wybranej zmiennej **Lokalizacja** w wynikach.

Zmienna czasu
Wybierz zmienne czasu, które mają być używane w predykcjach.

Uwaga: Można wybrać wyłącznie spośród zmiennych, których poziom pomiaru jest ilościowy, zaś typ składowania to czas, data, znacznik czasu lub liczba całkowita.

Predyktory (dane wejściowe)
Wybierz jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Uwaga: Można wybrać wyłącznie spośród zmiennych, których poziom pomiaru to Ilościowa.

Predykcja przestrzenno-czasowa — przedziały czasowe

W panelu Przedziały czasowe można wybrać opcje umożliwiające ustawienie przedziału czasowego oraz wszelkich wymaganych agregacji w czasie.

Konwersja zmiennych czasu do indeksu przed przystąpieniem do budowy modelu STP wymaga przygotowania danych; aby konwersja była możliwa, miedzy rekordami zmiennej czasu muszą występować stałe przedziały czasowe. Jeśli dane użytkownika nie zawierają jeszcze tych informacji, aby możliwe było użycie węzła modelowania, należy ustawić ten przedział za pomocą opcji w tym panelu.

Przedział czasu Wybierz przedział, na który chcesz, aby dane zostały przekonwertowane. Dostępność opcji zależy od typu składowania zmiennej wybranej jako **Zmienna czasu** dla modelu na karcie Zmienne.

- **Okresy** Ta opcja jest dostępna tylko dla zmiennych czasu będących liczbami całkowitymi; jest to szereg przedziałów z równomiernymi przedziałami między poszczególnymi pomiarami, nieodpowiadającymi dostępnym opcjom.
- **Lata** Ta opcja jest dostępna tylko dla zmiennych czasu Data i Znacznik czasu.
- **Kwartały** Ta opcja jest dostępna tylko dla zmiennych czasu Data i Znacznik czasu. Po wybraniu tej opcji wyświetlany jest monit o wybór wartości **Miesiąc początkowy** dla pierwszego kwartału.
- **Miesiące** Ta opcja jest dostępna tylko dla zmiennych czasu Data i Znacznik czasu.
- **Tygodnie** Ta opcja jest dostępna tylko dla zmiennych czasu Data i Znacznik czasu.
- **Dni** Ta opcja jest dostępna tylko dla zmiennych czasu Data i Znacznik czasu.
- **Godziny** Ta opcja jest dostępna tylko dla zmiennych czasu Data i Znacznik czasu.
- **Minuty** Ta opcja jest dostępna tylko dla zmiennych czasu Data i Znacznik czasu.
- **Sekundy** Ta opcja jest dostępna tylko dla zmiennych czasu Data i Znacznik czasu.

Po wybraniu opcji **Przedział czasu** użytkownik jest monitowany o wprowadzenie kolejnych zmiennych. Są one zależne od przedziału czasu i typu składowania. Zmienne, które mogą być wyświetlane, umieszczono na poniższej liście.

- **Liczba dni w tygodniu**
- **Liczba godzin w dniu**
- **Tydzień zaczyna się od** Pierwszy dzień tygodnia
- **Dzień zaczyna się o godzinie** Godzina, którą uznaje się za rozpoczynającą nowy dzień.
- **Wartość przedziału** Można wybrać jedną z następujących opcji: 1, 2, 3, 4, 5, 6, 10, 12, 15, 20 lub 30.
- **Miesiąc początkowy** Miesiąc, w którym rozpoczyna się rok finansowy.
- **Okres początkowy** Jeśli wybrano opcję **Okresy**, należy wybrać okres początkowy.

Dane odpowiadają określonym ustawieniom przedziałów czasowych Jeśli dane zawierają już poprawne informacje dotyczące przedziałów czasowych i nie wymagają konwersji, należy zaznaczyć to pole wyboru. Po zaznaczeniu tego pola wyboru zmienne w obszarze **Agregacja** będą niedostępne.

Agregacja

Ta opcja jest dostępna wyłącznie po usunięciu zaznaczenia pola wyboru **Dane odpowiadają określonym ustawieniom przedziałów czasowych**; należy określić opcje agregacji zmiennych w celu dopasowania ich do podanego przedziału. Na przykład w przypadku mieszaniny danych tygodniowych i miesięcznych możliwa jest agregacja lub zwinięcie

wartości tygodniowych w celu uzyskania równomiernych przedziałów miesięcznych. Należy wybrać ustawienia domyślne, które mają być używane na potrzeby agregacji różnych typów zmiennych, a następnie utworzyć wszelkie ustawienia niestandardowe niezbędne dla dowolnych wybranych zmiennych.

- **Ilościowa** Umożliwia ustawienie domyślnej metody agregacji stosowanej do wszystkich zmiennych ilościowych nieokreślonych indywidualnie. Można wybrać spośród kilku metod:
  - Suma
  - Średnia
  - Minimum
  - Maksimum
  - Mediana
  - Pierwszy kwartyl
  - Trzeci kwartyl

Ustawienia niestandardowe dla określonych zmiennych W celu zastosowania do poszczególnych zmiennych określonej funkcji agregacji należy wybrać je w tej tabeli, a następnie wybrać metodę agregacji.

- **Zmienna** Przycisk **Dodaj zmienną** powoduje wyświetlenie okna dialogowego Wybierz zmienne i wybór wymaganych zmiennych. Wybrane zmienne są wyświetlane w tej kolumnie.
- **Funkcja agregująca** Z rozwijanej listy wybierz funkcję agregującą w celu przekonwertowania zmiennej w określony przedział czasu.

Predykcja przestrzenno-czasowa — podstawowe opcje budowy

Ustawienia w tym oknie dialogowym umożliwiają ustawienie podstawowych opcji budowania modelu.

Ustawienia modelu

Uwzględnij wyraz wolny

Uwzględnienie wyrazu wolnego (stały składnik modelu) może spowodować zwiększenie ogólnej dokładności rozwiązania. Jeśli można założyć, że dane przechodzą przez początek układu współrzędnych, wyraz wolny można wyłączyć z modelu.

Kolejność autoregresji maksimum

Rzędy autoregresji określają, które z poprzednich wartości są używane do przewidywania bieżących wartości. Opcja ta umożliwia określenie liczby poprzednich rekordów używanych do obliczenia nowej wartości. Można wybrać dowolną liczbę całkowitą z przedziału od 1 do 5.

Kowariancja przestrzenna

Metoda estymacji

Wybierz metodę estymacji, która ma być używana; można wybrać albo opcję **Parametryczna**, albo **Nieparametryczna**. W przypadku metody **Parametryczna** można wybrać jeden z trzech typów **Model**:

- **Gaussa**
- **Wykładniczy**
- **Wykładniczo-potęgowy** W przypadku wybrania tej opcji należy także określić żądany poziom **Potęga**. Ten poziom może być dowolną wartością z przedziału od 1 do 2, ze zmianą o przyroście 0,1.

Predykcja przestrzenno-czasowa — zaawansowane opcje budowy

Użytkownicy ze szczegółową znajomością predykcji przestrzenno-czasowej mogą skorzystać z poniższych opcji w celu precyzyjnego dostosowania procesu budowy modelu do swoich potrzeb.

Maksymalny procent brakujących wartości

Umożliwia określenie maksymalnej wartości procentowej rekordów zawierających braki danych, która może zostać uwzględniona w modelu.

Poziom istotności dla testowania hipotez w procesie budowania modelu
Określ wartość poziomu istotności, jaka ma być używana dla wszystkich testów w ramach oszacowania modelu STP, w tym dwu testów Dobroć dopasowania, testów Efekt F i Współczynnik T. Ten poziom może być dowolną wartością z przedziału od 0 do 1, ze zmianą o przyroście 0,01.

Predykcja przestrzenno-czasowa — wynik

Przed przystąpieniem do budowy modelu należy za pomocą opcji w tym panelu wybrać wynik, który ma być wyświetlany w oknie wynikowym.

Informacje o modelu

Specyfikacja modelu
Wybór tej opcji pozwala uwzględnić informacje dotyczące specyfikacji modelu w wyniku modelu.

Podsumowanie informacji o definicji czasu
Wybór tej opcji pozwala uwzględnić podsumowanie informacji dotyczących czasu w wyniku modelu.

Ewaluacja

Jakość modelu
Wybór tej opcji pozwala uwzględnić jakość modelu w wyniku modelu.

Test efektów w modelu strukturalnym średniej
Wybór tej opcji pozwala uwzględnić informacje dotyczące testu wpływów w wyniku modelu.

Interpretacja

Struktura średniej współczynników modelu
Wybór tej opcji pozwala uwzględnić współczynniki struktury średnich modelu w wyniku modelu.

Współczynniki autoregresyjne
Wybór tej opcji pozwala uwzględnić informacje dotyczące współczynników autoregresji w wyniku modelu.

Testy spadku przestrzennego
Wybór tej opcji pozwala uwzględnić w wyniku modelu dane testu kowariancji przestrzennej lub dane o spadku przestrzennym.

Wykres parametrów parametrycznego modelu kowariancji przestrzennej
Wybór tej opcji pozwala uwzględnić w wyniku modelu dane wykresu parametrów parametrycznego modelu kowariancji.

Uwaga: Ta opcja jest dostępna wyłącznie, jeśli wybrano metodę estymacji **Parametryczna** na karcie Podstawowe.

Mapa natężeń korelacji
Wybór tej opcji pozwala uwzględnić w wyniku modelu mapę wartości przewidywanych.

Uwaga: Jeśli model obejmuje więcej niż 500 lokalizacji, mapa nie zostanie utworzona.

Mapa korelacji
Wybór tej opcji pozwala uwzględnić w wyniku modelu mapę korelacji.

Uwaga: Jeśli model obejmuje więcej niż 500 lokalizacji, mapa nie zostanie utworzona.

Grupy lokalizacji
Wybór tej opcji pozwala uwzględnić informacje dotyczące grupowania lokalizacji w wyniku modelu. Uwzględniane są tylko wyniki niewymagające dostępu do danych mapy.

Uwaga: Wynik można tworzyć tylko dla nieparametrycznego modelu kowariancji przestrzennej.

Po wybraniu tej opcji można ustawić następujące elementy:

- **Wartość graniczna podobieństwa** Pozwala wybrać wartość, przy której grupy wyników będą uznawane za wystarczająco podobne, aby mogły zostać scalone w jedną grupę.
- **Maksymalna liczba grup do wyświetlenia** Pozwala ustawić górny limit dla liczby klastrów, które mają zostać uwzględnione w wyniku modelu.

Predykcja przestrzenno-czasowa — opcje modelu

Nazwa modelu Można automatycznie generować nazwę modelu na podstawie zmiennych przewidywanych lub podać nazwę użytkownika. Automatycznie wygenerowana nazwa jest nazwą zmiennej docelowej.

Współczynnik niepewności (%) Współczynnik niepewności to wartość procentowa reprezentująca wzrost niepewności związany z prognozowaniem przyszłości. Górna i dolna granica niepewności prognozy zwiększa się o tę wartość procentowo wraz z każdym krokiem w przyszłość. Ustaw współczynnik niepewności stosowany do wyników modelu; powoduje to ustawienie górnej i dolnej granicy dla przewidywanych wartości.

Model użytkowy predykcji przestrzenno-czasowej

Model użytkowy predykcji przestrzenno-czasowej przedstawia szczegółowe informacje dotyczące modelu na karcie Model Raportu wynikowego. Więcej informacji na temat korzystania z okna raportów wynikowych można znaleźć w temacie „Praca z oknem wyników" w Podręczniku użytkownika programu Modeler (ModelerUsersGuide.pdf).

Operacja modelowania predykcji przestrzenno-czasowej umożliwia tworzenie pewnej liczby nowych zmiennych o prefiksie $STP- zgodnie z danymi w poniższej tabeli.

Tabela 26. Nowe zmienne tworzone w ramach operacji modelowania STP

Nazwa zmiennej	Opis
$STP-<Time>	Zmienna czasu tworzona jako część budowanego modelu. Ustawienia w panelu Przedziały czasowe na karcie Opcje budowania określają sposób tworzenia tej zmiennej. <Time> to oryginalna nazwa zmiennej wybrana jako **Zmienna czasu** na karcie Zmienne. **Uwaga:** To pole jest tworzone wyłącznie, jeśli oryginalna **Zmienna czasu** została przekonwertowana jako część budowanego modelu.
$STP-<Target>	Ta zmienna zawiera predykcje dla wartości przewidywanej. <Target> jest nazwą oryginalnej zmiennej **Target** dla danego modelu
$STPVAR-<Target>	Ta zmienna zawiera wartości wariancji predykcji dla punktu. <Target> jest nazwą oryginalnej zmiennej **Target** dla danego modelu
$STPLCI-<Target>	Ta zmienna zawiera wartości mniejsze od przedziału predykcji, to jest wartości mniejsze niż dolna granica ufności. <Target> jest nazwą oryginalnej zmiennej **Target** dla danego modelu
$STPUCI-<Target>	Ta zmienna zawiera wartości większe od przedziału predykcji, to jest wartości większe niż górna granica ufności. <Target> jest nazwą oryginalnej zmiennej **Target** dla danego modelu

Ustawienia modelu predykcji przestrzenno-czasowej

Karta Ustawienia pozwala kontrolować poziom niepewności uważany za akceptowalny w ramach operacji modelowania.

Współczynnik niepewności (%) Współczynnik niepewności to wartość procentowa reprezentująca wzrost niepewności związany z prognozowaniem przyszłości. Górna i dolna granica niepewności prognozy zwiększa się o tę wartość procentowo wraz z każdym krokiem w przyszłość. Ustaw współczynnik niepewności stosowany do wyników modelu; powoduje to ustawienie górnej i dolnej granicy dla przewidywanych wartości.

Węzeł TCM

Ten węzeł umożliwia tworzenie modelu przyczynowego szeregów czasowych (TCM).

Modele przyczynowe szeregów czasowych

Modelowanie przyczynowe szeregów czasowych jest próbą wykrycia kluczowych zależności przyczynowych w danych o szeregach czasowych. W procesie modelowania przyczynowego szeregów czasowych użytkownik określa zbiór szeregów przewidywanych i zbiór potencjalnych zmiennych wejściowych dla tych szeregów przewidywanych. Następnie procedura buduje autoregresyjny model każdego szeregu przewidywanego i uwzględnia tylko te zmienne wejściowe, które z przewidywanym szeregiem łączą relacje przyczynowe. Ta strategia różni się od tradycyjnego modelowania szeregów czasowych, w którym użytkownik musi jawnie określić predyktory dla przewidywanego szeregu. Ponieważ w modelowaniu przyczynowym szeregów czasowych zazwyczaj buduje się modele dla wielu powiązanych szeregów czasowych, wynik tego modelowania nazywamy *systemem modelu*.

W kontekście modelowania przyczynowego szeregów czasowych termin *przyczynowe* odnosi się do przyczynowości w sensie Grangera. Szereg czasowy X jest uznawany za „przyczynę w sensie Grangera" innego szeregu czasowego Y, jeśli regresja dla Y przy przeszłych wartościach szeregów X i Y zwraca lepszy model dla Y niż regresja tylko przy przeszłych wartościach dla Y.

Uwaga: Węzeł modelowania przyczynowego szeregów czasowych nie obsługuje etapów Ocena modelu ani Champion Challenger programu IBM SPSS Collaboration and Deployment Services.

Przykłady

Osoby podejmujące decyzje biznesowe mogą korzystać z modelowania przyczynowego szeregów czasowych do odkrywania relacji przyczynowych w dużych zestawach metryk w oparciu o czas opisujących procesy biznesowe. Efektem analizy może być zidentyfikowanie możliwych do zmiany danych wejściowych mających największy wpływ na kluczowe wskaźniki wydajności.

Menedżerowie dużych systemów IT mogą korzystać z modelowania przyczynowego szeregów czasowych do wykrywania anomalii w dużych zbiorach powiązanych ze sobą metryk operacyjnych. Model przyczynowy umożliwia wówczas więcej niż tylko wykrywanie anomalii i odkrywanie najbardziej prawdopodobnych ich przyczyn.

Wymagania dotyczące zmiennych

Wymagana jest co najmniej jedna zmienna przewidywana. Domyślnie zmienne o predefiniowanej roli Brak nie są używane.

Struktura danych

Modelowanie przyczynowe szeregów czasowych obsługuje dwa rodzaje struktur danych.

Dane kolumnowe
: W przypadku danych kolumnowych każda zmienna szeregu czasowego zawiera dane dla jednego szeregu czasowego. Struktura ta jest tradycyjną strukturą danych szeregów czasowych, stosowaną przez kreator modeli szeregów czasowych.

Dane wielowymiarowe
: W przypadku danych wielowymiarowych każda zmienna szeregu czasowego zawiera dane dla wielu szeregów czasowych. Oddzielne szeregi czasowe w ramach danej zmiennej są wówczas identyfikowane na podstawie zestawu wartości zmiennych jakościowych, zwanych także zmiennymi *wymiarów*. Na przykład dane sprzedaży dla dwu różnych kanałów sprzedaży (detalicznego i internetowego) mogą być przechowywane w jednej zmiennej *sales*. Zmienna wymiarowa o nazwie *channel* i wartościach 'retail' oraz 'web' określa rekordy powiązane z każdym z tych dwu kanałów sprzedaży.

Uwaga: Do utworzenia modelu przyczynowego szeregów czasowych potrzebna jest odpowiednio duża liczba punktów. W programie obowiązuje ograniczenie:

```
m>(L + KL + 1)
```

gdzie m jest liczbą punktów danych, L jest liczbą opóźnień, a K jest liczbą predyktorów. Dane muszą być na tyle obszerne, by liczba punktów danych (m) spełniała ten warunek.

Szeregi czasowe dla modelowania

Na karcie Zmienne należy użyć ustawień **Szeregi czasowe**, aby określić szeregi, jakie będą uwzględniane w systemie modelu.

Należy wybrać opcję dla struktury danych, która jest odpowiednia dla przetwarzanych danych. W przypadku danych wielowymiarowych należy kliknąć opcję **Wybierz wymiary**, aby określić zmienne wymiarów. Kolejność określona w zmiennych wymiaru definiuje kolejność, w jakiej będą wyświetlane wszystkie następne okna dialogowe i wyniki. Kolejność zmiennych wymiarów można zmieniać za pomocą przycisków strzałek w górę i w dół w podrzędnym oknie dialogowym Wybierz wymiary.

W przypadku danych kolumnowych pojęcie *szeregi* ma takie samo znaczenie jak pojęcie *zmienna*. W przypadku danych wielowymiarowych zmienne, które zawierają szeregi czasowe, są nazywane zmiennymi *metryk*. Szeregi czasowe, dla danych wielowymiarowych, są definiowane przez zmienną metryk i wartość dla każdej zmiennej wymiaru. Przedstawione poniżej rozważania mają zastosowanie do danych kolumnowych oraz do danych wielowymiarowych.

- Szeregi określane jako potencjalne zmienne wejściowe oraz jako przewidywane i wejściowe są uwzględniane w celu dołączenia do modelu dla każdej zmiennej przewidywanej. Model dla każdej zmiennej przewidywanej zawsze uwzględnia wartości opóźnione samej zmiennej przewidywanej.
- Szeregi określane jako wymuszone zmienne wejściowe zawsze są uwzględniane w modelu dla każdej zmiennej przewidywanej.
- Co najmniej jeden szereg musi być określony jako przewidywany lub jako przewidywany i wejściowy.
- Jeśli zaznaczono opcję **Użyj wstępnie zdefiniowanych ról**, zmienne z rolą Zmienna wejściowa są ustawiane jako potencjalne zmienne wejściowe. Żadna wstępnie zdefiniowana rola nie jest odwzorowywana na potencjalną zmienną wejściową.

Dane wielowymiarowe

W przypadku danych wielowymiarowych zmienne metryk i powiązane role są określane w siatce, w której każdy wiersz określa pojedynczą metrykę lub rolę. Domyślnie system modelu obejmuje szeregi dla wszystkich kombinacji zmiennych wymiarów dla każdego wiersza w siatce. Jeśli na przykład dostępne są wymiary dla *region* i *brand*, domyślnie określenie metryki *sales* jako przewidywanej oznacza, że istnieje osobny szereg przewidywany sprzedaży dla każdej kombinacji wartości *region* i *brand*.

Dla każdego wiersza w siatce można dostosować zestaw wartości dowolnej zmiennej wymiaru, klikając przycisk wielokropka danego wymiaru. Ta czynność otworzy podrzędne okno dialogowe Wybierz wartości wymiarów. Wiersze siatki można również dodawać, usuwać lub kopiować.

W kolumnie **Liczebność szeregu** wyświetlana jest liczba zestawów wartości wymiarów, jakie są aktualnie określone dla powiązanej metryki. Wyświetlana wartość może być większa niż rzeczywista liczba szeregów (jeden szereg na zestaw). Taka sytuacja ma miejsce, jeśli niektóre z określonych kombinacji wartości wymiarów nie odpowiadają szeregom objętym przez powiązaną metrykę.

Wybierz wartości wymiarów: W przypadku danych wielowymiarowych można dostosować analizę poprzez określenie, które wartości wymiarów mają zastosowanie do konkretnej zmiennej metryki dla konkretnej roli. Na przykład, jeśli *sales* jest zmienną metryki a *channel* jest wymiarem z wartościami „retail” i „web”, można określić, że „web” sales jest zmienną wejściową, a „retail” sales jest zmienną przewidywaną. Można również określić podzbiory

wymiarów, jakie będą miały zastosowanie do wszystkich zmiennych metryk użytych w analizie. Na przykład, jeśli *region* jest zmienną wymiaru, która określa region geograficzny, wówczas można ograniczyć analizę do konkretnych regionów.

Wszystkie wartości
Ta opcja określa, że uwzględniane są wszystkie wartości bieżącej zmiennej wymiaru. Jest to opcja domyślna.

Wybierz wartości do uwzględnienia lub wykluczenia
Tej opcji należy użyć, aby określić zbiory wartości dla bieżącej zmiennej wymiaru. Jeśli dla opcji **Tryb** wybrana jest wartość **Uwzględnij**, uwzględniane będą tylko wartości określone na liście **Wybrane wartości**. Jeśli dla opcji **Tryb** wybrana jest wartość **Wyklucz**, wówczas uwzględniane są wszystkie wartości inne niż te określone na liście **Wybrane wartości**.

Wartości, z których można dokonywać wyboru, można filtrować. Wartości, które spełniają warunki filtrowania, są wyświetlane na karcie **Dopasowane**, a wartości, które nie spełniają warunków filtrowania, są wyświetlane na karcie **Niedopasowane** w postaci listy **Niewybrane wartości**. Karta **Wszystkie** zawiera listę wszystkich niewybranych wartości, niezależnie od warunków filtrowania.

- Podczas określania filtru można użyć gwiazdki (*) jako symbolu wieloznacznego.
- Aby wyczyścić bieżący filtr, dla poszukiwanego terminu w oknie dialogowym Filtruj wyświetlane wartości należy wprowadzić pustą wartość.

Obserwacje

Na karcie Zmienne należy użyć ustawień **Obserwacje**, aby określić zmienne definiujące obserwacje.

Obserwacje definiowane jako data/czas

Można określić, czy obserwacje będą definiowane na podstawie zmiennej daty, czasu lub daty/czasu. Oprócz zmiennej, która definiuje obserwacje, należy wybrać odpowiedni przedział czasowy, w którym będą opisywane obserwacje. W zależności od określonego przedziału czasowego można również dokonać innych ustawień, takich jak przedział między obserwacjami (przyrost) lub liczba dni w tygodniu. Poniższe rozważania mają zastosowanie do przedziałów czasowych:

- Wartości **Nieregularny** należy użyć, jeśli obserwacje są nierównomiernie rozmieszczone w czasie, na przykład są wykonywane w chwili, gdy następuje przetwarzanie zamówienia sprzedaży. Jeśli wybrana zostanie opcja **Nieregularny**, należy określić przedział czasowy, jaki będzie używany do analizy, wybierając ustawienia **Przedział czasowy** na karcie Specyfikacja danych.
- Jeśli obserwacje reprezentują datę i czas, a przedział czasowy to godziny, minuty lub sekundy, wówczas należy wybrać opcję **Godziny dziennie**, **Minuty dziennie** lub **Sekundy dziennie**. Jeśli obserwacje reprezentują czas (trwanie) bez odniesienia do daty, a przedział czasowy to godziny, minuty lub sekundy, wówczas należy użyć opcji **Godziny (nieokresowo)**, **Minuty (nieokresowo)** lub **Sekundy (nieokresowo)**.
- Na podstawie wybranego przedziału czasowego procedura może wykryć brakujące obserwacje. Wykrywanie brakujących obserwacji jest konieczne, ponieważ procedura zakłada, że wszystkie obserwacje są równomiernie rozłożone w czasie i że nie ma brakujących obserwacji. Na przykład, jeśli przedziałem czasu są dni, a po dacie 2014-10-27 występuje 2014-10-29, istnieje brakująca obserwacja dla 2014-10-28. Dla wszystkich brakujących obserwacji wprowadzane są wartości. Ustawienia obsługi braków danych można wprowadzić na karcie Specyfikacja danych.
- Określone przedziały czasowe umożliwiają procedurze wykrycie wielu obserwacji w jednym przedziale czasowym, które muszą być zagregowane i które są dopasowane do obserwacji dla granicy przedziału (np. pierwszy dzień miesiąca), dzięki czemu obserwacje będą równomiernie rozmieszczone. Na przykład, jeśli przedziałem czasu są miesiące, to zagregowane zostanie wiele dat z tego samego miesiąca. Ten typ agregacji jest nazywany *grupowaniem*. Domyślnie obserwacje są sumowane podczas grupowania. Można określić inną metodę grupowania, np. średnia z obserwacji; ustawienia **Agregacja i rozkład** można wprowadzić na karcie Specyfikacja danych.
- Przy niektórych przedziałach czasowych istnieją dodatkowe ustawienia umożliwiające zdefiniowanie odstępów w zwykle równomiernie rozmieszczonych przedziałach. Na przykład, jeśli przedziałem czasu są dni, ale istotne są tylko dni robocze, można określić pięciodniowy tydzień rozpoczynający się od poniedziałku.

Obserwacje zdefiniowane jako okresy lub okresy cykliczne

Obserwacje mogą być zdefiniowane przez co najmniej jedną zmienną całkowitą, która reprezentuje okresy lub powtarzające się cyklicznie okresy, aż do dowolnej liczby poziomów cyklicznych. Taka struktura pozwala opisać serie obserwacji, które nie pasują do jednego ze standardowych przedziałów czasowych. Przykładowo, rok fiskalny trwający tylko 10 miesięcy może być opisany przez zmienną cyklu, która reprezentuje lata i zmienną okresu, która reprezentuje miesiące, przy czym długość jednego cyklu wynosi 10.

Zmienne, które określają okresy cykliczne, definiują hierarchię poziomów okresowości, w której najniższy poziom jest definiowany przez zmienną **Okres**. Następny wyższy poziom jest określany przez zmienną cyklu z poziomem 1, po której następuje zmienna cyklu z poziomem 2 itd. Wartości zmiennych dla każdego poziomu, z wyjątkiem najwyższego, muszą być okresowe w odniesieniu do kolejnego najwyższego poziomu. Wartości dla najwyższego poziomu nie mogą być okresowe. Na przykład, dla 10-miesięcznego roku fiskalnego miesiące występują okresowo w latach, ale lata nie są okresowe.

- Długość cyklu na poszczególnych poziomach stanowi okresowość dla kolejnego najniższego poziomu. W przykładzie dot. roku fiskalnego istnieje tylko jeden poziom cyklu, a długość cyklu wynosi 10, ponieważ kolejny najniższy poziom reprezentuje miesiące, a w określonym roku fiskalnym jest 10 miesięcy.
- Należy określić wartość początkową dla każdej zmiennej okresowej, która nie rozpoczyna się od 1. To ustawienie jest niezbędne dla wykrywania braków danych. Na przykład, jeśli zmienna okresowa rozpoczyna się od 2, ale wartość początkowa jest określona jako 1, wówczas procedura zakłada, że istnieje brak danych dla pierwszego okresu w każdym cyklu dla tej zmiennej.

Przedział czasowy analizy

Przedział czasowy, który jest używany do analizy, może różnić się od przedziału czasowego dla obserwacji. Na przykład, jeśli przedział czasowy obserwacji to dni, jako przedział czasowy dla analizy można wybrać miesiące. Wówczas przed zbudowaniem modelu dane agregowane są z dziennych na miesięczne. Można również rozłożyć dane z dłuższego przedziału czasu na krótszy. Przykładowo, jeśli obserwacje są przeprowadzane kwartalnie, wówczas można rozłożyć dane z kwartalnych na miesięczne.

Opcje możliwe do wyboru dla przedziału czasowego, w jakim wykonywana jest analiza, zależą od sposobu zdefiniowania obserwacji oraz wyznaczonego dla nich przedziału czasowego. W szczególności, jeśli obserwacje są zdefiniowane przez okresy cykliczne, wówczas obsługiwana jest tylko agregacja. W takim przypadku przedział czasowy dla analizy musi być większy od przedziału czasowego dla obserwacji lub mu równy.

Przedział czasowy dla analizy można określić w ustawieniach **Przedział czasowy** na karcie Specyfikacja danych. Metoda, którą dane są agregowane lub rozkładane, jest określana w ustawieniach **Agregacja i rozkład** na karcie Specyfikacja danych.

Agregacja i rozkład

Funkcje agregacji

Jeśli przedział czasowy użyty dla analizy jest dłuższy niż przedział czasowy dla obserwacji, dane wejściowe zostają zagregowane. Przykładowo agregacja jest przeprowadzana, kiedy przedział czasowy dla obserwacji to dni, a przedział czasowy dla analizy to miesiące. Dostępne są następujące funkcje agregacji: średnia, suma, dominanta, wartość minimalna lub maksymalna.

Funkcje rozkładu

Jeśli przedział czasowy użyty dla analizy jest krótszy niż przedział czasowy dla obserwacji, dane wejściowe zostają rozłożone. Przykładowo rozkład jest przeprowadzany, kiedy przedział czasowy dla obserwacji to kwartały, a przedział czasowy dla analizy to miesiące. Dostępne są następujące funkcje rozkładu: średnia lub suma.

Funkcje grupujące

Grupowanie jest stosowane, kiedy obserwacje są definiowane przez datę/czas i w tym samym przedziale czasowym występuje wiele obserwacji. Na przykład, jeśli przedział czasowy dla obserwacji to miesiące, wówczas wiele dat z tego samego miesiąca jest grupowanych i tworzone jest ich powiązanie z miesiącem, w którym występują. Dostępne są następujące funkcje grupowania: średnia, suma, dominanta, wartość

minimalna lub maksymalna. Grupowanie jest zawsze przeprowadzane, kiedy obserwacje są zdefiniowane przez datę/czas, a przedział czasowy dla obserwacji jest określony jako Nieregularny.

Uwaga: Chociaż grupowanie jest formą agregacji, jest przeprowadzane przed rozpoczęciem obsługi braków danych, podczas gdy formalna agregacja jest wykonywana po zakończeniu obsługi braków danych. Jeśli przedział czasowy dla obserwacji jest określony jako Nieregularny, agregacja jest wykonywana tylko za pomocą funkcji grupowania.

Agreguj obserwacje przekraczające granice dnia do dnia poprzedniego
Określa, czy obserwacje, których czas przekracza granicę dnia, są agregowane na wartości dla dnia poprzedniego. Przykładowo, dla obserwacji godzinowych trwających osiem godzin dziennie i rozpoczynających się o godzinie 20:00, to ustawienie określi, czy obserwacje od godziny 00:00 do 04:00 będą uwzględniane w zagregowanych wynikach dla poprzedniego dnia. To ustawienie ma zastosowanie tylko w przypadku, kiedy przedział czasowy dla obserwacji to Godziny dziennie, Minuty dziennie lub Sekundy dziennie, a przedział czasowy dla analizy to dni.

Ustawienia niestandardowe dla określonych zmiennych
Funkcje agregacji, rozkładu i grupowania dla zmiennej można określić na podstawie zmiennej. Ustawienia te zastępują domyślne ustawienia funkcji agregacji, rozkładu i grupowania.

Brakujące wartości

Brakujące wartości w danych wejściowych są zastępowane przez wartość podstawianą. Dostępne są zastępujące metody zastępowania:

Interpolacja liniowa
Powoduje zastąpienie braków danych przy wykorzystaniu interpolacji liniowej. W interpolacji używana jest ostatnia ważna wartość przed brakiem danych oraz pierwsza ważna za brakiem. Jeśli pierwsza lub ostatnia obserwacja w szeregu zawiera brakujące wartości, wówczas używane są dwie najbliższe niebrakujące wartości na początku i na końcu serii.

Średnia szeregu
Zastępuje braki danych średnią obliczoną ze wszystkich obserwacji.

Średnia z sąsiednich punktów
Powoduje zastąpienie braków danych średnią z ważnych wartości sąsiednich. Rozpiętość sąsiednich punktów, to liczba poprawnych wartości występujących przed i po brakującej wartości, jakie są wykorzystywane do obliczenia średniej.

Mediana z sąsiednich punktów
Powoduje zastąpienie braków danych medianą ważnych wartości sąsiednich. Rozpiętość sąsiednich punktów, to liczba poprawnych wartości występujących przed i po brakującej wartości, jakie są wykorzystywane do obliczenia mediany.

Trend liniowy
Ta opcja wykorzystuje niebrakujące obserwacje w szeregu do dopasowania prostego modelu regresji liniowej, który jest następnie używany w celu przypisania brakujących wartości.

Inne ustawienia:

Maksymalny procent braków danych (%)
Określa maksymalną wartość procentową braków danych, jaka jest dozwolona w szeregu. Szeregi z większą liczbą braków danych od określonego maksimum są wykluczane z analizy.

Ogólne opcje danych

Maksymalna liczba odrębnych wartości na zmienną wymiaru
To ustawienie dotyczy danych wielowymiarowych i określa maksymalną liczbę odrębnych wartości, jaka jest dozwolona dla dowolnej zmiennej wymiaru. Domyślnie ograniczenie jest ustawione na 10000, ale wartość tę można zmienić na dowolnie dużą liczbę.

Ogólne opcje budowania

Szerokość przedziału ufności (%)
To ustawienie decyduje o przedziałach ufności dla prognoz i parametrów modelu. Można określić dowolną wartość dodatnią mniejszą do 100. Domyślnie ustawiony jest 95-procentowy przedział ufności.

Maksymalna liczba zmiennych wejściowych na zmienną przewidywaną
To ustawienie określa maksymalną liczbę zmiennych wejściowych, jaka jest dozwolona w modelu dla każdej zmiennej przewidywanej. Można określić liczbę całkowitą z zakresu od 1 do 20. Model dla każdej zmiennej przewidywanej zawsze uwzględnia wartości opóźnione samej zmiennej przewidywanej; ustawienie tej wartości na 1 spowoduje, że tylko zmienna wejściowa jest samą zmienną przewidywaną.

Tolerancja modelu
To ustawienie kontroluje proces iteracyjny, jaki jest stosowany do określenia najlepszego zestawu zmiennych wejściowych dla każdej zmiennej przewidywanej. Można określić dowolną wartość większą od zera. Domyślną wartością jest 0,001. Tolerancja modelu jest kryterium zatrzymania wyboru predyktorów. Może wpłynąć na liczbę predyktorów uwzględnionych w ostatecznym modelu. Jeśli jednak zmienna przewidywana bardzo dobrze przewiduje sama siebie, to pozostałe predyktory mogą nie zostać uwzględnione w ostatecznym modelu. Konieczne może być postępowanie metodą prób i błędów (np. jeśli wartość tego ustawienia jest wysoka, można ją zmniejszyć, aby sprawdzić, czy inne predyktory mogą zostać uwzględnione, czy nie).

Próg (%) dla wartości odstającej
Obserwacja jest oznaczana jako odstająca, jeśli obliczone wg modelu prawdopodobieństwo, że jest odstająca, przekracza wyznaczony próg. Można określić wartość z zakresu od 50 do 100.

Liczba opóźnień dla każdej zmiennej wejściowej
To ustawienie określa liczbę opóźnień dla każdej zmiennej wejściowej w modelu dla każdej zmiennej przewidywanej. Domyślnie liczba opóźnień jest określana automatycznie na podstawie przedziału czasowego używanego do analizy. Przykładowo, jeśli przedział czasowy to miesiące (z przyrostem o jeden miesiąc), wówczas liczba opóźnień wynosi 12. Opcjonalnie można jawnie określić liczbę opóźnień. Podana wartość musi być liczbą całkowitą z zakresu od 1 do 20.

Kontynuuj oszacowanie, używając istniejących modeli
Jeśli wygenerowano już modeli przyczynowy szeregów czasowych, tę opcję należy wybrać, aby zamiast budowania nowego modelu, ponownie użyć ustawień kryteriów, jakie są określone dla tego modelu. Można w ten sposób zaoszczędzić czas, ponownie oszacowując i tworząc nową prognozę w oparciu o te same ustawienia modelu, co wcześniej, lecz na podstawie bardziej aktualnych danych.

Prezentacja szeregów

Te opcje określają szeregi (przewidywane lub wejściowe), dla których wyświetlany jest wynik. Zawartość wyniku dla określonego szeregu jest określana przez ustawienia **Opcje wyników**.

Przedstaw zmienne przewidywane w powiązaniu z najlepszymi modelami
Domyślnie wynik jest wyświetlany dla zmiennych przewidywanych, które są powiązane z 10 najlepszymi modelami, określonymi na podstawie wartości R-kwadrat. Można określić inną stałą liczbę najlepszych modeli lub podać odsetek tych modeli. Można również wybrać jedną z następujących miar dopasowania:

R-kwadrat
Miara dobroci dopasowania modelu liniowego, czasami nazywana współczynnikiem determinacji. Jest to część zmienności w zmiennej przewidywanej wyjaśniona przez model. Przyjmuje wartości z przedziału od 0 do 1. Małe wartości statystyki wskazują na słabe dopasowanie modelu do danych.

Pierwiastek procentowego błędu średniokwadratowego
Miara tego, jak bardzo wartość przewidywana przez model może się różnić od wartości obserwowanej szeregu. Jest niezależna od używanych jednostek i tym samym może być używana do porównywania szeregów używających różnych jednostek.

Pierwiastek błędu średniokwadratowego
Pierwiastek kwadratowy obliczany z przeciętnego odchylenia kwadratowego. Mierzy, jak bardzo szereg zależny odbiega od poziomu przewidywanego przez model; miara wyrażona w jednostkach używanych przez szereg zależny.

BIC Bayesowskie kryterium informacyjne. Miara wybierania i porównywania modeli tworzonych na podstawie zredukowanego -2 logarytmu wiarygodności. Mniejsze wartości oznaczają lepszy model. Wartość BIC „karze" także modele przeparametryzowane (na przykład złożone modele z dużą liczbą danych wejściowych), jednak silniej niż miara AIC.

AIC Kryterium informacyjne Akaike. Miara wybierania i porównywania modeli tworzonych na podstawie zredukowanego -2 logarytmu wiarygodności. Mniejsze wartości oznaczają lepszy model. Wartość AIC „karze" modele przeparametryzowane (na przykład złożone modele z dużą liczbą danych wejściowych).

Określ poszczególne szeregi

Można określić poszczególne szeregi, dla których potrzebny jest wynik.

- W przypadku danych kolumnowych należy określić zmienne, które zawierają wymagane szeregi. Kolejność określonych zmiennych definiuje kolejność, w jakiej są wyświetlane w wyniku.
- W przypadku danych wielowymiarowych należy określić konkretny szereg, dodając wpis do siatki dla zmiennej metryki, która zawiera szereg. Następnie należy określić wartości zmiennych wymiaru, które definiują szereg.
  - Można wprowadzić wartość dla każdej zmiennej wymiaru bezpośrednio do siatki lub wybrać ją z listy dostępnych wartości wymiarów. Aby wybrać wartość z listy, należy kliknąć przycisk wielokropka w komórce wybranego wymiaru. Ta czynność otworzy podrzędne okno dialogowe Wybierz zmienną wymiaru.
  - Listę wartości wymiaru w podrzędnym oknie dialogowym Wybierz zmienną wymiaru można przeszukać, klikając ikonę lornetki i wprowadzając poszukiwany termin. Spacje są traktowane jako część poszukiwanego terminu. Gwiazdka (*) w poszukiwanym terminie nie oznacza symbolu wieloznacznego.
  - Kolejność szeregów w siatce definiuje kolejność, w jakiej są wyświetlane w wyniku.

Wynik dla danych kolumnowych i danych wielowymiarowych jest ograniczony do 30 szeregów. Ten limit uwzględnia poszczególne szeregi (wejściowe i przewidywane) określone przez użytkownika i przewidywane powiązane z najlepszymi modelami. Szeregi określone przez użytkownika mają pierwszeństwo w stosunku do przewidywanych powiązanych z najlepszymi modelami.

Opcje wyników

Te opcje określają zawartość wyniku. Opcje z grupy **Wyniki dla zmiennych przewidywanych** generują wynik dla zmiennych przewidywanych, które są powiązane z najlepszymi modelami w ustawieniach **Prezentacja szeregów**. Opcje z grupy **Wyniki dla szeregów czasowych** generują wyniki dla poszczególnych szeregów, które zostały określone w ustawieniach **Prezentacja szeregów**.

System modelu ogólnego

Wyświetla graficzną reprezentację relacji przyczynowych pomiędzy szeregami w systemie modelu. Tabele statystyk dopasowania modelu i wartości odstających dla wyświetlanych przewidywanych są uwzględniane jako część element wyniku. Jeśli ta opcja zostanie wybrana w grupie **Wyniki dla szeregów czasowych**, tworzony jest osobny element wyniku dla każdego pojedynczego szeregu, jaki został określony w ustawieniach **Prezentacja szeregów**.

Relacje przyczynowe pomiędzy szeregami mają przypisany poziom istotności, gdzie niższy poziom istotności oznacza bardziej istotny związek. Istnieje możliwość ukrycia relacji, których poziom istotności jest większy od określonej wartości.

Statystyki dopasowania modelu i wartości odstające
Tabele statystyk dopasowania modelu i wartości odstających dla szeregów przewidywanych, jakie zostały

wybrane do wyświetlenia. Tabele te zawierają te same informacje, co tabele w wizualizacji Ogólny system modelu. Obsługują one wszystkie standardowe funkcje przestawiania i edytowania tabel.

Efekty i parametry modelu
Tabele testów efektów modelu i parametrów modelu dla szeregów przewidywanych, jakie zostały wybrane do wyświetlenia. Testy efektów modelu obejmują statystykę F i powiązaną wartość istotności dla każdej zmiennej wejściowej uwzględnianej w modelu.

Diagram wpływu

Przedstawia graficzną reprezentację relacji przyczynowych pomiędzy szeregami będącymi przedmiotem zainteresowania a innymi szeregami, na które one wpływają lub które mają na nie wpływ. Szeregi, które wpływają na szeregi będące przedmiotem zainteresowania, są nazywane *przyczynami*. Wybór opcji **Efekty** powoduje wygenerowanie diagramu wpływu, który jest inicjowany w celu wyświetlenia efektów. Wybór opcji **Przyczyny** powoduje wygenerowanie diagramu wpływu, który jest inicjowany w celu wyświetlenia przyczyn. Wybór opcji **Przyczyny i skutki** powoduje wygenerowanie dwóch osobnych diagramów wpływu; jeden jest inicjowany dla przyczyn, a drugi dla efektów. Przyczyny i efekty można w sposób interaktywny przełączać w elemencie wyniku, który wyświetla diagram wpływu.

Można określić liczbę poziomów przyczyn i efektów do wyświetlenia, gdzie pierwszy poziom stanowi szereg będący przedmiotem zainteresowania. Każdy dodatkowy poziom przedstawia bardziej pośrednie przyczyny lub efekty szeregu będącego przedmiotem zainteresowania. Przykładowo, trzeci poziom wyświetlanych efektów składa się z szeregów z drugiego poziomu stanowiących bezpośrednie dane wejściowe. Na szeregi z trzeciego poziomu szeregi będące przedmiotem zainteresowania mają więc pośredni wpływ, ponieważ szeregi te stanowią bezpośrednie dane wejściowe dla szeregów z drugiego poziomu.

Wykres szeregu
Wykresy obserwowanych i przewidywanych wartości dla szeregów przewidywanych, które zostały wybrane do wyświetlenia. Jeśli żądanie obejmowało prognozy, wykres przedstawia również wartości prognozowane i przedziały ufności dla prognoz.

Wykres reszt
Wykresy reszt modelu dla szeregów przewidywanych, jakie zostały wybrane do wyświetlenia.

Najlepsze zmienne wejściowe
Wykresy poszczególnych wyświetlanych przewidywanych w czasie wraz z 3 najlepszymi zmiennymi wejściowymi dla przewidywanych. Najlepsze zmienne wejściowe to zmienne z najniższą wartością istotności. W celu dostosowania różnych skal zmiennych wejściowych i przewidywanych na osi y prezentowane są oceny z z każdego szeregu.

Tabela prognoz
Tabele zawierające przewidziane wartości i przedziały ufności dla tych prognoz określane dla przewidywanych szeregów, jakie zostały wybrane do wyświetlenia.

Analiza podstawowych przyczyn wartości odstających
Określa, które szeregi z największym prawdopodobieństwem będą powodowały powstawanie poszczególnych wartości odstających w szeregu będącym przedmiotem zainteresowania. Analiza podstawowych przyczyn wartości odstających jest przeprowadzana dla każdego szeregu przewidywanego, który jest uwzględniony na liście poszczególnych szeregów w ustawieniach **Prezentacja szeregów**.

Raport

Interaktywna tabela i wykres wartości odstających
Tabela i wykres odstających wartości oraz podstawowych przyczyn powstawania tych wartości dla każdego szeregu będącego przedmiotem zainteresowania. Każda wartość jest zapisywana w tabeli w osobnym wierszu. Wykres ten przedstawia diagram wpływu. Wybór wiersza w tabeli powoduje wyróżnienie ścieżki (na diagramie wpływu) prowadzącej od szeregu będącego przedmiotem zainteresowania do szeregu, który z największym prawdopodobieństwem był przyczyną powstania powiązanej wartości odstającej.

Tabela przestawna wartości odstających
Tabela odstających wartości oraz głównych przyczyn powstawania tych wartości dla

każdego szeregu będącego przedmiotem zainteresowania. Ta tabela zawiera te same informacje, co tabela na ekranie interaktywnym. Tabela ta obsługuje wszystkie standardowe funkcje przestawiania i edytowania tabel.

Poziomy przyczynowe
Użytkownik może określić liczbę poziomów, jakie będą uwzględnione w wyszukiwaniu podstawowych przyczyn. Informacje na temat poziomów, jakie są tutaj używane, są takie same, jak w przypadku diagramów wpływu.

Dopasowanie wszystkich modeli
Histogram dopasowania modelu dla wszystkich modeli i wybranych statystyk dopasowania. Dostępne są następujące statystyki dopasowania:

R-kwadrat
Miara dobroci dopasowania modelu liniowego, czasami nazywana współczynnikiem determinacji. Jest to część zmienności w zmiennej przewidywanej wyjaśniona przez model. Przyjmuje wartości z przedziału od 0 do 1. Małe wartości statystyki wskazują na słabe dopasowanie modelu do danych.

Pierwiastek procentowego błędu średniokwadratowego
Miara tego, jak bardzo wartość przewidywana przez model może się różnić od wartości obserwowanej szeregu. Jest niezależna od używanych jednostek i tym samym może być używana do porównywania szeregów używających różnych jednostek.

Pierwiastek błędu średniokwadratowego
Pierwiastek kwadratowy obliczany z przeciętnego odchylenia kwadratowego. Mierzy, jak bardzo szereg zależny odbiega od poziomu przewidywanego przez model; miara wyrażona w jednostkach używanych przez szereg zależny.

BIC Bayesowskie kryterium informacyjne. Miara wybierania i porównywania modeli tworzonych na podstawie zredukowanego -2 logarytmu wiarygodności. Mniejsze wartości oznaczają lepszy model. Wartość BIC „karze" także modele przeparametryzowane (na przykład złożone modele z dużą liczbą danych wejściowych), jednak silniej niż miara AIC.

AIC Kryterium informacyjne Akaike. Miara wybierania i porównywania modeli tworzonych na podstawie zredukowanego -2 logarytmu wiarygodności. Mniejsze wartości oznaczają lepszy model. Wartość AIC „karze" modele przeparametryzowane (na przykład złożone modele z dużą liczbą danych wejściowych).

Wartości odstające w czasie
Wykres słupkowy liczby wartości odstających dla wszystkich zmiennych przewidywanych i każdego przedziału czasowego w okresie oszacowania.

Transformacje szeregów
Tabela wszystkich transformacji, jakie zostały przeprowadzone dla szeregu w systemie modelu. Możliwe transformacje to podstawianie brakujących wartości, agregacja i rozkład.

Okres estymacji

Domyślnie okres estymacji zaczyna się od czasu z najwcześniejszą obserwacją, a kończy w czasie z najpóźniejszą obserwacją we wszystkich szeregach.

Wyznaczony przez czas początkowy i końcowy
Można określić datę rozpoczęcia i zakończenia okresu estymacji lub można określić tylko datę rozpoczęcia lub tylko datę zakończenia. Jeśli rozpoczęcie lub zakończenie okresu estymacji zostanie pominięte, użyta będzie wartość domyślna.

- Jeśli obserwacje są zdefiniowane przez określenie zmiennej daty/czasu, wówczas wartości rozpoczęcia i zakończenia okresu należy wprowadzić w takim samym formacie, jaki został użyty dla zmiennej daty/czasu.
- W przypadku obserwacji definiowanych na podstawie okresów cyklicznych należy określić wartość dla każdej zmiennej okresu cyklicznego. Każda zmienna jest wyświetlana w osobnej kolumnie.

Wyznaczony przez najwcześniejszy lub najpóźniejszy przedział czasowy

Definiuje okres estymacji jako określoną liczbę przedziałów czasowych, która rozpoczyna się od najwcześniejszego przedziału czasowego lub kończy na najpóźniejszym przedziale czasowym określonym w danych, z opcjonalnym przesunięciem. W tym kontekście przedział czasowy odnosi się do przedziału czasowego dla analizy. Przykładowo, załóżmy, że obserwacje są przeprowadzane miesięcznie, ale przedział czasowy dla analizy to kwartały. Określenie wartości **Najpóźniejszy** i wartości 24 dla opcji **Liczba przedziałów czasowych** będzie oznaczało ostatnie 24 kwartały.

Opcjonalnie można wykluczyć określoną liczbę przedziałów czasowych. Przykładowo, określenie ostatnich 24 przedziałów czasowych i 1 do wykluczenia oznacza, że okres oszacowania składa się z 24 przedziałów, które poprzedzają ostatni.

Opcje modelu

Nazwa modelu

Można określić niestandardową nazwę modelu lub zaakceptować nazwę wygenerowaną automatycznie, czyli *TCM*.

Prognoza

Opcja **Rozszerz rekordy na przedziały z przyszłości** pozwala ustawić liczbę przedziałów do prognozowania poza koniec okresu estymacji. Przedział czasowy jest w tym przypadku przedziałem czasowym dla analizy, określonym na karcie Specyfikacja danych. Z chwilą wywołania prognoz następuje automatyczna budowa modeli autoregresji dla szeregów wejściowych niebędących jednocześnie wartościami przewidywanymi. Modele te są następnie używane do generowania wartości dla tych szeregów wejściowych w okresie prognozy. Dla tego ustawienia nie ma maksymalnego limitu.

Wyjście interaktywne

Wynik modelowania przyczynowego szeregów czasowych zawiera interaktywne obiekty wynikowe. Interaktywne właściwości są dostępne po aktywowaniu (dwukrotnym kliknięciu) obiektu wykresu w oknie wynikowym.

System modelu ogólnego

Wyświetla relacje przyczynowe między szeregami w systemie modelu. Wszystkie linie łączące konkretną zmienną przewidywaną z jej zmiennymi wejściowymi mają ten sam kolor. Grubość linii oznacza istotność związku przyczynowego, przy czym grubsze linie oznaczają bardziej istotny związek. Zmienne wejściowe, które nie są jednocześnie przewidywanymi, oznaczone są czarnym kwadratem.

- Istnieje możliwość wyświetlenia relacji najlepszych modeli, określonego szeregu, wszystkich szeregów lub modeli bez danych wejściowych. Najlepsze modele są tymi, które spełniają kryteria najlepiej dopasowanych modeli wprowadzone na stronie **Szereg do wyświetlenia**.
- Można wygenerować diagramy wpływu dla jednego lub wielu szeregów, wybierając nazwę szeregu w tabeli, klikając ją prawym przyciskiem myszy i wybierając z menu kontekstowego polecenie **Utwórz diagram wpływu**.
- Istnieje możliwość ukrycia relacji przyczynowych, których poziom istotności jest większy od określonej wartości. Mniejszy poziom istotności oznacza bardziej znaczącą relację przyczynową.
- Można wyświetlić relacje dla konkretnego szeregu, wybierając w tabeli nazwę szeregu, klikając ją prawym przyciskiem i wybierając z menu kontekstowego polecenie **Podświetl relacje dla szeregu**.

Diagram wpływu

Przedstawia graficzną reprezentację relacji przyczynowych pomiędzy szeregami będącymi przedmiotem zainteresowania a innymi szeregami, na które one wpływają lub które mają na nie wpływ. Szeregi, które wpływają na szeregi będące przedmiotem zainteresowania, są nazywane *przyczynami*.

- Analizowany szereg można zmienić, wpisując nazwę nowego potrzebnego szeregu. Dwukrotne kliknięcie dowolnego węzła na diagramie wpływu zmieni analizowany szereg na ten, który jest powiązany z danym węzłem.
- Istnieje możliwość przełączenia obrazu między przyczynami i skutkami oraz zmiany liczby wyświetlanych poziomów przyczyn lub skutków.
- Jednokrotne kliknięcie dowolnego węzła wyświetla szczegółowy diagram kolejności dla szeregu powiązanego z węzłem.

Analiza podstawowych przyczyn wartości odstających

Określa, które szeregi z największym prawdopodobieństwem będą powodowały powstawanie poszczególnych wartości odstających w szeregu będącym przedmiotem zainteresowania.

- Podstawową przyczynę dowolnej wartości odstającej można określić, wybierając jej wiersz w tabeli Wartości odstające. Można ją także wyświetlić, klikając ikonę wartości odstającej w wykresie sekwencji.
- Jednokrotne kliknięcie dowolnego węzła wyświetla szczegółowy diagram kolejności dla szeregu powiązanego z węzłem.

Jakość modelu ogólnego

Histogram dopasowania modelu dla wszystkich modeli, w szczególności dla statystyki dopasowania. Kliknięcie słupka na wykresie słupkowym filtruje wykres punktowy tak, aby wyświetlane były wyłącznie modele powiązane z wybranym słupkiem. Model dla określonej serii docelowej na wykresie punktowym można znaleźć, określając nazwę szeregu.

Rozkład wartości odstających

Wykres słupkowy liczby wartości odstających dla wszystkich zmiennych przewidywanych i każdego przedziału czasowego w okresie oszacowania. Kliknięcie słupka na wykresie słupkowym filtruje wykres punktowy tak, aby wyświetlane były wyłącznie wartości odstające powiązane z wybranym słupkiem.

Model użytkowy TCM

Operacja modelowania przyczynowego szeregów czasowych umożliwia tworzenie pewnej liczby nowych zmiennych o prefiksie $TCM- zgodnie z danymi w poniższej tabeli.

Tabela 27. Nowe zmienne tworzone w ramach operacji modelowania TCM

Nazwa zmiennej	Opis
$TCM-*colname*	Wartość prognozowana przez model dla każdego przewidywanego szeregu.
$TCMLCI-*colname*	Dolne przedziały ufności dla każdego prognozowanego szeregu.
$TSUCI-*colname*	Górne przedziały ufności dla każdego prognozowanego szeregu.
$TCMResidual-*colname*	Wartość reszty modelu dla każdej kolumny generowanych danych modelu.

Ustawienia modelu użytkowego TCM

Karta Ustawienia zawiera dodatkowe opcje dla modelu użytkowego TCM.

Prognoza

Opcja **Rozszerz rekordy na przedziały z przyszłości** pozwala ustawić liczbę przedziałów do prognozowania poza koniec okresu estymacji. Przedział czasu jest w tym przypadku okresem czasu analizy, określonym na karcie Specyfikacja danych węzła TCM. Z chwilą wywołania prognoz następuje automatyczna budowa modeli autoregresji dla szeregów wejściowych niebędących jednocześnie wartościami przewidywanymi. Modele te są następnie używane do generowania wartości dla tych szeregów wejściowych w okresie prognozy.

Udostępnij do oceniania

Utwórz dla każdego modelu nowe zmienne do oceniania. Umożliwia określenie nowych zmiennych, które mają zostać utworzone dla każdego ocenianego modelu.

- **Reszty szumów.** Zaznaczenie tej opcji powoduje utworzenie nowej zmiennej (z prefiksem domyślnym $TCM-) dla reszt modelu dla każdej zmiennej przewidywanej, wraz z sumą tych wartości.
- **Górny i dolny przedział ufności.** Zaznaczenie tej opcji powoduje utworzenie nowych zmiennych (o prefiksie domyślnym $TCM-) odpowiednio dla dolnego i górnego przedziału ufności, wraz z sumami tych wartości.

Zmienne przewidywane uwzględnione w ocenie. Umożliwia wybór dostępnych wartości przewidywanych do uwzględnienia w ocenie modelu.

Scenariusze modelowania przyczynowego szeregów czasowych

Procedura Scenariusze modelowania przyczynowego szeregów czasowych uruchamia zdefiniowane przez użytkownika scenariusze dla systemu modelowania przyczynowego szeregów czasowych wraz z danymi z aktywnego zbioru danych. Zmienna *scenario* jest definiowana przez szereg czasowy będący szeregiem źródłowym (*root series*) oraz przez zestaw zdefiniowanych przez użytkownika wartości dla tego szeregu w podanym zakresie czasu. Podane wartości są następnie używane do generowania predykcji dla szeregów czasowych, na które wpływa szereg źródłowy. Procedura wymaga pliku systemu modelu utworzonego w procedurze Modelowanie przyczynowe szeregów czasowych. Zakłada się, że aktywny zbiór danych to te same dane, które były używane do utworzenia pliku systemu modelu.

Przykład

Korzystając z procedury modelowania przyczynowego szeregów czasowych, osoba odpowiedzialna za decyzje biznesowe zidentyfikowała kluczową metrykę wpływającą na pewną liczbę ważnych wskaźników wydajności. Metryka poddaje się kontroli, dlatego osoba odpowiedzialna za decyzje zamierza zbadać wpływ różnych zestawów wartości metryki na dane w następnym kwartale. Można to łatwo zrobić, wczytując plik systemu modelu do procedury modelowania przyczynowego szeregów czasowych i określając zestaw wartości dla kluczowej metryki.

Definiowanie okresu scenariusza

Okres scenariusza to okres, w którym użytkownik określa wartości używane do uruchamiania scenariuszy. Powinien on rozpoczynać się przed okresem estymacji lub po jego zakończeniu. Opcjonalnie można określić predykcję poza koniec okresu scenariusza. Domyślnie predykcje są generowane po zakończeniu okresu scenariusza. Wszystkie scenariusze korzystają z tego samego okresu scenariusza oraz ze specyfikacji określających, jak daleko ma sięgać predykcja.

Uwaga: Predykcje rozpoczynają się w pierwszym okresie czasu po rozpoczęciu okresu scenariusza. Na przykład, jeśli okres scenariusza rozpoczyna się z datą 2014-11-01, zaś przedział czasu jest mierzony w miesiącach, wówczas pierwsza predykcja ma datę 2014-12-01.

Wyznaczony przez początek, koniec i czas predykcji

- Jeśli obserwacje są zdefiniowane przez określenie zmiennej daty/czasu, wówczas wartości rozpoczęcia, zakończenia i predykcji należy wprowadzić w takim samym formacie, jaki został użyty dla zmiennej daty/czasu. Wartości zmiennych daty/czasu są dopasowane do rozpoczęcia powiązanego przedziału czasowego. Na przykład, jeśli przedział czasowy dla analizy to miesiące, wówczas wartość 10/10/2014 jest korygowana na 10/01/2014, co oznacza rozpoczęcie miesiąca.
- W przypadku obserwacji definiowanych na podstawie okresów cyklicznych należy określić wartość dla każdej zmiennej okresu cyklicznego. Każda zmienna jest wyświetlana w osobnej kolumnie.

Wyznaczony przez przedziały zależne od końca okresu estymacji

Definiuje rozpoczęcie i zakończenie w odniesieniu do liczby przedziałów czasowych odpowiadających końcowi okresu estymacji, gdzie przedział czasowy jest przedziałem czasowym dla analizy. Koniec okresu estymacji jest definiowany jako przedział czasowy 0. Przedziały czasowe przed zakończeniem okresu estymacji mają ujemne wartości, a przedziały po zakończeniu okresu estymacji mają wartości dodatnie. Można również określić liczbę przedziałów dla predykcji po zakończeniu okresu scenariusza. Domyślną wartością jest 0.

Załóżmy na przykład, że przedział czasowy dla analizy to miesiące, jako początek przedziału określono wartość 1, wartość 3 dla zakończenia przedziału oraz wartość 1 dla określenia, jak daleko poza koniec okresu scenariusza ma sięgać predykcja. Okres scenariusza jest wówczas ustawiony na 3 miesiące po zakończeniu okresu estymacji. Predykcje są generowane dla drugiego i trzeciego miesiąca okresu scenariusza i dla jeszcze 1 miesiąca poza koniec okresu scenariusza.

Dodawanie scenariuszy i grup scenariuszy

Na karcie Scenariusze określane są scenariusze, jakie mają zostać uruchomione. Aby zdefiniować scenariusz, konieczne jest zdefiniowanie najpierw okresu scenariusza poprzez kliknięcie przycisku **Zdefiniuj okres scenariusza**. Scenariusze i grupy scenariuszy (dotyczy tylko danych wielowymiarowych) są tworzone poprzez kliknięcie powiązanego przycisku **Dodaj scenariusz** lub **Dodaj grupę scenariuszy**. Wybierając określony scenariusz lub grupę scenariuszy w powiązanej siatce, można je edytować, utworzyć kopię lub usunąć.

Dane kolumnowe

W kolumnie **Zmienna źródłowa** w siatce określane są zmienne szeregów czasowych, których wartości są zastępowane przez wartości scenariusza. W kolumnie **Wartości scenariusza** wyświetlane są określone wartości scenariusza w kolejności od najwcześniejszych do najpóźniejszych. Jeśli wartości scenariusza są zdefiniowane przez wyrażenie, wówczas w kolumnie wyświetlane jest wyrażenie.

Dane wielowymiarowe

Indywidualne scenariusze

W każdym wierszu w siatce Indywidualne scenariusze określane są szeregi czasowe, których wartości są zastępowane przez określone wartości scenariusza. Szereg jest definiowany przez kombinację zmiennej, jaka została określona w kolumnie **Metryka źródłowa** oraz określonej wartości dla każdej zmiennej wymiaru. Zawartość kolumny **Wartości scenariusza** jest taka sama, jak w przypadku danych kolumnowych.

Grupy scenariuszy

Grupa scenariusza definiuje zestaw scenariuszy, które zostały utworzone w oparciu o jedną zmienną metryki źródłowej i wiele zbiorów wartości wymiaru. Każdy zbiór wartości wymiaru (jedna wartość dla zmiennej wymiaru) dla określonej zmiennej metryki definiuje szereg czasowy. Pojedynczy scenariusz jest wówczas generowany dla każdego szeregu czasowego, którego wartości są następnie zastępowane przez wartości scenariusza. Wartości scenariusza dla grupy scenariuszy są określane przez wyrażenie, które jest następnie stosowane do każdego szeregu czasowego w grupie.

W kolumnie **Liczba szeregów** wyświetlana jest liczba zbiorów wartości wymiarów, które są powiązane z grupą scenariuszy. Wyświetlana wartość może być większa niż rzeczywista liczba szeregów, które są powiązane z grupą scenariusza (jeden szereg dla zbioru). Taka sytuacja ma miejsce, jeśli niektóre z określonych kombinacji wartości wymiarów nie odpowiadają szeregom zawartym w metryce źródłowej dla grupy.

Przykładem grupy scenariusza może być zmienna metryki *advertising* oraz dwie zmienne wymiarów *region* i *brand*. Można zdefiniować grupę scenariusza dla zmiennej *advertising* stanowiącej metrykę źródłową i obejmującej wszystkie kombinacje z innych *region* i *brand*. Następnie można określić wartość advertising*1.2 jako wyrażenie do zbadania efektu zwiększenia wartości *advertising* o 20 procent dla wszystkich szeregów czasowych, jakie są powiązane ze zmienną *advertising*. Jeśli istnieją 4 wartości *region* i 2 wartości *brand*, wówczas istnieje 8 takich szeregów czasowych i 8 scenariuszy zdefiniowanych przez grupę.

Definicja scenariusza: Ustawienia dla definiowania scenariusza zależą od tego, czy dane są danymi kolumnowymi czy wielowymiarowymi.

Szereg początkowy

Określa szereg początkowy dla scenariusza. Każdy scenariusz jest oparty na jednym szeregu początkowym. W przypadku danych kolumnowych wybierana jest zmienna, która definiuje szereg początkowy. W przypadku danych wielowymiarowych należy określić szereg początkowy, dodając wpis do siatki dla zmiennej metryki, która zawiera szereg. Następnie należy określić wartości zmiennych wymiaru, które definiują szereg źródłowy. Następujące opcje mają zastosowanie podczas określania wartości wymiarów:

- Można wprowadzić wartość dla każdej zmiennej wymiaru bezpośrednio do siatki lub wybrać ją z listy dostępnych wartości wymiarów. Aby wybrać wartość z listy, należy kliknąć przycisk wielokropka w komórce wybranego wymiaru. Ta czynność otworzy podrzędne okno dialogowe Wybierz zmienną wymiaru.
- Listę wartości wymiaru w podrzędnym oknie dialogowym Wybierz zmienną wymiaru można przeszukać, klikając ikonę lornetki i wprowadzając poszukiwany termin. Spacje są traktowane jako część poszukiwanego terminu. Gwiazdka (*) w poszukiwanym terminie nie oznacza symbolu wieloznacznego.

Określ zmienione zmienne przewidywane

Tej opcji należy użyć, jeśli znane są konkretne zmienne przewidywane, na które wpływają szeregi czasowe oraz użytkownik zamierza zbadać efekty tylko dla tych zmiennych przewidywanych. Domyślnie zmienne przewidywane, na które wpływa szereg początkowy, są ustalane automatycznie. Można określić zasięg szeregu, na który ma wpływ scenariusz, używając opcji na karcie Opcje.

W przypadku danych kolumnowych należy wybrać dowolne zmienne przewidywane. W przypadku danych wielowymiarowych należy określić szereg przewidywany, dodając wpis do siatki dla przewidywanej zmiennej metryki, która zawiera szereg. Domyślnie uwzględniane są wszystkie szeregi, które są zawarte w określonej zmiennej metryki. Można dostosować zbiór uwzględnionych szeregów poprzez dostosowanie wartości dla co najmniej jednej zmiennej wymiaru. Aby dostosować uwzględnione wartości wymiaru, należy kliknąć przycisk wielokropka dla wybranego wymiaru. Ta czynność otworzy okno dialogowe Wybierz wartości wymiarów.

W kolumnie **Liczebność szeregu** (dla danych wielowymiarowych) wyświetlana jest liczba zestawów wartości wymiarów, jakie są aktualnie określone dla powiązanej metryki przewidywanej. Wyświetlana wartość może być większa niż rzeczywista liczba szeregów przewidywanych (jeden szereg na zestaw). Taka sytuacja ma miejsce, jeśli niektóre z określonych kombinacji wartości wymiarów nie odpowiadają szeregom objętym przez powiązaną metrykę przewidywaną.

Identyfikator scenariusza
Każdy scenariusz musi mieć unikalny identyfikator. Identyfikator jest wyświetlany w wyniku, który jest powiązany ze scenariuszem. Nie ma żadnych ograniczeń, poza unikalnością, dla wartości identyfikatora.

Określ wartości scenariusza dla szeregu źródłowego
Tej opcji należy użyć, aby określić jawne wartości dla szeregów początkowych w okresie scenariusza. Należy określić wartość liczbową dla każdego przedziału czasowego, jaki został wymieniony w siatce. Można uzyskać wartości szeregów początkowych (rzeczywistych i przewidywanych) dla każdego przedziału z okresu scenariusza, klikając opcje **Odczyt**, **Prognoza** lub **Odczyt\Prognoza**.

Określ wyrażenie dla wartości scenariusza dla szeregu źródłowego
Można zdefiniować wyrażenie do obliczania wartości dla szeregów początkowych w okresie scenariusza. Można wprowadzić wyrażenie bezpośrednio lub kliknąć przycisk kalkulatora i utworzyć wyrażenie za pomocą konstruktora wyrażeń wartości scenariusza.

- Wyrażenie może zawierać dowolną zmienną przewidywaną lub przewidywaną w systemie modelu.
- Jeśli okres scenariusza będzie rozszerzony poza istniejące dane, wówczas wyrażenie jest stosowane do wartości prognozowanych dla zmiennych z wyrażenia.
- W przypadku danych wielowymiarowych każda zmienna w wyrażeniu określa szereg czasowy, który jest definiowany przez zmienną oraz wartości wymiaru, jakie zostały określone dla metryki źródłowej. Są to szeregi czasowe, które są używane do oszacowania wyrażenia.

Załóżmy na przykład, że zmienną źródłową jest *advertising*, a wyrażenie to advertising*1.2. Wartości dla *advertising*, jakie zostały użyte w scenariuszu, reprezentują 20 procent wzrostu wobec istniejących wartości.

Uwaga: Scenariusze są tworzone poprzez kliknięcie **Dodaj scenariusz** na karcie Scenariusze.

Wybierz wartości wymiarów: W przypadku danych wielowymiarowych można dostosować wartości wymiarów, które definiują zmienne przewidywane, na które wpływa scenariusz lub grupa scenariuszy. Można również dostosować wartości wymiarów, które definiują zbiór szeregów początkowych dla grupy scenariusza.

Wszystkie wartości
Ta opcja określa, że uwzględniane są wszystkie wartości bieżącej zmiennej wymiaru. Jest to opcja domyślna.

Wybierz wartości
Tej opcji należy użyć, aby określić zbiory wartości dla bieżącej zmiennej wymiaru. Wartości, z których można dokonywać wyboru, można filtrować. Wartości, które spełniają warunki filtrowania, są wyświetlane na karcie **Dopasowane**, a wartości, które nie spełniają warunków filtrowania, są wyświetlane na karcie **Niedopasowane** w postaci listy **Niewybrane wartości**. Karta **Wszystkie** zawiera listę wszystkich niewybranych wartości, niezależnie od warunków filtrowania.

- Podczas określania filtru można użyć gwiazdki (*) jako symbolu wieloznacznego.
- Aby wyczyścić bieżący filtr, dla poszukiwanego terminu w oknie dialogowym Filtruj wyświetlane wartości należy wprowadzić pustą wartość.

Aby dostosować wartości wymiaru dla zmienionych zmiennych przewidywanych:

1. W oknie dialogowym Definicja scenariusza lub Definicja grupy scenariuszy należy wybrać metrykę przewidywaną, dla której wartości wymiaru mają zostać dostosowane.
2. Należy kliknąć przycisk wielokropka w kolumnie wymiaru, jaki ma być dostosowany.

Aby dostosować wartości wymiaru dla szeregu początkowego w grupie scenariuszy:

1. W oknie dialogowym Definicja grupy scenariuszy należy kliknąć przycisk wielokropka (w siatce szeregów początkowych) dla wymiaru, jaki ma być dostosowany.

Definicja grupy scenariuszy:

Szereg początkowy

Określa zbiór szeregów początkowych dla grupy scenariuszy. Pojedynczy scenariusz jest generowany dla każdego szeregu czasowego w zbiorze. Należy określić szereg początkowy, dodając wpis do siatki dla zmiennej metryki, która zawiera wybrany szereg. Następnie należy określić wartości zmiennych wymiaru, które definiują zbiór. Domyślnie uwzględniane są wszystkie szeregi, które są zawarte w określonej zmiennej metryki źródłowej. Można dostosować zbiór uwzględnionych szeregów poprzez dostosowanie wartości dla co najmniej jednej zmiennej wymiaru. Aby dostosować uwzględnione wartości wymiaru, należy kliknąć przycisk wielokropka dla wybranego wymiaru. Ta czynność otworzy okno dialogowe Wybierz wartości wymiarów.

W kolumnie **Liczebność szeregu** wyświetlana jest liczba zestawów wartości wymiarów, jakie są aktualnie uwzględniane dla powiązanej metryki źródłowej. Wyświetlana wartość może być większa niż rzeczywista liczba szeregów dla grupy scenariusza (jeden szereg dla zbioru). Taka sytuacja ma miejsce, jeśli niektóre z określonych kombinacji wartości wymiarów nie odpowiadają szeregom objętym przez metrykę źródłową.

Określ zmieniony szereg przewidywany

Tej opcji należy użyć, jeśli znane są konkretne zmienne przewidywane, na które wpływa zbiór szeregów początkowych oraz użytkownik zamierza zbadać efekty tylko dla tych zmiennych przewidywanych. Domyślnie zmienne przewidywane, na które wpływa każdy szereg początkowy, są ustalane automatycznie. Można określić zasięg szeregu, na który ma wpływ każdy pojedynczy scenariusz, używając opcji na karcie Opcje.

Należy określić szereg przewidywany, dodając wpis do siatki dla przewidywanej zmiennej metryki, która zawiera szereg. Domyślnie uwzględniane są wszystkie szeregi, które są zawarte w określonej zmiennej metryki. Można dostosować zbiór uwzględnionych szeregów poprzez dostosowanie wartości dla co najmniej jednej zmiennej wymiaru. Aby dostosować uwzględnione wartości wymiaru, należy kliknąć przycisk wielokropka dla wybranego wymiaru. Ta czynność otworzy okno dialogowe Wybierz wartości wymiarów.

W kolumnie **Liczebność szeregu** wyświetlana jest liczba zestawów wartości wymiarów, jakie są aktualnie określone dla powiązanej metryki przewidywanej. Wyświetlana wartość może być większa niż rzeczywista liczba szeregów przewidywanych (jeden szereg na zestaw). Taka sytuacja ma miejsce, jeśli niektóre z określonych kombinacji wartości wymiarów nie odpowiadają szeregom objętym przez powiązaną metrykę przewidywaną.

Przedrostek identyfikatora scenariusza

Każda grupa scenariuszy musi mieć unikalny przedrostek. Przedrostek służy do budowania identyfikatora, jaki jest wyświetlany w wyniku powiązanym z poszczególnymi pojedynczymi scenariuszami w grupie scenariuszy. Identyfikator dla pojedynczego scenariusza składa się z przedrostka, po którym znajduje się znak podkreślenia, a następnie wartość zmiennej wymiaru, która identyfikuje szereg początkowy. Wartości wymiarów są oddzielane znakami podkreślenia. Nie ma żadnych ograniczeń, poza unikalnością, dla wartości przedrostka.

Wyrażenie dla wartości scenariusza dla szeregu początkowego

Wartości scenariusza dla grupy scenariuszy są określane za pomocą wyrażenia, które jest następnie używane do wyliczenia wartości dla każdego szeregu początkowego w grupie. Można wprowadzić wyrażenie bezpośrednio lub kliknąć przycisk kalkulatora i utworzyć wyrażenie za pomocą konstruktora wyrażeń wartości scenariusza.

- Wyrażenie może zawierać dowolną zmienną przewidywaną lub przewidywaną w systemie modelu.
- Jeśli okres scenariusza będzie rozszerzony poza istniejące dane, wówczas wyrażenie jest stosowane do wartości prognozowanych dla zmiennych z wyrażenia.

- Dla każdego szeregu początkowego w grupie zmienne w wyrażeniu określają szeregi czasowe, które są definiowane przez te zmienne oraz wartości wymiaru, które definiują szereg początkowy. Są to szeregi czasowe, które są używane do oszacowania wyrażenia. Przykładowo, jeśli szereg początkowy jest definiowany przez wyrażenie region='north' oraz brand='X', wówczas szeregi czasowe, używane w wyrażeniu są definiowane przez te same wartości wymiaru.

Załóżmy na przykład, że zmienna metryki źródłowej to *advertising* i dostępne są dwie zmienne wymiaru *region* i *brand*. Należy również założyć, że grupa scenariuszy zawiera również wszystkie kombinacje wartości zmiennych wymiarów. Następnie można określić wartość advertising*1.2 jako wyrażenie do zbadania efektu zwiększenia wartości *advertising* o 20 procent dla wszystkich szeregów czasowych, jakie są powiązane ze zmienną *advertising*.

Uwaga: Grupy scenariuszy mają zastosowanie tylko do danych wielowymiarowych i tworzone są poprzez kliknięcie opcji **Dodaj grupę scenariuszy** na karcie Scenariusze.

Opcje

Maksymalny poziom dla zmienionych zmiennych przewidywanych
Określa maksymalną liczbę poziomów zmienionej zmiennej przewidywanej. Każdy kolejny poziom, maksymalnie do 5, obejmuje zmienne przewidywane, na które szereg początkowy ma bardziej pośredni wpływ Szczególnie dotyczy to pierwszego poziomu obejmującego zmienne przewidywane, dla których szereg początkowy stanowi bezpośrednią wartość wejściową. Bezpośrednie dane wejściowe zmiennych przewidywanych na drugim poziomie stanowią zmienne przewidywane z pierwszego itd. Zwiększenie wartości tego ustawienia powoduje również zwiększenie złożoności obliczeń i może wpływać na wydajność.

Maksimum automatycznie wykrywanych zmiennych przewidywanych
Określa maksymalną liczbę zmiennych przewidywanych, które są automatycznie wykrywane dla każdego szeregu początkowego. Zwiększenie wartości tego ustawienia powoduje również zwiększenie złożoności obliczeń i może wpływać na wydajność.

Diagram wpływu
Wyświetla graficzną reprezentację relacji przyczynowych pomiędzy szeregiem początkowym dla każdego scenariusza a szeregiem przewidywanym, na który ma on wpływ. Tabele dla wartości scenariusza i wartości przewidywanych dla zmienionych zmiennych przewidywanych są uwzględniane jako część elementu wyniku. Wykres obejmuje wykresy przewidywanych wartości zmienionych zmiennych przewidywanych. Jednokrotne kliknięcie dowolnego węzła w diagramie wpływu otwiera szczegółowy diagram kolejności dla szeregu powiązanego z węzłem. Dla każdego scenariusza generowany jest osobny diagram wpływu.

Wykresy szeregów
Generuje wykresy szeregów przewidywanych wartości dla wszystkich zmienionych zmiennych przewidywanych w scenariuszu.

Tabele prognoz i scenariuszy
Tabele przewidywanych wartości i wartości scenariusza dla każdego scenariusza. Tabele te zawierają te same informacje, co tabele w diagramie wpływu. Obsługują one wszystkie standardowe funkcje przestawiania i edytowania tabel.

Uwzględniaj przedziały ufności na wykresach i w tabelach
Określa, czy przedziały ufności dla predykcji scenariusza są uwzględniane na wykresie i w wyniku.

Szerokość przedziału ufności (%)
To ustawienie kontroluje przedziały ufności dla predykcji scenariusza. Można określić dowolną wartość dodatnią mniejszą do 100. Domyślnie ustawiony jest 95-procentowy przedział ufności.

Węzeł Szeregi czasowe

Węzeł szeregów czasowych może być używany z danymi w środowisku lokalnym lub rozproszonym; w środowisku rozproszonym można wykorzystać możliwości produktu IBM SPSS Analytic Server. Ten węzeł umożliwia estymację i stworzenie modelu wygładzania wykładniczego, modelu autoregresyjnej zintegrowanej średniej ruchomej (ARIMA) jednej zmiennej lub modelu ARIMA wielu zmiennych (lub funkcji przenoszenia) dla danych szeregów czasowych i generuje prognozy w oparciu o dane szeregu czasowego.

Wygładzanie wykładnicze to metoda prognozowania wykorzystująca wartości ważone poprzednich obserwacji szeregu do predykcji przyszłych wartości. Jako takie, wygładzanie wykładnicze nie bazuje na teoretycznej interpretacji danych. Prognozuje ono jeden punkt naraz, dopasowując jego prognozy w miarę dochodzenia nowych danych. Technika ta jest szczególnie przydatna w przypadku prognozowania szeregów wykazujących trend, sezonowość lub oba te zjawiska. Można wybrać spośród szeregu modeli wygładzania wykładniczego, różniących się między sobą sposobem traktowania trendów i sezonowości.

Modele **ARIMA** oferują bardziej wyrafinowane metody modelowania składników trendu i sezonowości, niż modele wygładzania wykładniczego, oraz, w szczególności, mają dodatkową zaletę polegającą na uwzględnianiu w modelu zmiennych niezależnych (predykcyjnych). Obejmuje to jawne określanie kolejności autoregresji i średnich ruchomych, a także stopnia różnicowania. Użytkownik może uwzględnić zmienne predykcyjne i zdefiniować funkcje przenoszenia dla dowolnych lub wszystkich z nich, a także wybrać automatyczne wykrywanie wartości odstających lub jawnie wybrać ich zestaw.

Uwaga: W praktyce modele ARIMA są najbardziej użyteczne w sytuacji, gdy pożądane jest uwzględnienie predyktorów, które mogą pomóc w wyjaśnieniu zachowania prognozowanego szeregu, takich jak liczba rozesłanych katalogów czy liczba odwiedzin na stronie WWW firmy. Modele wygładzania wykładniczego opisują przebieg szeregu czasowego bez próby zrozumienia przyczyn takiego przebiegu. Na przykład szereg o historycznej wartości szczytowej powtarzającej się co 12 miesięcy prawdopodobnie będzie kształtował się w taki sposób nadal, nawet jeśli nie wiemy, dlaczego się tak dzieje.

Program **Expert Modeler**, który pozwala próbować automatycznie identyfikować i estymować modele najlepszego dopasowania ARIMA oraz modele wygładzania wykładniczego do jednej lub więcej zmiennych przewidywanych, eliminując konieczność znajdowania odpowiedniego modelu metodą prób i błędów. W razie wątpliwości należy użyć programu Expert Modeler.

Jeśli określono zmienne predykcyjne, program Expert Modeler wybiera te zmienne, które mają statystycznie znaczącą relację z szeregiem zależnym. Zmienne modelu są przekształcane odpowiednio do potrzeb, za pomocą transformacji różnicującej i/lub pierwiastka kwadratowego lub logarytmu naturalnego. Domyślnie program Expert Modeler uwzględnia wszystkie modele wygładzania wykładniczego oraz wszystkie modele ARIMA i wybiera najlepszy spośród nich model dla każdej zmiennej przewidywanej. Można jednak ograniczyć możliwości programu Expert Modeler w taki sposób, aby umożliwiał on tylko wybór najlepszego z modeli wygładzania wykładniczego lub tylko wybór najlepszego z modeli ARIMA. Można także zdecydować o automatycznym wykrywaniu wartości odstających.

Węzeł szeregów czasowych — opcje zmiennych

Na karcie Zmienne można zdecydować, czy mają zostać użyte ustawienia roli zmiennej już zdefiniowane w węzłach poprzedzających, czy też przypisania zmiennych mają zostać dokonane ręcznie.

Użyj wstępnie zdefiniowanych ról Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Tę opcję należy wybrać, aby ręcznie przypisać zmienne przewidywane, predyktory i inne role.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Przewidywane. Można wybrać jedną lub więcej zmiennych jako przewidywane dla predykcji.

Potencjalne zmienne wejściowe. Można wybrać jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Zdarzenia i interwencje. Ten obszar umożliwia wyznaczenie pewnych zmiennych wejściowych jako zmiennych typu zdarzenie lub interwencja. W ten sposób zmienna jest identyfikowana jako zawierająca dane szeregu czasowego, na które mogą mieć wpływ zdarzenia (przewidywalne sytuacje powtarzalne; na przykład promocje) lub interwencje (jednorazowe incydenty, na przykład przerwa w zasilaniu lub strajk pracowniczy). Wybierane zmienne muszą być flagami, które mają składowanie z użyciem liczby całkowitej.

Węzeł szeregów czasowych — opcje specyfikacji danych

Na karcie Specyfikacje danych można ustawić wszystkie opcje dla danych, które mają zostać uwzględnione w modelu. Jeśli zostaną określone wartości pól **Zmienna typu data/czas** i **Przedział czasowy**, można kliknąć przycisk **Uruchom** w celu zbudowania modelu z wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik potrzebuje zwykle dostosować proces budowy do swoich celów..

Karta zawiera kilka różnych okien, w których można dostosować ustawienia odpowiednio do specyfiki własnego modelu.

Węzeł szeregów czasowych — obserwacje

Należy użyć ustawień na tym panelu, aby określić zmienne definiujące obserwacje.

Obserwacje określone przez zmienną typu data/czas

Można określić, czy obserwacje będą definiowane na podstawie zmiennej daty, czasu lub daty/czasu. Oprócz zmiennej, która definiuje obserwacje, należy wybrać odpowiedni przedział czasowy, w którym będą opisywane obserwacje. W zależności od określonego przedziału czasowego można również dokonać innych ustawień, takich jak przedział między obserwacjami (przyrost) lub liczba dni w tygodniu. Poniższe rozważania mają zastosowanie do przedziałów czasowych:

- Wartości **Nieregularny** należy użyć, jeśli obserwacje są nierównomiernie rozmieszczone w czasie, na przykład są wykonywane w chwili, gdy następuje przetwarzanie zamówienia sprzedaży. Jeśli wybrana zostanie opcja **Nieregularny**, należy określić przedział czasowy, jaki będzie używany do analizy, wybierając ustawienia **Przedział czasowy** na karcie Specyfikacja danych.
- Jeśli obserwacje reprezentują datę i czas, a przedział czasowy to godziny, minuty lub sekundy, wówczas należy wybrać opcję **Godziny dziennie**, **Minuty dziennie** lub **Sekundy dziennie**. Jeśli obserwacje reprezentują czas (trwanie) bez odniesienia do daty, a przedział czasowy to godziny, minuty lub sekundy, wówczas należy użyć opcji **Godziny (nieokresowo)**, **Minuty (nieokresowo)** lub **Sekundy (nieokresowo)**.
- Na podstawie wybranego przedziału czasowego procedura może wykryć brakujące obserwacje. Wykrywanie brakujących obserwacji jest konieczne, ponieważ procedura zakłada, że wszystkie obserwacje są równomiernie rozłożone w czasie i że nie ma brakujących obserwacji. Na przykład, jeśli przedziałem czasu są dni, a po dacie 2015-10-27 występuje 2015-10-29, istnieje brakująca obserwacja dla 2015-10-28. Wartości są podstawiane dla wszystkich brakujących obserwacji; obszar **Traktowanie braków danych** na karcie Specyfikacje danych umożliwia określenie ustawień traktowania brakujących wartości.
- Określone przedziały czasowe umożliwiają procedurze wykrycie wielu obserwacji w jednym przedziale czasowym, które muszą być zagregowane i które są dopasowane do obserwacji dla granicy przedziału (np. pierwszy dzień miesiąca), dzięki czemu obserwacje będą równomiernie rozmieszczone. Na przykład, jeśli przedziałem czasu są miesiące, to zagregowane zostanie wiele dat z tego samego miesiąca. Ten typ agregacji jest nazywany *grupowaniem*. Domyślnie obserwacje są sumowane podczas grupowania. Można określić inną metodę grupowania, np. średnia z obserwacji; ustawienia **Agregacja i rozkład** można wprowadzić na karcie Specyfikacja danych.

- Przy niektórych przedziałach czasowych istnieją dodatkowe ustawienia umożliwiające zdefiniowanie odstępów w zwykle równomiernie rozmieszczonych przedziałach. Na przykład, jeśli przedziałem czasu są dni, ale istotne są tylko dni robocze, można określić pięciodniowy tydzień rozpoczynający się od poniedziałku.

Obserwacje zdefiniowane jako okresy lub okresy cykliczne

Obserwacje mogą być zdefiniowane przez co najmniej jedną zmienną całkowitą, która reprezentuje okresy lub powtarzające się cyklicznie okresy, aż do dowolnej liczby poziomów cyklicznych. Taka struktura pozwala opisać serie obserwacji, które nie pasują do jednego ze standardowych przedziałów czasowych. Przykładowo, rok fiskalny trwający tylko 10 miesięcy może być opisany przez zmienną cyklu, która reprezentuje lata i zmienną okresu, która reprezentuje miesiące, przy czym długość jednego cyklu wynosi 10.

Zmienne, które określają okresy cykliczne, definiują hierarchię poziomów okresowości, w której najniższy poziom jest definiowany przez zmienną **Okres**. Następny wyższy poziom jest określany przez zmienną cyklu z poziomem 1, po której następuje zmienna cyklu z poziomem 2 itd. Wartości zmiennych dla każdego poziomu, z wyjątkiem najwyższego, muszą być okresowe w odniesieniu do kolejnego najwyższego poziomu. Wartości dla najwyższego poziomu nie mogą być okresowe. Na przykład, dla 10-miesięcznego roku fiskalnego miesiące występują okresowo w latach, ale lata nie są okresowe.

- Długość cyklu na poszczególnych poziomach stanowi okresowość dla kolejnego najniższego poziomu. W przykładzie dot. roku fiskalnego istnieje tylko jeden poziom cyklu, a długość cyklu wynosi 10, ponieważ kolejny najniższy poziom reprezentuje miesiące, a w określonym roku fiskalnym jest 10 miesięcy.
- Należy określić wartość początkową dla każdej zmiennej okresowej, która nie rozpoczyna się od 1. To ustawienie jest niezbędne dla wykrywania braków danych. Na przykład, jeśli zmienna okresowa rozpoczyna się od 2, ale wartość początkowa jest określona jako 1, wówczas procedura zakłada, że istnieje brak danych dla pierwszego okresu w każdym cyklu dla tej zmiennej.

Węzeł szeregów czasowych — przedział czasowy dla analizy

Przedział czasowy, który jest używany do analizy, może różnić się od przedziału czasowego dla obserwacji. Na przykład, jeśli przedział czasowy dla obserwacji to dni, jako przedział czasowy dla analizy można wybrać miesiące. Wówczas przed zbudowaniem modelu dane agregowane są z dziennych na miesięczne. Można również rozłożyć dane z dłuższego przedziału czasu na krótszy. Przykładowo, jeśli obserwacje są przeprowadzane kwartalnie, wówczas można rozłożyć dane z kwartalnych na miesięczne.

Należy użyć ustawień na tym panelu, aby określić przedział czasowy dla analizy. Metoda, którą dane są agregowane lub rozkładane, jest określana w ustawieniach **Agregacja i rozkład** na karcie Specyfikacja danych.

Opcje możliwe do wyboru dla przedziału czasowego, w jakim wykonywana jest analiza, zależą od sposobu zdefiniowania obserwacji oraz wyznaczonego dla nich przedziału czasowego. W szczególności, jeśli obserwacje są zdefiniowane przez okresy cykliczne, wówczas obsługiwana jest tylko agregacja. W takim przypadku przedział czasowy dla analizy musi być większy od przedziału czasowego dla obserwacji lub mu równy.

Węzeł szeregów czasowych — opcje rozkładu i agregacji

Należy użyć ustawień na tym panelu, aby określić ustawienia agregacji i rozkładu danych wejściowych odpowiednio do przedziałów czasowych dla obserwacji.

Funkcje agregacji
: Jeśli przedział czasowy użyty dla analizy jest dłuższy niż przedział czasowy dla obserwacji, dane wejściowe zostają zagregowane. Przykładowo agregacja jest przeprowadzana, kiedy przedział czasowy dla obserwacji to dni, a przedział czasowy dla analizy to miesiące. Dostępne są następujące funkcje agregacji: średnia, suma, dominanta, wartość minimalna lub maksymalna.

Funkcje rozkładu
: Jeśli przedział czasowy użyty dla analizy jest krótszy niż przedział czasowy dla obserwacji, dane wejściowe zostają rozłożone. Przykładowo rozkład jest przeprowadzany, kiedy przedział czasowy dla obserwacji to kwartały, a przedział czasowy dla analizy to miesiące. Dostępne są następujące funkcje rozkładu: średnia lub suma.

Funkcje grupujące

Grupowanie jest stosowane, kiedy obserwacje są definiowane przez datę/czas i w tym samym przedziale czasowym występuje wiele obserwacji. Na przykład, jeśli przedział czasowy dla obserwacji to miesiące, wówczas wiele dat z tego samego miesiąca jest grupowanych i tworzone jest ich powiązanie z miesiącem, w którym występują. Dostępne są następujące funkcje grupowania: średnia, suma, dominanta, wartość minimalna lub maksymalna. Grupowanie jest zawsze przeprowadzane, kiedy obserwacje są zdefiniowane przez datę/czas, a przedział czasowy dla obserwacji jest określony jako Nieregularny.

Uwaga: Chociaż grupowanie jest formą agregacji, jest przeprowadzane przed rozpoczęciem obsługi braków danych, podczas gdy formalna agregacja jest wykonywana po zakończeniu obsługi braków danych. Jeśli przedział czasowy dla obserwacji jest określony jako Nieregularny, agregacja jest wykonywana tylko za pomocą funkcji grupowania.

Agreguj obserwacje przekraczające granice dnia do dnia poprzedniego

Określa, czy obserwacje, których czas przekracza granicę dnia, są agregowane na wartości dla dnia poprzedniego. Przykładowo, dla obserwacji godzinowych trwających osiem godzin dziennie i rozpoczynających się o godzinie 20:00, to ustawienie określi, czy obserwacje od godziny 00:00 do 04:00 będą uwzględniane w zagregowanych wynikach dla poprzedniego dnia. To ustawienie ma zastosowanie tylko w przypadku, kiedy przedział czasowy dla obserwacji to Godziny dziennie, Minuty dziennie lub Sekundy dziennie, a przedział czasowy dla analizy to dni.

Ustawienia niestandardowe dla określonych zmiennych

Funkcje agregacji, rozkładu i grupowania dla zmiennej można określić na podstawie zmiennej. Ustawienia te zastępują domyślne ustawienia funkcji agregacji, rozkładu i grupowania.

Węzeł szeregów czasowych — opcje braków danych

Należy użyć ustawień na tym panelu, aby określić, w jaki sposób brakujące wartości w danych wejściowych mają zostać zastąpione wartością podstawianą. Dostępne są zastępujące metody zastępowania:

Interpolacja liniowa

Powoduje zastąpienie braków danych przy wykorzystaniu interpolacji liniowej. W interpolacji używana jest ostatnia ważna wartość przed brakiem danych oraz pierwsza ważna za brakiem. Jeśli pierwsza lub ostatnia obserwacja w szeregu zawiera brakujące wartości, wówczas używane są dwie najbliższe niebrakujące wartości na początku i na końcu serii.

Średnia szeregu

Zastępuje braki danych średnią obliczoną ze wszystkich obserwacji.

Średnia z sąsiednich punktów

Powoduje zastąpienie braków danych średnią z ważnych wartości sąsiednich. Rozpiętość sąsiednich punktów, to liczba poprawnych wartości występujących przed i po brakującej wartości, jakie są wykorzystywane do obliczenia średniej.

Mediana z sąsiednich punktów

Powoduje zastąpienie braków danych medianą ważnych wartości sąsiednich. Rozpiętość sąsiednich punktów, to liczba poprawnych wartości występujących przed i po brakującej wartości, jakie są wykorzystywane do obliczenia mediany.

Trend liniowy

Ta opcja wykorzystuje niebrakujące obserwacje w szeregu do dopasowania prostego modelu regresji liniowej, który jest następnie używany w celu przypisania brakujących wartości.

Inne ustawienia:

Najniższa ocena jakości danych (%)

Oblicza miary jakości danych zmiennej czasu i danych wejściowych odpowiadających każdemu z szeregów czasowych. Jeśli jakość danych jest niższa od podanego tutaj progu, odpowiedni szereg czasowy zostanie odrzucony.

Węzeł szeregów czasowych — okres oszacowania

W panelu Okres oszacowania można określić przedział rekordów, które zostaną użyte w oszacowaniu modelu. Domyślnie okres estymacji zaczyna się od czasu z najwcześniejszą obserwacją, a kończy w czasie z najpóźniejszą obserwacją we wszystkich szeregach.

Wyznaczony przez czas początkowy i końcowy

Można określić datę rozpoczęcia i zakończenia okresu estymacji lub można określić tylko datę rozpoczęcia lub tylko datę zakończenia. Jeśli rozpoczęcie lub zakończenie okresu estymacji zostanie pominięte, użyta będzie wartość domyślna.

- Jeśli obserwacje są zdefiniowane przez określenie zmiennej daty/czasu, wówczas wartości rozpoczęcia i zakończenia okresu należy wprowadzić w takim samym formacie, jaki został użyty dla zmiennej daty/czasu.
- W przypadku obserwacji definiowanych na podstawie okresów cyklicznych należy określić wartość dla każdej zmiennej okresu cyklicznego. Każda zmienna jest wyświetlana w osobnej kolumnie.

Wyznaczony przez najwcześniejszy lub najpóźniejszy przedział czasowy

Definiuje okres estymacji jako określoną liczbę przedziałów czasowych, która rozpoczyna się od najwcześniejszego przedziału czasowego lub kończy na najpóźniejszym przedziale czasowym określonym w danych, z opcjonalnym przesunięciem. W tym kontekście przedział czasowy odnosi się do przedziału czasowego dla analizy. Przykładowo, załóżmy, że obserwacje są przeprowadzane miesięcznie, ale przedział czasowy dla analizy to kwartały. Określenie wartości **Najpóźniejszy** i wartości 24 dla opcji **Liczba przedziałów czasowych** będzie oznaczało ostatnie 24 kwartały.

Opcjonalnie można wykluczyć określoną liczbę przedziałów czasowych. Przykładowo, określenie ostatnich 24 przedziałów czasowych i 1 do wykluczenia oznacza, że okres oszacowania składa się z 24 przedziałów, które poprzedzają ostatni.

Węzeł szeregów czasowych — opcje budowania

Na karcie Opcje budowania można ustawić wszystkie opcje budowania modelu. Można oczywiście po prostu kliknąć przycisk **Uruchom** w celu zbudowania modelu z wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik potrzebuje zwykle dostosować proces budowy do swoich celów.

Karta zawiera dwa różne panele, w których można dostosować ustawienia odpowiednio do specyfiki własnego modelu.

Węzeł szeregów czasowych — ogólne opcje budowania

Dostępność opcji w tym panelu zależy od tego, które z poniższych trzech ustawień zostaną wybrane na liście **Metoda**:

- **Automatyczny dobór modelu.** Wybór tej opcji powoduje użycie funkcji Automatyczny dobór modelu, która automatycznie znajduje model o najlepszym dopasowaniu dla każdego szeregu zależnego.
- **Wygładzanie wykładnicze.** Ta opcja umożliwia określenie niestandardowego modelu wygładzania wykładniczego.
- **ARIMA.** Ta opcja umożliwia określenie niestandardowego modelu ARIMA.

Automatyczny dobór modelu

W obszarze **Typ modelu** wybierz typ modeli, które chcesz tworzyć:

- **Wszystkie modele.** Funkcja automatycznego doboru modeli uwzględnia zarówno modele ARIMA, jak i modele wygładzania wykładniczego.
- **Tylko modele wygładzania wykładniczego.** Automatyczny dobór modeli uwzględnia tylko modele wygładzania wykładniczego.
- **Tylko modele ARIMA.** Automatyczny dobór modeli uwzględnia tylko modele ARIMA.

Automatyczny dobór modelu uwzględnia modele sezonowe. Ta opcja jest aktywna tylko wówczas, jeśli dla aktywnego zbioru danych zdefiniowano okresowość. Po zaznaczeniu tej opcji automatyczny dobór modeli uwzględnia zarówno modele sezonowe, jak i niesezonowe. W przypadku niezaznaczenia tej opcji automatyczny dobór modeli uwzględnia tylko modele niesezonowe.

Automatyczny dobór modelu uwzględnia wyrafinowane metody wygładzania wykładniczego. Gdy ta opcja jest wybrana, podczas automatycznego doboru modelu program przeszukuje łącznie 13 modeli wygładzania wykładniczego (7 z nich istniało w pierwotnym węźle Szereg czasowy, a 6 dodano w wersji 18.1). W przypadku niezaznaczenia tej opcji automatyczny dobór modeli uwzględnia tylko pierwotnych 7 modeli wygładzania wykładniczego.

W obszarze **Wartości odstające** dokonaj wyboru spośród następujących opcji

Automatycznie wykryj wartości odstające. Domyślnie automatyczne wykrywanie wartości odstających nie jest przeprowadzane. Zaznacz tę opcję, aby automatycznie wykrywać wartości odstające, a następnie wybierz żądany typ wartości odstających.

Aby zmienne wejściowe zostały uwzględnione na tej liście, muszą mieć poziom pomiaru *Flaga*, *Nominalny* lub *Porządkowy* i muszą być numeryczne (na przykład 1/0 zamiast Prawda/Fałsz w przypadku zmiennej flagi).

Automatyczny dobór modeli uwzględnia tylko regresję prostą, pomijając dowolne funkcje przenoszenia dla danych wejściowych zidentyfikowanych jako zmienne typu zdarzenie czy interwencja na karcie **Zmienne**.

Wygładzanie wykładnicze

Typ modelu. Modele wygładzania wykładniczego są klasyfikowane jako sezonowe lub niesezonowe.[1] Modele sezonowe są dostępne tylko wówczas, jeśli okresowość zdefiniowana za pomocą węzła Przedziały czasowe na karcie Specyfikacja danych jest sezonowa. Okresowości sezonowe to: okresy cykliczne, lata, kwartały, miesiące, dni tygodnia, godziny dnia, minuty dnia oraz sekundy dnia. Dostępne są następujące typy modeli:

- **Prosty.** Ten model jest odpowiedni w przypadku szeregu, w którym brak trendu lub sezonowości. Jedynym odpowiednim parametrem wygładzania jest poziom. Proste wygładzanie wykładnicze jest najbardziej podobne do modelu ARIMA bez rzędów autoregresji, z jednym rzędem różnicowania, jednym rzędem ruchomej średniej i brakiem stałych.
- **Trend liniowy Holta.** Model ten jest odpowiedni dla szeregów, w których istnieje trend liniowy i nie ma sezonowości. Właściwe dla niego parametry wygładzania to poziom i trend; w przypadku tego modelu nie są one ograniczone wzajemnie swoimi wartościami. Model Holta jest bardziej ogólny niż model Browna, ale w przypadku dłuższych szeregów przeliczanie zajmuje więcej czasu. Wygładzanie wykładnicze Holta jest najbardziej podobne do modelu ARIMA bez rzędu autoregresji, z dwoma rzędami różnicowania i dwoma rzędami średniej ruchomej.
- **Trend wygasający.** Model ten jest odpowiedni dla szeregów, w których istnieje trend liniowy, który wygasa i nie ma sezonowości. Właściwe dla niego parametry wygładzania to: poziom, trend i trend osłabiający. Wygasające wygładzanie wykładnicze jest najbardziej podobne do modelu ARIMA z jednym rzędem autoregresji, jednym rzędem różnicowania i dwoma rzędami średniej ruchomej.
- **Trend multiplikatywny.** Model ten jest odpowiedni dla szeregów, w których występuje trend zmienny wraz z modułem szeregu i nie ma sezonowości. Właściwe dla niego parametry wygładzania to: poziom i trend. Wygładzanie wykładnicze trendu multiplikatywnego nie jest podobne do modelu ARIMA.
- **Trend liniowy Browna.** Model ten jest odpowiedni dla szeregów, w których istnieje trend liniowy i nie ma sezonowości. Właściwe dla niego parametry wygładzania to poziom i trend; w tym modelu zakłada się jednak, że są one sobie równe. Zatem model Browna jest specjalnym przypadkiem modelu Holta. Wygładzanie wykładnicze Browna jest podobne do modelu ARIMA z zerowym rzędem autoregresji, dwoma rzędami różnic oraz dwoma rzędami ruchomych średnich ze współczynnikiem drugiego rzędu ruchomej średniej równym kwadratowi połowy współczynnika pierwszego rzędu.
- **Prosty sezonowy.** Model ten jest odpowiedni dla szeregów, w których nie ma trendu, a efekt sezonowości jest stały w czasie. Właściwe dla niego parametry wygładzania to: poziom i sezon. Proste sezonowe wygładzanie wykładnicze jest podobne do modelu ARIMA bez rzędów autoregresji, jednym rzędem różnicowania oraz jednym rzędem różnicowania sezonowego i rzędami 1, p i p+1 średniej ruchomej, gdzie p jest liczbą okresów w przedziale sezonowym. W przypadku danych miesięcznych p = 12.
- **Addytywny model Wintersa.** Model ten jest odpowiedni dla szeregów, w których nie ma trendu liniowego, a efekt sezonowości jest stały w czasie. Właściwe dla niego parametry wygładzania to: poziom, trend i sezon. Wygładzanie

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

wykładnicze addytywnego modelu Wintersa jest podobne do modelu ARIMA bez rzędu autoregresji, z jednym rzędem różnicowania oraz jednym rzędem różnicowania sezonowego i p+1 rzędami średniej ruchomej, gdzie p jest liczbą okresów w przedziale sezonowym. W przypadku danych miesięcznych p = 12.

- **Trend wygasający ze składnikiem sezonowym addytywnym.** Model ten jest odpowiedni dla szeregów, w których istnieje trend liniowy, który wygasa, a efekt sezonowości jest stały w czasie. Właściwe dla niego parametry wygładzania to: poziom, trend, trend gasnący i sezon. Trend gasnący i wygładzanie wykładnicze nie są podobne do modelu ARIMA.
- **Trend multiplikatywny ze składnikiem sezonowym addytywnym.** Model ten jest odpowiedni dla szeregów, w których występuje trend zmienny wraz z modułem szeregu, a efekt sezonowości jest stały w czasie. Właściwe dla niego parametry wygładzania to: poziom, trend i sezon. Trend multiplikatywny i wygładzanie wykładnicze nie są podobne do modelu ARIMA.
- **Multiplikatywny ze składnikiem sezonowym.** Model ten jest odpowiedni dla szeregów, w których nie ma trendu, a efekt sezonowości jest zmienny wraz z modułem szeregu. Właściwe dla niego parametry wygładzania to: poziom i sezon. Wygładzanie wykładnicze trendu multiplikatywnego ze składnikiem sezonowym nie jest podobne do modelu ARIMA.
- **Multiplikatywny model Wintersa.** Model ten jest odpowiedni dla szeregów, w których występuje trend liniowy oraz efekt sezonowości zmienny wraz z modułem szeregu. Właściwe dla niego parametry wygładzania to: poziom, trend i sezon. Wygładzanie wykładnicze multiplikatywnego modelu Wintersa nie jest podobne do modelu ARIMA.
- **Trend wygasający ze składnikiem sezonowym multiplikatywnym.** Model ten jest odpowiedni dla szeregów, w których istnieje trend liniowy, który wygasa, oraz efekt sezonowości zmienny wraz z modułem szeregu. Właściwe dla niego parametry wygładzania to: poziom, trend, trend gasnący i sezon. Trend gasnący i wygładzanie wykładnicze trendu multiplikatywnego ze składnikiem sezonowym nie są podobne do modelu ARIMA.
- **Trend multiplikatywny ze składnikiem sezonowym multiplikatywnym.** Model ten jest odpowiedni dla szeregów, w których występuje trend, a efekt sezonowości jest zmienny wraz z modułem szeregu. Właściwe dla niego parametry wygładzania to: poziom, trend i sezon. Trend multiplikatywny i wygładzanie wykładnicze trendu multiplikatywnego ze składnikiem sezonowym nie są podobne do modelu ARIMA.

Transformacja zmiennej przewidywanej. Istnieje możliwość określenia transformacji do wykonania dla każdej zmiennej zależnej przed jej zamodelowaniem.

- **Brak.** Nie jest wykonywana żadna transformacja.
- **Pierwiastek kwadratowy.** Wykonywana jest transformacja pierwiastkiem kwadratowym.
- **Logarytm naturalny.** Wykonywana jest transformacja logarytmem naturalnym.

ARIMA

Określ strukturę niestandardowego modelu ARIMA.

Rzędy ARIMA. Należy wprowadzić wartości dla różnych składników ARIMA modelu do odpowiednich komórek siatki. Wszystkie wartości muszą być nieujemnymi liczbami całkowitymi. W przypadku składników autoregresja i średnia ruchoma wartość ta reprezentuje rząd maksymalny. W modelu zostaną uwzględnione wszystkie niższe rzędy dodatnie. Na przykład w przypadku podania wartości 2 model będzie obejmował rzędy 2 i 1. Komórki w kolumnie Sezonowa są aktywne tylko, jeśli dla aktywnego zbioru danych zdefiniowano okresowość.

- **Autoregresja (p).** Liczba rzędów autoregresji w modelu. Rzędy autoregresji określają, które z poprzednich wartości są używane do przewidywania bieżących wartości. Na przykład rząd autoregresji 2 oznacza, że do przewidywania bieżącej wartości zostanie użyta wartość dwu okresów czasu szeregu w przeszłości.
- **Różnica (d).** Określa rząd różnicowania stosowany względem szeregu przed oszacowaniem modeli. Różnicowanie jest niezbędne w przypadku obecności trendów (szeregi z trendami są zwykle niestacjonarne, zaś modelowanie ARIMA zakłada stacjonarność) i służy do usuwania ich wpływu. Rząd różnicowania odpowiada stopniowi trendu szeregu — różnicowanie pierwszego rzędu jest uwzględniane w przypadku trendów liniowych, różnicowanie drugiego rzędu w przypadku trendów kwadratowych itd.
- **Średnia ruchoma (q).** Liczba rzędów średniej ruchomej w modelu. Rzędy średniej ruchomej określają, w jaki sposób odchylenia od średniej szeregu dla poprzednich wartości są używane do przewidywania bieżących wartości.

Na przykład rzędy średniej ruchomej 1 i 2 oznaczają, że odchylenia od wartości średniej szeregu z każdego z dwu ostatnich okresów czasu będą uwzględniane podczas przewidywania bieżących wartości szeregu.

Sezonowe. Autoregresja sezonowa, średnia ruchoma i różnicowanie odgrywają takie same role, jak ich niesezonowe odpowiedniki. W przypadku rzędów sezonowości na bieżące wartości szeregu wpływają jednak poprzednie wartości szeregu, rozdzielone jednym lub większą liczbą okresów sezonowości. Na przykład w przypadku danych miesięcznych (okres sezonowości 12) rząd sezonowości 1 oznacza, że na bieżącą wartość szeregu wpływa 12 okresów wartości szeregu poprzedzających okres bieżący. Rząd sezonowości 1, w przypadku danych miesięcznych, jest wówczas taki sam, jak w przypadku rzędu niesezonowości wynoszącego 12.

Automatycznie wykryj wartości odstające. Wybór tej opcji pozwala automatycznie wykrywać wartości odstające i wybrać jeden lub więcej spośród dostępnych typów wartości odstających.

Typy wykrywanych wartości odstających. Wybierz typ(y) wartości odstających, które chcesz wykrywać. Obsługiwane typy to:

- Addytywne (domyślny)
- Przesunięcie poziomu (domyślny)
- Innowacyjne
- Przemijające
- Sezonowo addytywne
- Trend lokalny
- Wiązka addytywna

Rzędy i transformacje funkcji przenoszenia. W celu zdefiniowania przenoszenia oraz funkcji przenoszenia dla wybranych lub wszystkich zmiennych wejściowych w modelu ARIMA kliknij opcję **Ustaw**; zostanie wyświetlone osobne okno dialogowe, w którym można wprowadzić szczegółowe dane dotyczące przenoszenia i przekształcania.

Dołącz stałe do modelu. Uwzględnienie stałej to praktyka standardowa, o ile użytkownik ma pewność, że ogólna wartość średniej szeregu wynosi 0. Wykluczenie stałej zaleca się w przypadku stosowania różnicowania.

Dalsze szczegóły

- Więcej informacji na temat typów wartości odstających można znaleźć w temacie “Wartości odstające” na stronie 290.
- Więcej informacji na temat funkcji przenoszenia i transformacji można znaleźć w temacie “Funkcje przenoszenia i transformacji”.

Funkcje przenoszenia i transformacji: W celu zdefiniowania funkcji przenoszenia dla wybranych lub wszystkich zmiennych wejściowych w modelu ARIMA służy okno dialogowe Rzędy i transformacje funkcji przenoszenia.

Transformacja zmiennej przewidywanej. Istnieje możliwość określenia transformacji do wykonania dla każdej zmiennej przewidywanej przed jej zamodelowaniem.

- **Brak.** Nie jest wykonywana żadna transformacja.
- **Pierwiastek kwadratowy.** Wykonywana jest transformacja pierwiastkiem kwadratowym.
- **Logarytm naturalny.** Wykonywana jest transformacja logarytmem naturalnym.

Funkcje przenoszenia i transformacje potencjalnych zmiennych wejściowych. Funkcje przenoszenia pozwalają określać sposób wykorzystania poprzednich wartości zmiennych wejściowych do prognozowania przyszłych wartości szeregu przewidywanego. Lista po lewej stronie panelu zawiera wszystkie zmienne wejściowe. Pozostałe informacje w tym panelu są charakterystyczne dla wybranej zmiennej wejściowej.

Polecenia wykonania funkcji transferu. Należy wprowadzić wartości dla różnych składników funkcji przenoszenia do odpowiednich komórek siatki **Struktura** . Wszystkie wartości muszą być nieujemnymi liczbami całkowitymi. W przypadku składników, takich jak licznik i mianownik, wartość ta reprezentuje rząd maksymalny. W modelu zostaną

uwzględnione wszystkie niższe rzędy dodatnie. Ponadto w przypadku składników typu licznik zawsze uwzględniany jest rząd 0. Na przykład w przypadku wskazania dla licznika wartości 2 model będzie obejmował rzędy 2, 1 i 0. W przypadku wskazania wartości 3 dla mianownika model będzie obejmował rzędy 3, 2 i 1. Komórki w kolumnie Sezonowa są aktywne tylko, jeśli dla aktywnego zbioru danych zdefiniowano okresowość.

Licznik. Rząd licznika funkcji przenoszenia określa, które poprzednie wartości z wybranego szeregu niezależnego (predyktora) są używane do przewidywania bieżących wartości szeregu zależnego. Na przykład rząd licznika równy 1 oznacza, że wartość szeregu niezależnego jeden okres czasu wstecz oraz bieżąca wartość szeregu niezależnego służą do przewidywania bieżącej wartości każdego szeregu zależnego.

Mianownik. Rząd mianownika funkcji przenoszenia określa, jak odchylenia od średniej szeregu, dla poprzednich wartości wybranych szeregów niezależnych (predyktora) są używane do przewidywania bieżących wartości szeregu zależnego. Na przykład rząd mianownika 1 oznacza, że odchylenia od wartości średniej szeregu niezależnego jeden okres czasu wstecz zostaną uwzględnione podczas przewidywania bieżącej wartości każdego szeregu zależnego.

Różnica. Określa rząd różnicowania stosowany względem wybranego szeregu niezależnego (predykcyjnego) przed oszacowaniem modeli. Różnicowanie jest niezbędne w przypadku obecności trendów i służy do usuwania ich wpływu.

Sezonowe. Składniki, takie jak licznik, mianownik i różnicowanie odgrywają takie same role, jak ich niesezonowe odpowiedniki. W przypadku rzędów sezonowości na bieżące wartości szeregu wpływają jednak poprzednie wartości szeregu, rozdzielone jednym lub większą liczbą okresów sezonowości. Na przykład w przypadku danych miesięcznych (okres sezonowości 12) rząd sezonowości 1 oznacza, że na bieżącą wartość szeregu wpływa 12 okresów wartości szeregu poprzedzających okres bieżący. Rząd sezonowości 1, w przypadku danych miesięcznych, jest wówczas taki sam, jak w przypadku rzędu niesezonowości wynoszącego 12.

Opóźnienie. Ustawienie opóźnienia powoduje opóźnienie wpływu zmiennej wejściowej o podaną liczbę przedziałów. Na przykład ustawienie opóźnienia na wartość 5 oznacza, że wartość zmiennej wejściowej w chwili *t* nie wpływa na prognozy aż do chwili upłynięcia pięciu okresów (*t* + 5).

Transformacja. Specyfikacja funkcji przenoszenia dla zestawu zmiennych niezależnych obejmuje także opcjonalną transformację do wykonania na tych zmiennych.

- **Brak.** Nie jest wykonywana żadna transformacja.
- **Pierwiastek kwadratowy.** Wykonywana jest transformacja pierwiastkiem kwadratowym.
- **Logarytm naturalny.** Wykonywana jest transformacja logarytmem naturalnym.

Węzeł szeregów czasowych — opcje wyników budowania

Maksimum przesunięć w wynikach ACF i PACF. Autokorelacja (ACF) i autokorelacja cząstkowa (PACF) to miary związków między bieżącymi i przeszłymi wartościami szeregów określające, które przeszłe wartości szeregów są najbardziej użyteczne przy przewidywaniu przyszłych wartości. Można tutaj ustawić maksymalną liczbę przesunięć wyświetlanych w tabelach i na wykresach autokorelacji i autokorelacji cząstkowych.

Oblicz ważność predyktora. W przypadku modeli generujących odpowiednią miarę ważności możliwe jest wyświetlenie tabeli wskazującej ważność względną każdego predyktora w procesie estymacji modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Obliczenie ważności predyktora może potrwać dłużej dla niektórych modeli, szczególnie w przypadku pracy z dużymi zbiorami danych, i domyślnie ta opcja dla niektórych modeli jest wyłączona.

Węzeł szeregów czasowych — opcje modelu

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Szerokość przedziału ufności (%). Przedziały ufności są obliczane dla predykcji modelu i autokorelacji reszt. Można określić dowolną wartość dodatnią mniejszą do 100. Domyślnie ustawiony jest 95-procentowy przedział ufności.

Kontynuuj oszacowanie za pomocą istniejących modeli. Jeśli model szeregów czasowych był już generowany, wybór tej opcji pozwala ponownie użyć ustawień kryteriów określonych dla wcześniejszego modelu i generować węzeł nowego modelu w palecie Modele zamiast budowania nowego modelu od początku. Można w ten sposób zaoszczędzić czas, ponownie oszacowując i tworząc nową prognozę w oparciu o te same ustawienia modelu, co wcześniej, lecz na podstawie bardziej aktualnych danych. Dlatego na przykład, jeśli model oryginalny dla określonego szeregu czasowego był trendem liniowym Holta, do ponownego oszacowania i prognozy dla tych danych używany jest ten sam typ modelu. System nie ponawia próby znalezienia najlepszego typu modelu dla nowych danych.

Zbuduj tylko model oceniający. Zaznaczenie tego pola wyboru pozwala zmniejszyć ilość danych przechowywanych w modelu. Może to zwiększyć wydajność budowy modeli o bardzo dużej liczbie szeregów czasowych (liczonej w dziesiątkach tysięcy). Nadal można oceniać dane w zwykły sposób.

Rozszerz rekordy na przedziały z przyszłości. Włącza sekcję **Wartości przyszłe używane w prognozowaniu**, w której można ustawić liczbę przedziałów czasowych do prognozowania poza koniec okresu estymacji. Przedział czasowy jest w tym przypadku przedziałem czasowym dla analizy, określonym na karcie Specyfikacja danych. Dla tego ustawienia nie ma maksymalnego limitu. Dzięki następującym opcjom można automatycznie obliczać przyszłe wartości zmiennych wejściowych lub ręcznie określić wartości prognozowane dla jednego lub więcej predyktorów.

Wartości przyszłe używane w prognozowaniu

- **Oblicz przyszłe wartości zmiennych wejściowych** W przypadku zaznaczenia tej opcji automatycznie obliczane są wartości prognozy dla predyktorów, predykcje szumu, oszacowanie wariancji oraz przyszłe wartości czasu. Z chwilą wywołania prognoz następuje automatyczna budowa modeli autoregresji dla szeregów wejściowych niebędących jednocześnie wartościami przewidywanymi. Modele te są następnie używane do generowania wartości dla tych szeregów wejściowych w okresie prognozy.
- **Wybierz zmienne, których wartości mają być dodane do danych.** W przypadku każdego rekordu, jaki ma zostać objęty prognozą (z wyjątkiem wstrzymań), korzystanie ze zmiennych predykcyjnych (z rolą ustawioną na Dane wejściowe) pozwala na określenia oszacowanych wartości dla okresu prognozy każdego predyktora. Wartości można określić ręcznie lub wybrać je z listy.
  - **Zmienna.** Należy kliknąć przycisk Selektor zmiennych i wybrać zmienne, które mogą być używane jako predyktory. Należy pamiętać, że wybrane tutaj zmienne mogą, ale nie muszą być używane w modelowaniu; aby rzeczywiście użyć zmiennej jako predyktora, musi być ona wybrana w dalszym węźle modelowania. To okno dialogowe pozwala na określanie przyszłych wartości w wygodny sposób, dzięki czemu mogą one być współużytkowane przez wiele kolejnych węzłów modelowania bez konieczności określania ich w każdym węźle osobno. Należy także pamiętać, że lista dostępnych zmiennych może być ograniczona poprzez wybór na karcie Opcje budowania.

    Należy pamiętać, że jeśli przyszłe wartości są określone dla zmiennej, która nie jest dłużej dostępna w strumieniu (z powodu usunięcia lub aktualizacji wyboru na karcie Opcje budowania), zmienna jest przedstawiona na czerwono.
  - **Wartości.** Dla każdej zmiennej można wybrać funkcje z listy lub kliknąć **Określ**, aby wprowadzić wartości ręcznie lub wybrać je z listy predefiniowanych wartości. Jeśli zmienne predykcyjne odnoszą się do elementów objętych kontrolą użytkownika lub elementów, które są wcześniej znane z innych powodów, wartości należy wprowadzić ręcznie. Przykładowo, jeśli tworzona jest prognoza przychodów hotelu dla kolejnego miesiąca na podstawie liczby zarezerwowanych pokoi, można określić liczbę rezerwacji w rzeczywistości dokonanych dla tego okresu. I odwrotnie, jeśli zmienna predykcyjna odnosi się do elementów poza kontrolą użytkownika, na przykład cena akcji, należy użyć funkcji, takich jak ostatnia wartość lub średnia z ostatnich punktów.

    Dostępne funkcje zależą pod poziomu pomiaru zmiennej.

Tabela 28. Funkcje dostępne dla poziomów pomiaru

Poziom pomiaru	Funkcje
Zmienna ilościowa lub nominalna	Pusta Średnia z ostatnich punktów Ostatnia wartość Określ

Tabela 28. Funkcje dostępne dla poziomów pomiaru (kontynuacja)

Poziom pomiaru	Funkcje
Zmienna flagi	Pusta Ostatnia wartość Prawda Fałsz Określ

Średnia z ostatnich punktów oblicza przyszłą wartość na podstawie średniej z ostatnich trzech punktów danych.

Ostatnia wartość ustawia przyszłą wartość w oparciu o najnowszy punkt danych.

Prawda/Fałsz ustawia przyszłą wartość zmiennej typu flaga jako Prawda lub Fałsz.

Określ otwiera okno dialogowe w celu określenia przyszłych wartości ręcznie lub wybrania ich z predefinowanej listy.

Udostępnij do oceniania

W tym miejscu można ustawić wartości domyślne dla opcji oceniania, które będą widoczne w oknie dialogowym dla modelu użytkowego.

- **Oblicz górny i dolny przedział ufności.** Zaznaczenie tej opcji powoduje utworzenie nowych zmiennych (o prefiksach domyślnych $TSLCI- i $TSUCI-) dla dolnego i górnego przedziału ufności, dla każdej zmiennej przewidywanej.
- **Oblicz reszty szumów.** Zaznaczenie tej opcji powoduje utworzenie nowej zmiennej (z prefiksem domyślnym $TSResidual-) dla reszt modelu dla każdej zmiennej przewidywanej, wraz z sumą tych wartości.

Ustawienia modelu

Maksymalna liczba modeli wyświetlanych w wynikach. Określa maksymalną liczbę modeli, które mają być zawarte w wynikach. Należy pamiętać, że jeśli liczba zbudowanych modeli przekracza ten próg, modele nie są wyświetlane w wynikach, tylko nadal dostępne do oceny. Wartość domyślna to 10. Wyświetlenie dużej liczby modeli może spowodować pogorszenie wydajności lub niestabilność oprogramowania.

Model użytkowy szeregów czasowych

Wynik modelu użytkowego szeregów czasowych

Po utworzeniu modelu szeregów czasowych w oknie wyników dostępne są następujące informacje. Należy pamiętać o ustalonym dla modeli szeregów czasowych limicie 10 modeli, które można wyświetlać w przeglądarce wyników.

Podsumowanie informacji o definicji czasu

Podsumowanie zawiera informacje dotyczące:

- Zmiennej czasu
- Przedziału
- Punktu startowego i końcowego
- Liczby unikalnych punktów

Podsumowanie dotyczy wszystkich zmiennych przewidywanych.

Tabela Informacje o modelu

Tabela Informacje o modelu zawiera kluczowe informacje o modelu dla każdej zmiennej przewidywanej. Tabela zawiera niektóre ustawienia modelu wysokiego poziomu, takie jak:

- Nazwa zmiennej przewidywanej wybranej w węźle Typ lub na karcie Zmienne węzła szeregów czasowych.

- Metoda budowania modelu – na przykład Wygładzanie wykładnicze lub ARIMA.
- Liczba predyktorów wprowadzonych do modelu.
- Liczba rekordów użytych do dopasowania typu modelu. Przykładami różnych typów modeli mogą być: RMSE, MAE, AIC, BIC i R-kwadrat

Ponadto może być również wyświetlana statystyka Q Ljunga-Boxa, o ile dane spełniają wymagane warunki. Ta statystyka **nie** jest dostępna pod następującymi warunkami:

- Jeśli liczba niebrakujących punktów danych jest mniejsza lub równa liczbie żądanych składników sumowania (ustalonej na 18).
- Jeśli liczba parametrów jest większa lub równa liczbie żądanych składników sumowania.
- Jeśli obliczona liczba składników sumowania jest mniejsza niż najmniejsza akceptowalna wartość k (ustalona na 7).
- Jeśli tabela powtarza się dla każdej zmiennej przewidywanej.

Ważność predyktorów

Wykres Ważność predyktorów przedstawia ważność pierwszych 10 danych wejściowych (predyktorów) w modelu jako wykres słupkowy dla każdej zmiennej przewidywanej.

W przypadku, gdy na wykresie jest więcej niż 10 zmiennych, można zmienić wybór predyktorów uwzględnianych na wykresie, korzystając z suwaka pod wykresem. Wskaźniki na suwaku mają stałą szerokość, a każdy znak na suwaku reprezentuje 10 zmiennych. Wskaźniki można przemieszczać wzdłuż suwaka, wyświetlając w ten sposób 10 kolejnych lub poprzednich zmiennych, uporządkowanych według ważności predyktora.

Dwukrotne kliknięcie wykresu powoduje otwarcie osobnego okna dialogowego, w którym można edytować ustawienia wykresu. Można na przykład zmodyfikować cechy, takie jak wielkość wykresu, a także rozmiar i kolor używanych czcionek. Po zamknięciu tego osobnego okna dialogowego do edycji zmiany są odzwierciedlane na wykresie wyświetlanym na karcie Wynik.

Korelogram

Korelogram lub wykres autokorelacji jest przedstawiany dla każdej zmiennej przewidywanej i przedstawia funkcję autokorelacji (ACF) lub funkcję autokorelacji cząstkowej (PACF) reszt (różnic między wartościami oczekiwanymi a rzeczywistymi) względem opóźnień czasowych. Przedział ufności jest przedstawiany jako podświetlenie na wykresie.

Oszacowania parametrów

Powtarzana dla każdej zmiennej przewidywanej tabela Oszacowania parametrów przedstawia (tam, gdzie ma to zastosowanie) następujące dane szczegółowe:

- Nazwa zmiennej przewidywanej.
- Zastosowana transformacja.
- Opóźnienia zastosowane do tego parametru w modelu (ARIMA)
- Wartość współczynnika
- Błąd standardowy oszacowania parametru.
- Wartość oszacowania parametru podzielona przez błąd standardowy.
- Poziom istotności dla oszacowania parametru.

Ustawienia modelu użytkowego szeregów czasowych

Karta Ustawienia zawiera dodatkowe opcje dla modelu użytkowego Szeregi czasowe.

Prognoza

Opcja **Rozszerz rekordy na przedziały z przyszłości.** pozwala ustawić liczbę przedziałów czasowych do prognozowania poza koniec okresu estymacji. Przedział czasu jest w tym przypadku okresem czasu analizy,

określonym na karcie Specyfikacja danych węzła Szeregi czasowe. Z chwilą wywołania prognoz następuje automatyczna budowa modeli autoregresji dla szeregów wejściowych niebędących jednocześnie wartościami przewidywanymi. Modele te są następnie używane do generowania wartości dla tych szeregów wejściowych w okresie prognozy.

Oblicz przyszłe wartości zmiennych wejściowych. W przypadku zaznaczenia tej opcji obliczane są wartości prognozy dla predyktorów, predykcje szumu, oszacowanie wariancji oraz przyszłe wartości czasu.

Wartości przyszłe używane w prognozowaniu

- **Oblicz przyszłe wartości zmiennych wejściowych** W przypadku zaznaczenia tej opcji automatycznie obliczane są wartości prognozy dla predyktorów, predykcje szumu, oszacowanie wariancji oraz przyszłe wartości czasu. Z chwilą wywołania prognoz następuje automatyczna budowa modeli autoregresji dla szeregów wejściowych niebędących jednocześnie wartościami przewidywanymi. Modele te są następnie używane do generowania wartości dla tych szeregów wejściowych w okresie prognozy.
- **Wybierz zmienne, których wartości mają być dodane do danych.** W przypadku każdego rekordu, jaki ma zostać objęty prognozą (z wyjątkiem wstrzymań), korzystanie ze zmiennych predykcyjnych (z rolą ustawioną na Dane wejściowe) pozwala na określenia oszacowanych wartości dla okresu prognozy każdego predyktora. Wartości można określić ręcznie lub wybrać je z listy.
  - **Zmienna.** Należy kliknąć przycisk Selektor zmiennych i wybrać zmienne, które mogą być używane jako predyktory. Należy pamiętać, że wybrane tutaj zmienne mogą, ale nie muszą być używane w modelowaniu; aby rzeczywiście użyć zmiennej jako predyktora, musi być ona wybrana w dalszym węźle modelowania. To okno dialogowe pozwala na określanie przyszłych wartości w wygodny sposób, dzięki czemu mogą one być współużytkowane przez wiele kolejnych węzłów modelowania bez konieczności określania ich w każdym węźle osobno. Należy także pamiętać, że lista dostępnych zmiennych może być ograniczona poprzez wybór na karcie Opcje budowania.

    Należy pamiętać, że jeśli przyszłe wartości są określone dla zmiennej, która nie jest dłużej dostępna w strumieniu (z powodu usunięcia lub aktualizacji wyboru na karcie Opcje budowania), zmienna jest przedstawiona na czerwono.
  - **Wartości.** Dla każdej zmiennej można wybrać funkcje z listy lub kliknąć **Określ**, aby wprowadzić wartości ręcznie lub wybrać je z listy predefiniowanych wartości. Jeśli zmienne predykcyjne odnoszą się do elementów objętych kontrolą użytkownika lub elementów, które są wcześniej znane z innych powodów, wartości należy wprowadzić ręcznie. Przykładowo, jeśli tworzona jest prognoza przychodów hotelu dla kolejnego miesiąca na podstawie liczby zarezerwowanych pokoi, można określić liczbę rezerwacji w rzeczywistości dokonanych dla tego okresu. I odwrotnie, jeśli zmienna predykcyjna odnosi się do elementów poza kontrolą użytkownika, na przykład cena akcji, należy użyć funkcji, takich jak ostatnia wartość lub średnia z ostatnich punktów.

    Dostępne funkcje zależą pod poziomu pomiaru zmiennej.

Tabela 29. Funkcje dostępne dla poziomów pomiaru

Poziom pomiaru	Funkcje
Zmienna ilościowa lub nominalna	Pusta Średnia z ostatnich punktów Ostatnia wartość Określ
Zmienna flagi	Pusta Ostatnia wartość Prawda Fałsz Określ

Średnia z ostatnich punktów oblicza przyszłą wartość na podstawie średniej z ostatnich trzech punktów danych.

Ostatnia wartość ustawia przyszłą wartość w oparciu o najnowszy punkt danych.

Prawda/Fałsz ustawia przyszłą wartość zmiennej typu flaga jako Prawda lub Fałsz.

Określ otwiera okno dialogowe w celu określenia przyszłych wartości ręcznie lub wybrania ich z predefinowanej listy.

Udostępnij do oceniania

Utwórz dla każdego modelu nowe zmienne do oceniania. Umożliwia określenie nowych zmiennych, które mają zostać utworzone dla każdego ocenianego modelu.

- **Reszty szumów.** Zaznaczenie tej opcji powoduje utworzenie nowej zmiennej (z prefiksem domyślnym $TSResidual-) dla reszt modelu dla każdej zmiennej przewidywanej, wraz z sumą tych wartości.
- **Górny i dolny przedział ufności.** Zaznaczenie tej opcji powoduje utworzenie nowych zmiennych (o prefiksach domyślnych $TSLCI- i $TSUCI-) dla dolnego i górnego przedziału ufności, wraz z sumami tych wartości.

Zmienne przewidywane uwzględnione w ocenie. Umożliwia wybór dostępnych wartości przewidywanych do uwzględnienia w ocenie modelu.

Rozdział 14. Modele węzłów odpowiedzi samonauczania

węzeł SLRM

Węzeł **SLRM** (model odpowiedzi samonauczania) umożliwia utworzenie modelu, który można w sposób ciągły aktualizować lub ponownie oceniać w miarę rozwijania zbioru danych, bez konieczności ponownego tworzenia modelu za każdym razem z użyciem kompletnego zbioru danych. Jest to na przykład użyteczne w przypadku dysponowania kilkoma produktami i potrzeby wytypowania produktu, który klient najprawdopodobniej kupi w przypadku złożenia mu oferty. Model ten umożliwia predykcję najlepszego dopasowania ofert do klientów oraz prawdopodobieństwa zaakceptowania tych ofert przez klientów.

Tworzenie modelu można rozpocząć od niewielkiego zbioru danych z losowo złożonymi ofertami oraz odpowiedziami na te oferty. W miarę poszerzania się zbioru danych model może być aktualizowany, co zwiększy jego możliwości w zakresie predykcji najlepszego dopasowania ofert do klientów oraz prawdopodobieństwa zaakceptowania tych ofert w oparciu o pozostałe zmienne wejściowe, takie jak wiek, płeć, zawód i przychód. Dostępność ofert można zmodyfikować, dodając lub usuwając je z okna dialogowego węzła, bez konieczności zmiany zmiennej przewidywanej dla zbioru danych.

Po połączeniu z programem IBM SPSS Collaboration and Deployment Services można skonfigurować automatyczne, regularne aktualizacje modelu. Proces ten, niewymagający inicjatywy ani nadzoru człowieka, stanowi elastyczne i efektywne kosztowo rozwiązanie dla organizacji i zastosowań, w których interwencje ze strony specjalisty ds. eksploracji danych są zbędne lub niemożliwe.

Przykład. Instytucja finansowa chce uzyskać wyższy zysk, dopasowując ofertę o najwyższym prawdopodobieństwie akceptacji do każdego ze swoich klientów. W celu zidentyfikowania charakterystyki klientów, którzy z największym prawdopodobieństwem odniosą się do oferty pozytywnie, można użyć modelu samonauczania oraz poprzednich promocji, a następnie aktualizować model w czasie rzeczywistym, na podstawie ostatnich odpowiedzi klientów.

Opcje zmiennych węzła SLRM

Przed wykonaniem węzła SLRM konieczne jest określenie na karcie Zmienne węzła zarówno zmiennej przewidywanej, jak i zmiennej przewidywanej odpowiedzi.

Zmienna przewidywana. Wybierz zmienną przewidywaną z listy; na przykład zmienna nominalna (ustawiona) zawierająca różne produkty, które mają zostać zaoferowane klientom.

Uwaga: Zmienna przewidywana musi mieć składowanie łańcuchowe, a nie numeryczne.

Zmienna przewidywana odpowiedzi. Wybierz zmienną przewidywaną odpowiedzi z listy. Na przykład: Zaakceptowano lub Odrzucono.

Uwaga: Ta zmienna musi być typu Flaga. Wartość prawda flagi oznacza akceptację oferty, zaś wartość fałsz oznacza jej odrzucenie.

Pozostałe pola w tym oknie dialogowym są polami standardowymi używanymi w programie IBM SPSS Modeler. Więcej informacji można znaleźć w temacie “Opcje zmiennych węzła modelowania” na stronie 31.

Uwaga: Jeśli dane źródłowe obejmują zakresy, które mają być używane jako ilościowe zmienne wejściowe (zakresy liczbowe), wówczas konieczne jest zapewnienie, aby metadane obejmowały zarówno wartość minimalną, jak i maksymalną dla każdego z zakresów.

Opcje modelu węzła SLRM

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Kontynuuj uczenie istniejącego modelu. Domyślnie po każdym wykonaniu węzła modelowania tworzony jest całkowicie nowy model. Jeśli ta opcja jest zaznaczona, uczenie jest kontynuowane z użyciem ostatniego modelu pomyślnie utworzonego przez węzeł. Dzięki temu możliwa jest aktualizacja lub odświeżenie istniejącego modelu bez konieczności uzyskania dostępu do oryginalnych danych, co może skutkować znacznie bardziej wydajnym działaniem, ponieważ *tylko* nowe lub zaktualizowane rekordy są podawane do strumienia. Szczegóły dotyczące poprzedniego modelu są zapisywane z węzłem modelowania, umożliwiając używanie tej opcji nawet, jeśli poprzednie wartościowe informacje z modelu są już niedostępne w strumieniu lub w palecie Modeli.

Wartości zmiennej przewidywanej Domyślnie ta opcja ma wartość **Użyj wszystkich**, co oznacza, że zostanie utworzony model zawierający każdą ofertę powiązaną z wybraną wartością zmiennej przewidywanej. W celu wygenerowania modelu zawierającego tylko niektóre z ofert zmiennych przewidywanych kliknij opcję **Określ** i za pomocą przycisków **Dodaj**, **Edycja** i **Usuń** wprowadź lub zmodyfikuj nazwy ofert, dla których chcesz utworzyć model. Na przykład po wybraniu zmiennej przewidywanej zawierającej listę wszystkich oferowanych produktów można użyć tej zmiennej do ograniczenia oferty produktów do kilku tutaj wprowadzonych.

Ocena modelu. Zmienne w tym panelu są niezależne od modelu, w którym nie wpływają one na ocenianie. Zamiast tego umożliwiają one wizualizację stopnia dokładności przewidywania wyników przez model.

Uwaga: W celu wyświetlenia wyników oceny modelu w modelu użytkowym konieczne jest także zaznaczenie pola wyboru **Ocena modelu**.

- **Uwzględnij ocenę modelu.** Zaznacz to pole wyboru, aby utworzyć wykresy przedstawiające predykowaną dokładność modelu dla każdej wybranej oferty.
- **Ustaw wartość początkową generatora liczb losowych.** W przypadku szacowania dokładności modelu w oparciu o losową wartość procentową opcja ta pozwala na zduplikowanie tych samych wyników w innej sesji. Określenie wartości początkowej używanej przez generator liczb pseudolosowych zapewni, że podczas każdego wykonywania węzła przypisywane będą te same rekordy. Wprowadź żądaną wartość startową generatora. Jeśli nie wybrano tej opcji, przy każdej próbie wykonania węzła wygenerowana zostanie inna próba.
- **Symulowana wielkość próby.** Określ liczbę rekordów, jaka ma być używana w próbie podczas oceniania modelu. Domyślną wartością jest 100.
- **Liczba iteracji.** Opcja ta umożliwia przerwanie tworzenia oceny modelu po osiągnięciu wskazanej liczy iteracji. Określ maksymalną liczbę iteracji; wartość domyślna to 20.

  Uwaga: Należy pamiętać, że próby o dużych rozmiarach oraz wysokiej liczbie iteracji zwiększą ilość czasu niezbędną do utworzenia modelu.

Wyświetl ocenę modelu. Zaznacz tę opcję, aby wyświetlić graficzną reprezentację wyników w modelu użytkowym.

Opcje ustawień węzła SLRM

Opcje ustawień węzła umożliwiają precyzyjne dostosowanie procesu tworzenia modelu.

Maksymalna liczba predykcji na rekord. Ta opcja umożliwia ograniczenie liczby predykcji dla każdego rekordu w zbiorze danych. Domyślną wartością jest 3.

Na przykład dostępnych może być sześć ofert (takich jak oszczędności, hipoteka, kredyt samochodowy, emerytura, karta kredytowa i ubezpieczenie), ale użytkownik chce zarekomendować tylko dwie najlepsze; w tym przypadku

zmienną należy ustawić na 2. Podczas budowania modelu i dołączania go do tabeli widoczne będą dwie kolumny predykcji (i powiązana ufność dla prawdopodobieństwa zaakceptowania oferty) dla rekordu. Predykcje mogą być utworzone dla dowolnej z sześciu możliwych ofert.

Poziom losowości. Aby zapobiec odchyleniom — na przykład w małym lub niekompletnym zbiorze danych — i traktować jednakowo wszystkie potencjalne oferty, można dodać poziom losowości dla wyboru ofert i prawdopodobieństwa, że zostaną uwzględnione jako oferty rekomendowane. Losowość jest wyrażana jako wartość procentowa, przedstawiana jako wartości dziesiętne z przedziału od 0,0 (brak losowości) do 1,0 (całkowicie losowo). Domyślną wartością jest 0,0.

Ustaw wartość początkową generatora liczb losowych. Podczas dodawania poziomu losowości do wyboru oferty ta opcja umożliwi zduplikowanie tych samych wyników w kolejnej sesji. Określenie wartości początkowej używanej przez generator liczb pseudolosowych zapewni, że podczas każdego wykonywania węzła przypisywane będą te same rekordy. Wprowadź żądaną wartość startową generatora. Jeśli nie wybrano tej opcji, przy każdej próbie wykonania węzła wygenerowana zostanie inna próba.

Uwaga: Jeśli używana jest opcja **Ustaw wartość początkową generatora liczb losowych** w przypadku rekordów odczytanych z bazy danych, przed przeprowadzeniem próby konieczne może być sortowanie węzła, aby po każdym wykonaniu węzła uzyskany wynik był taki sam. Wynika to z faktu, że wartość początkowa generatora liczb losowych zależy od kolejności rekordów, która w relacyjnej bazie danych nie musi pozostawać jednakowa.

Porządek sortowania. Należy wybrać porządek, w jakim oferty będą wyświetlane w modelu budowania:

- **Malejąco.** Model wyświetla oferty z najwyższymi ocenami jako pierwsze. Są to oferty z najwyższym prawdopodobieństwem zaakceptowania.
- **Rosnąco.** Model wyświetla oferty z najniższymi ocenami jako pierwsze. Są to oferty z najwyższym prawdopodobieństwem odrzucenia. Może to być na przykład przydatne podczas podejmowania decyzji, których klientów należy usunąć z kampanii marketingowej dla konkretnej oferty.

Preferencje dla zmiennych przewidywanych. Podczas budowania modelu mogą wystąpić pewne aspekty dotyczące danych, które użytkownik chce aktywnie promować lub usuwać. Przykładowo, jeśli budowany jest model, który wybiera najlepszą ofertę finansową, jaka będzie promowana dla klienta, użytkownik chce mieć pewność, że jedna konkretna oferta będzie zawsze uwzględniona, niezależnie od jej wyniku w odniesieniu do poszczególnych klientów.

Aby uwzględnić ofertę w tym panelu i edytować jej preferencje, należy kliknąć przycisk **Dodaj**, wpisać nazwę oferty (na przykład Oszczędności lub Hipoteka) i kliknąć przycisk **OK**.

- **Wartość.** Wyświetla nazwę dodanej oferty.
- **Preferencja.** Określa poziom preferencji, jaki będzie zastosowany do oferty. Preferencje są wyrażane jako wartość procentowa, wyrażona jako wartości dziesiętne z przedziału od 0,0 (nie preferowane) do 1,0 (najbardziej preferowane). Domyślną wartością jest 0,0.
- **Zawsze uwzględniaj.** Aby upewnić się, że konkretna oferta będzie zawsze uwzględniona w predykcjach, należy zaznaczyć to pole wyboru.

  Uwaga: Jeśli opcja **Preferencje** jest ustawiona na wartość 0,0, ustawienie **Zawsze uwzględniaj** jest ignorowane.

Uwzględniaj rzetelność modelu. Dobrze ustrukturyzowany, zawierający wiele danych model, który został dostosowany poprzez kilkukrotne ponowne wygenerowanie zawsze powinien zapewniać bardziej dokładne wyniki w porównaniu do całkiem nowego modelu, który zawiera niewiele danych. Aby skorzystać ze zwiększonej rzetelności bardziej dojrzałego modelu, należy zaznaczyć to pole wyboru.

Modele użytkowe SLRM

Uwaga: Wyniki są wyświetlane na tej karcie tylko pod warunkiem zaznaczenia na karcie opcji modelu opcji **Uwzględnij ocenę modelu** i **Wyświetl ocenę modelu**.

Po uruchomieniu strumienia zawierającego model SLRM węzeł szacuje dokładność predykcji dla każdej wartości zmiennej przewidywanej (oferty) oraz ważność każdego użytego predyktora.

Uwaga: Jeśli wybrano opcję **Kontynuuj uczenie istniejącego modelu** na karcie Model węzła modelowania, informacje wyświetlane w modelu użytkowym są aktualizowane każdorazowo podczas ponownego generowania modelu.

W przypadku modeli tworzonych za pomocą programu IBM SPSS Modeler 12.0 lub nowszej wersji karta Model modelu użytkowego jest podzielona na dwie kolumny:

Lewa kolumna.

- **Widok.** W przypadku dysponowania więcej niż jedną ofertą wybierz taką, dla której mają zostać wyświetlone wyniki.
- **Wynik działania modelu.** Przedstawia szacowaną dokładność modelu dla każdej oferty. Zestaw testowy jest generowany w drodze symulacji.

Prawa kolumna.

- **Widok.** Zaznacz, czy chcesz wyświetlić dane **Powiązanie z odpowiedzią** czy **Ważność zmiennych**.
- **Powiązanie z odpowiedzią.** Wyświetla powiązanie (korelację) każdego predyktora ze zmienną przewidywaną.
- **Ważność predyktorów.** Określa względną wartość każdego predyktora przy oszacowywaniu modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Wykres ten można interpretować tak samo, jak w przypadku innych modeli, prezentujących ważność predyktorów; w przypadku opcji SLRM wykres jest jednak generowany w drodze symulacji, przez algorytm SLRM. W tym celu z modelu usuwany jest każdy kolejny predyktor, i obserwowany jest wpływ tego usunięcia na dokładność modelu. Więcej informacji można znaleźć w temacie “Ważność predyktorów” na stronie 43.

Ustawienia modelu SLRM

Karta Ustawienia modelu użytkowego SLRM określa opcje modyfikacji budowanego modelu. Na przykład węzeł SLRM może służyć do budowy kilku różnych modeli z użyciem tych samych danych i ustawień. Następnie, za pomocą tej karty dla każdego z modeli można nieznacznie zmodyfikować ustawienia, obserwując jednocześnie, jak wpłynie to na wyniki.

Uwaga: Ta karta jest dostępna tylko po dodaniu modelu użytkowego do strumienia.

Maksymalna liczba predykcji na rekord. Ta opcja umożliwia ograniczenie liczby predykcji dla każdego rekordu w zbiorze danych. Domyślną wartością jest 3.

Na przykład dostępnych może być sześć ofert (takich jak oszczędności, hipoteka, kredyt samochodowy, emerytura, karta kredytowa i ubezpieczenie), ale użytkownik chce zarekomendować tylko dwie najlepsze; w tym przypadku zmienną należy ustawić na 2. Podczas budowania modelu i dołączania go do tabeli widoczne będą dwie kolumny predykcji (i powiązana ufność dla prawdopodobieństwa zaakceptowania oferty) dla rekordu. Predykcje mogą być utworzone dla dowolnej z sześciu możliwych ofert.

Poziom losowości. Aby zapobiec odchyleniom — na przykład w małym lub niekompletnym zbiorze danych — i traktować jednakowo wszystkie potencjalne oferty, można dodać poziom losowości dla wyboru ofert i prawdopodobieństwa, że zostaną uwzględnione jako oferty rekomendowane. Losowość jest wyrażana jako wartość procentowa, przedstawiana jako wartości dziesiętne z przedziału od 0,0 (brak losowości) do 1,0 (całkowicie losowo). Domyślną wartością jest 0,0.

Ustaw wartość początkową generatora liczb losowych. Podczas dodawania poziomu losowości do wyboru oferty ta opcja umożliwi zduplikowanie tych samych wyników w kolejnej sesji. Określenie wartości początkowej używanej

przez generator liczb pseudolosowych zapewni, że podczas każdego wykonywania węzła przypisywane będą te same rekordy. Wprowadź żądaną wartość startową generatora. Jeśli nie wybrano tej opcji, przy każdej próbie wykonania węzła wygenerowana zostanie inna próba.

Uwaga: Jeśli używana jest opcja **Ustaw wartość początkową generatora liczb losowych** w przypadku rekordów odczytanych z bazy danych, przed przeprowadzeniem próby konieczne może być sortowanie węzła, aby po każdym wykonaniu węzła uzyskany wynik był taki sam. Wynika to z faktu, że wartość początkowa generatora liczb losowych zależy od kolejności rekordów, która w relacyjnej bazie danych nie musi pozostawać jednakowa.

Porządek sortowania. Należy wybrać porządek, w jakim oferty będą wyświetlane w modelu budowania:

- **Malejąco.** Model wyświetla oferty z najwyższymi ocenami jako pierwsze. Są to oferty z najwyższym prawdopodobieństwem zaakceptowania.
- **Rosnąco.** Model wyświetla oferty z najniższymi ocenami jako pierwsze. Są to oferty z najwyższym prawdopodobieństwem odrzucenia. Może to być na przykład przydatne podczas podejmowania decyzji, których klientów należy usunąć z kampanii marketingowej dla konkretnej oferty.

Preferencje dla zmiennych przewidywanych. Podczas budowania modelu mogą wystąpić pewne aspekty dotyczące danych, które użytkownik chce aktywnie promować lub usuwać. Przykładowo, jeśli budowany jest model, który wybiera najlepszą ofertę finansową, jaka będzie promowana dla klienta, użytkownik chce mieć pewność, że jedna konkretna oferta będzie zawsze uwzględniona, niezależnie od jej wyniku w odniesieniu do poszczególnych klientów.

Aby uwzględnić ofertę w tym panelu i edytować jej preferencje, należy kliknąć przycisk **Dodaj**, wpisać nazwę oferty (na przykład Oszczędności lub Hipoteka) i kliknąć przycisk **OK**.

- **Wartość.** Wyświetla nazwę dodanej oferty.
- **Preferencja.** Określa poziom preferencji, jaki będzie zastosowany do oferty. Preferencje są wyrażane jako wartość procentowa, wyrażona jako wartości dziesiętne z przedziału od 0,0 (nie preferowane) do 1,0 (najbardziej preferowane). Domyślną wartością jest 0,0.
- **Zawsze uwzględniaj.** Aby upewnić się, że konkretna oferta będzie zawsze uwzględniona w predykcjach, należy zaznaczyć to pole wyboru.

  Uwaga: Jeśli opcja **Preferencje** jest ustawiona na wartość 0,0, ustawienie **Zawsze uwzględniaj** jest ignorowane.

Uwzględniaj rzetelność modelu. Dobrze ustrukturyzowany, zawierający wiele danych model, który został dostosowany poprzez kilkukrotne ponowne wygenerowanie zawsze powinien zapewniać bardziej dokładne wyniki w porównaniu do całkiem nowego modelu, który zawiera niewiele danych. Aby skorzystać ze zwiększonej rzetelności bardziej dojrzałego modelu, należy zaznaczyć to pole wyboru.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Rozdział 15. Modele SVM

Informacje o algorytmie SVM

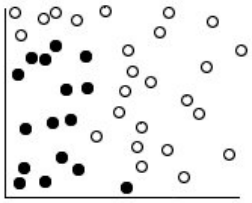
Algorytm SVM to technika klasyfikacji i regresji maksymalizująca dokładność predykcyjną modelu bez przeuczenia danych uczących. Algorytm SVM szczególnie dobrze nadaje się do analizowania danych o bardzo dużej liczbie (np. tysiącach) zmiennych predykcyjnych.

Algorytm SVM znajduje zastosowanie w wielu dziedzinach, takich jak zarządzanie relacjami z klientami (CRM), rozpoznawanie twarzy i innych obrazów, bioinformatyka, wyodrębnianie koncepcji w drodze eksploracji teksu, wykrywanie wtargnięć, przewidywanie struktury białek oraz rozpoznawanie głosu i mowy.

Sposób działania algorytmu SVM

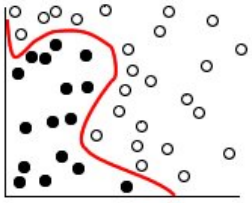
Działanie algorytmu SVM polega na mapowaniu danych na wielowymiarową przestrzeń właściwości w sposób umożliwiający kategoryzację punktów danych, nawet jeśli danych tych nie można w inny sposób liniowo oddzielić. Najpierw odszukiwany jest separator między kategoriami. Następnie dane są przekształcane w sposób umożliwiający wyrysowanie separatora jako hiperpłaszczyzny. Po wykonaniu tych czynności charakterystyki nowych danych mogą służyć do przewidywania grupy, do której powinien należeć nowy rekord.

Rozważmy na przykład poniższy rysunek, na którym punkty danych przypadają do dwu różnych kategorii.



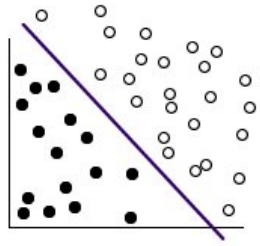
Rysunek 59. Oryginalny zbiór danych

Te dwie kategorie można oddzielić krzywą, tak jak pokazano na poniższym rysunku.



Rysunek 60. Dane z dodanym separatorem

Po dokonaniu przekształcenia granicę między dwiema kategoriami można zdefiniować za pomocą hiperpłaszczyzny, tak jak przedstawiono to na poniższym rysunku.



Rysunek 61. Przekształcone dane

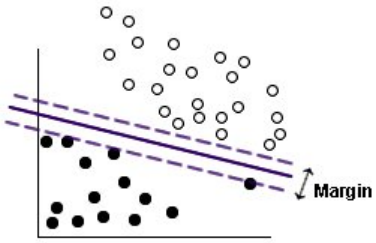
Funkcja matematyczna używana do przekształceń jest znana jako **algorytm domyślny**. Algorytm SVM w programie IBM SPSS Modeler obsługuje następujące typy algorytmów domyślnych:

- Liniowy
- Wielomianowy
- Radialna funkcja bazowa (RBF)
- Sigmoidalny

Funkcja algorytmu domyślnego jest zalecana, o ile liniowa separacja danych jest nieskomplikowana. W innych przypadkach należy użyć jednej z pozostałych funkcji. Konieczne będzie wypróbowanie różnych funkcji w celu uzyskania najlepszego modelu w przypadku każdej obserwacji, ponieważ każda z nich wykorzystuje inne algorytmy i parametry.

Precyzyjne dostosowywanie modelu SVM

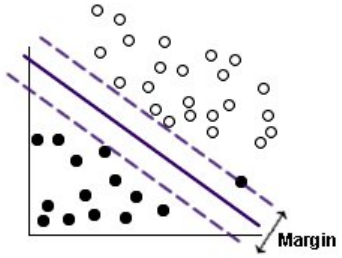
Poza linią oddzielającą kategorie model klasyfikacji SVM znajduje także linie marginesu definiujące przestrzeń między dwiema kategoriami.



Rysunek 62. Dane z modelem wstępnym

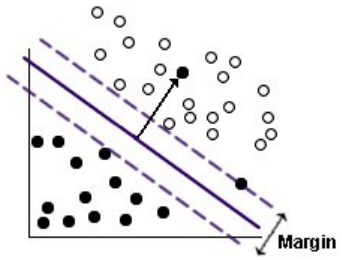
Punkty danych leżące na brzegach są zwane **wektorami pokrycia**.

Im szerszy margines między dwiema kategoriami, tym lepsze będą wyniki predykcji kategorii dla nowych rekordów. W poprzednim przykładzie margines nie był zbyt szeroki, a model został określony jako **przeuczony**. Istnieje możliwość akceptacji klasyfikacji błędnej w niewielkim zakresie z myślą o poszerzeniu marginesu; stosowny przykład zamieszczono na poniższym rysunku.



Rysunek 63. Dane z poprawionym modelem

W niektórych przypadkach separacja liniowa jest znacznie trudniejsza; przykład zamieszczono na poniższym rysunku.



Rysunek 64. Problem dotyczący separacji liniowej

W przypadkach takich jak ten celem jest znalezienie optymalnego zrównoważenia między szerokością marginesu a liczbą błędnie sklasyfikowanych punktów danych. Funkcja algorytmu domyślnego ma **parametr regularyzacji** (znany jako C) decydujący o wypośrodkowaniu między tymi dwiema wartościami. Znalezienie najlepszego modelu będzie prawdopodobnie wymagało wypróbowania różnych wartości tego i innych parametrów algorytmu domyślnego.

Węzeł SVM

Węzeł SVM umożliwia wykorzystanie algorytmu SVM do klasyfikacji danych. Algorytm SVM szczególnie dobrze nadaje się do użytku z obszernymi zbiorami danych, to jest, ze zbiorami o dużej liczbie zmiennych predykcyjnych. Ustawienia domyślne można wykorzystać w węźle do generowania relatywnie szybko modelu podstawowego. Można także użyć ustawień zaawansowanych do wypróbowania różnych typów algorytmów SVM.

Po zbudowaniu modelu można:

- Przeglądać model użytkowy w celu wyświetlenia względnej ważności zmiennych wejściowych w budowie modelu.
- Dołączyć węzeł Tabela do modelu użytkowego w celu wyświetlenia wyników modelu.

Przykład. Pracownik naukowo-badawczy zgromadził zbiór danych zawierający charakterystykę pewnej liczby prób komórek ludzkich pobranych od pacjentów z podejrzeniem nowotworu. Jak wykazała analiza oryginalnych danych, wiele z charakterystyk próbek z nowotworem złośliwym różniło się istotnie od próbek z nowotworem niezłośliwym. Pracownik chce opracować model SVM wykorzystujący wartości podobnych charakterystyk komórek w próbkach od innych pacjentów w celu wstępnego określania, czy ich próbki zawierają nowotwór złośliwy, czy niezłośliwy.

Opcje modelu węzła SVM

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Utwórz modele rozdzielone. Tworzy osobny model dla każdej możliwej wartości zmiennych wejściowych, jakie są określone jako zmienne podziału. Więcej informacji można znaleźć w “Budowanie modeli rozdzielonych” na stronie 28.

Opcje zaawansowane węzła SVM

Użytkownikom dysponującym gruntowną wiedzą na temat algorytmów SVM opcje zaawansowane umożliwiają precyzyjne dostosowanie procesu uczenia. W celu uzyskania dostępu do opcji zaawansowanych należy ustawić opcję Tryb na wartość **Zaawansowany** na karcie Zaawansowany.

Dołącz wszystkie prawdopodobieństwa (opcja poprawna tylko w przypadku przewidywanych zmiennych jakościowych). Zaznaczenie tej opcji powoduje, że dla każdego rekordu przetwarzanego przez węzeł wyświetlane są prawdopodobieństwa każdej możliwej wartości zmiennej przewidywanej nominalnej i typu flaga. Jeśli ta opcja nie zostanie zaznaczona, dla zmiennych przewidywanych nominalnych i typu flaga wyświetlane będzie tylko prawdopodobieństwo wartości predykcyjnej. Ustawienie tego pola wyboru określa domyślny stan odpowiedniego pola wyboru na ekranie modelu użytkowego.

Kryteria zatrzymania. Ta opcja określa moment zatrzymania algorytmu optymalizacji. Wartości mieszczą się w zakresie od 1,0E–1 do 1,0E–6; wartością domyślną jest 1,0E–3. Zmniejszenie tej wartości skutkuje większą dokładnością modelu; jednocześnie uczenie modelu może zająć więcej czasu.

Parametr Regularyzacja (C). Decyduje o wypośrodkowaniu między maksymalizacją marży a minimalizacją składnika błędu uczenia. Wartość powinna normalnie mieścić się w przedziale od 1 do 10 włącznie; wartość domyślna to 10. Zwiększenie wartości poprawia dokładność klasyfikacji (lub obniża błąd regresji) danych uczących, lecz może także doprowadzić do przeuczenia.

Precyzja regresji (epsilon). Jest używana tylko, jeśli poziom pomiaru zmiennej przewidywanej to *Ilościowa*. Powoduje akceptację błędów, pod warunkiem że są one mniejsze od podanej tutaj wartości. Zwiększenie wartości może skutkować szybszym modelowaniem, lecz za cenę dokładności.

Typ algorytmu domyślnego. Określa typ funkcji algorytmu domyślnego używanego do przekształceń. Różne typy algorytmu domyślnego powodują obliczanie separatora na różne sposoby, dlatego zaleca się wypróbowanie różnych opcji. Wartość domyślna to **RBF** (Radial Basis Function).

RBF gamma. Opcja aktywna tylko w przypadku ustawienia typu algorytmu domyślnego na wartość **RBF**. Wartość powinna mieścić się w przedziale od 3/*k* do 6/*k*, gdzie *k* jest liczbą zmiennych wejściowych. Na przykład w przypadku 12 zmiennych wejściowych warto wypróbować wartości z przedziału od 0,25 do 0,5. Zwiększenie wartości poprawia dokładność klasyfikacji (lub obniża błąd regresji) danych uczących, lecz może także doprowadzić do przeuczenia.

Gamma. Opcja aktywna tylko w przypadku ustawienia typu algorytmu domyślnego na wartość **Wielomianowy** lub **Sigmoidalny**. Zwiększenie wartości poprawia dokładność klasyfikacji (lub obniża błąd regresji) danych uczących, lecz może także doprowadzić do przeuczenia.

Odchylenie. Opcja aktywna tylko w przypadku ustawienia typu algorytmu domyślnego na wartość **Wielomianowy** lub **Sigmoidalny**. Ustawia wartość coef0 w funkcji algorytmu domyślnego. Wartość domyślna wynosi 0 i jest odpowiednia w większości przypadków.

Stopień. Opcja aktywna tylko w przypadku ustawienia typu algorytmu domyślnego na wartość **Wielomianowy**. Decyduje o złożoności (liczbie wymiarów) przestrzeni mapowania. W normalnym przypadku nie używa się wartości większej niż 10.

Model użytkowy SVM

Model użytkowy SVM tworzy pewną liczbę nowych zmiennych. Najważniejsza z nich to zmienna **$S-fieldname**, przedstawiająca wartość zmiennej przewidywanej predykowaną przez model.

Liczba i nazwy nowych zmiennych tworzonych przez model zależą od poziomu pomiaru zmiennej przewidywanej (zmienna ta jest określana w poniższych tabelach jako *fieldname*).

W celu wyświetlenia tych zmiennych oraz ich wartości należy dodać węzeł Tabela do modelu użytkowego SVM i uruchomić węzeł Tabela.

Tabela 30. Poziom pomiaru zmiennej przewidywanej to 'Nominalna' lub 'Flaga'.

Nazwa nowej zmiennej	**Opis**
$S-*fieldname*	Predykowana wartość zmiennej przewidywanej.
$SP-*fieldname*	Prawdopodobieństwo wartości przewidywanej.
$SP-*value*	Prawdopodobieństwo każdej możliwej wartości typu nominalna lub flaga (wyświetlane tylko, jeśli na karcie Ustawienia modelu użytkowego zaznaczono opcję **Dołącz wszystkie prawdopodobieństwa**).
$SRP-*value*	(Tylko zmienne przewidywane typu flaga) Nieprzetworzone (SRP) i skorygowane (SAP) oceny skłonności, określające prawdopodobieństwo wyniku „prawda" dla zmiennej przewidywanej. Oceny te są wyświetlane tylko w przypadku zaznaczenia pól wyboru na karcie Analiza węzła modelowania przed wygenerowaniem modelu. Więcej informacji można znaleźć w temacie "Opcje analizowania węzła modelowania" na stronie 34.
$SAP-*value*	

Tabela 31. Poziom pomiaru zmiennej przewidywanej to „Ilościowa"

Nazwa nowej zmiennej	**Opis**
$S-*fieldname*	Predykowana wartość zmiennej przewidywanej.

Ważność predyktorów

Opcjonalnie na karcie Model może być również wyświetlany wykres przedstawiający względną ważność poszczególnych predyktorów w oszacowaniu modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Należy zwrócić uwagę, że ten wykres jest dostępny tylko po wybraniu opcji **Oblicz ważność predyktora** na karcie Analiza przed wygenerowaniem modelu. Więcej informacji można znaleźć w temacie "Ważność predyktorów" na stronie 43.

Uwaga: ważność predyktorów może wydłużyć czas potrzebny na obliczenia wg algorytmu SVM w porównaniu z obliczeniami innych typów modeli. Dlatego domyślnie odpowiednia opcja na karcie Analiza jest niezaznaczona. Wybranie tej opcji może spowodować spowolnienie działania, szczególnie w przypadku dużych zbiorów danych.

Ustawienia modelu SVM

Karta Ustawienia umożliwia określenie dodatkowych zmiennych, jakie będą wyświetlane podczas przeglądania listy wyników (np. w wyniku wykonania węzła Tabela dołączonego do modelu użytkowego). Można zobaczyć efekt działania poszczególnych opcji; w tym celu należy je zaznaczyć i kliknąć przycisk Podgląd — aby zobaczyć dodatkowe zmienne należy przewinąć w prawo wyniki podglądu.

Dołącz wszystkie prawdopodobieństwa (tylko dla przewidywanych zmiennych jakościowych). Jeśli ta opcja jest zaznaczona, prawdopodobieństwa poszczególnych możliwych wartości zmiennej nominalnej lub zmiennej przewidywanej typu flaga są wyświetlane dla każdego rekordu przetworzonego przez węzeł. Jeśli ta opcja nie jest zaznaczona, wówczas dla zmiennych nominalnych i przewidywanych typu flaga wyświetlana będzie tylko wartość przewidywana i jej prawdopodobieństwo.

Ustawienie domyślne dla tego pola wyboru jest określane przez odpowiednie pole w modelu użytkowym.

Wylicz surowe oceny skłonności. W przypadku modeli z przewidywaną zmienną typu flaga (zwracających predykcje tak lub nie) można wyliczyć oceny skłonności definiujące prawdopodobieństwo prawdziwego wyniku określonego dla zmiennej przewidywanej. Stanowią one uzupełnienie pozostałych współczynników ufności i wartości predykcyjnych, jakie mogą zostać wygenerowane podczas oceniania.

Wylicz skorygowane oceny skłonności. Surowe oceny skłonności bazują na danych uczących i mogą być zbyt optymistyczne z uwagi na tendencję wielu modeli do przeuczania tych danych. Opcja wyliczania skłonności skorygowanych próbuje skompensować tę tendencję na drodze oceny wydajności modelu w teście lub w podzbiorze walidacyjnym. Opcja ta wymaga, aby zmienna dzieląca na podzbiory była zdefiniowana w strumieniu oraz aby skorygowane oceny skłonności były włączone w węźle modelowania przed przystąpieniem do generowania modelu.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślnie: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych** Po wybraniu ta opcja powoduje pobieranie danych z bazy danych i ocenia je w SPSS Modeler.

Węzeł LSVM

Węzeł LSVM umożliwia wykorzystanie algorytmu liniowej maszyny wektorów nośnych do klasyfikacji danych. Algorytm LSVM szczególnie dobrze nadaje się do użytku z obszernymi zbiorami danych, to jest, ze zbiorami o dużej liczbie zmiennych predykcyjnych. Ustawienia domyślne można wykorzystać w węźle do relatywnie szybkiego wygenerowania modelu podstawowego. Można także użyć opcji budowania do eksperymentowania z różnymi ustawieniami.

Węzeł LSVM jest podobny do węzła SVM, ale ma charakterystykę liniową i lepiej radzi sobie z obsługą dużej liczby rekordów.

Po zbudowaniu modelu można:

- Przeglądać model użytkowy w celu wyświetlenia względnej ważności zmiennych wejściowych w budowie modelu.
- Dołączyć węzeł Tabela do modelu użytkowego w celu wyświetlenia wyników modelu.

Przykład. Pracownik naukowo-badawczy zgromadził zbiór danych zawierający charakterystykę pewnej liczby prób komórek ludzkich pobranych od pacjentów z podejrzeniem nowotworu. Jak wykazała analiza oryginalnych danych, wiele z charakterystyk próbek z nowotworem złośliwym różniło się istotnie od próbek z nowotworem niezłośliwym. Naukowiec chce opracować model LSVM wykorzystujący wartości podobnych cech komórek w próbkach od innych pacjentów w celu wstępnego określania, czy próbki zawierają nowotwór złośliwy, czy łagodny.

Opcje modelu węzła LSVM

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Oblicz ważność predyktora. W przypadku modeli generujących odpowiednią miarę ważności możliwe jest wyświetlenie tabeli wskazującej ważność względną każdego predyktora w procesie estymacji modelu. Zazwyczaj działania modelujące mają koncentrować się na predyktorach, które są najważniejsze, a opuszczane lub ignorowane mają być te predyktory, które są najmniej ważne. Należy zauważyć, że obliczenie ważności predyktora może potrwać

dłużej dla niektórych modeli, szczególnie w przypadku pracy z dużymi zbiorami danych, i domyślnie ta opcja dla niektórych modeli jest wyłączona. Ważność predyktorów jest niedostępna dla modeli listy decyzyjnej. Więcej informacji można znaleźć w “Ważność predyktorów” na stronie 43.

Opcje budowania węzła LSVM

Ustawienia modelu

Uwzględnij wyraz wolny. Uwzględnienie wyrazu wolnego (stały składnik modelu) może spowodować zwiększenie ogólnej dokładności rozwiązania. Jeśli można założyć, że dane przechodzą przez początek układu współrzędnych, wyraz wolny można wyłączyć z modelu.

Porządek sortowania jakościowych zmiennych przewidywanych. Określa porządek sortowania jakościowych zmiennych przewidywanych. To ustawienie jest ignorowane w przypadku ilościowych zmiennych przewidywanych.

Precyzja regresji (epsilon). Jest używana tylko, jeśli poziom pomiaru zmiennej przewidywanej to *Ilościowa*. Powoduje akceptację błędów, pod warunkiem że są one mniejsze od podanej tutaj wartości. Zwiększenie wartości może skutkować szybszym modelowaniem, lecz za cenę dokładności.

Wyklucz rekordy z brakami danych. Gdy ta opcja jest ustawiona na **Prawda**, rekord jest wykluczany, jeśli brakuje choć jednej wartości.

Ustawienia kary

Funkcja kary. Określa typ używanej funkcji kary w celu zniwelowania ryzyka przeuczenia. Dostępne opcje to **E1** i **E2**.

Opcje **E1** i **E2** niwelują ryzyko przeuczenia, nakładając parametr kary na współczynniki. Różnica między nimi polega na tym, że w przypadku dużej liczby predyktorów opcja **E1** wybiera predyktora poprzez ustawienie dla niektórych współczynników wartości 0 podczas budowania modelu. Opcja **E2** nie ma takiej możliwości, dlatego nie należy jej stosować w przypadku dużej liczby predyktorów.

Parametr kary (lambda). Określa parametr kary (regularyzacji). To ustawienie jest aktywne, jeśli włączona jest **Funkcja kary**.

Model użytkowy LSVM (wyniki interaktywne)

Po uruchomieniu modelu LSVM dostępne są następujące wyniki.

Informacje o modelu

Widok Informacje o modelu zawiera kluczowe informacje o modelu. Tabela określa niektóre ustawienia modelu wysokiego poziomu, takie jak:

- Nazwa zmiennej przewidywanej określona na karcie Zmienne
- Metoda budowania modelu określona w ustawieniach Wybór modelu
- Liczba predyktorów w danych wejściowych
- Liczba predyktorów użytych w modelu końcowym
- Typ regularyzacji (L1 lub L2)
- Parametr kary (lambda). Jest to parametr regularyzacji.
- Precyzja regresji (epsilon). Akceptowane są błędy o wartości niższej od podanej tutaj. Podanie wyższej wartości może skutkować szybszym modelowaniem, lecz za cenę dokładności. Opcja używana tylko w sytuacji, gdy poziom pomiaru zmiennej przewidywanej to *Ilościowa*.
- Procentowo określona dokładność klasyfikacji. Ma zastosowanie tylko do klasyfikacji.
- Błąd średniokwadratowy. Ma zastosowanie tylko do regresji.

Podsumowanie rekordów

Widok Podsumowanie rekordów przedstawia informacje o liczbie i wartości procentowej rekordów (obserwacji) uwzględnionych w modelu i wykluczonych z modelu.

Ważność predyktorów

Zazwyczaj działania modelujące mają koncentrować się na zmiennych predyktorów, które są najważniejsze, a opuszczane lub ignorowane mają być te zmienne, które są najmniej ważne. Wykres ważności predyktorów pomaga osiągnąć ten cel przez wskazanie względnej ważności każdego predyktora przy szacowaniu modelu. Ponieważ wartości są względne, suma wartości wszystkich wyświetlanych predyktorów wynosi 1,0. Ważność predyktora nie jest powiązana z dokładnością modelu. Jest powiązana z ważnością każdego predyktora przy prognozach, a nie z tym, czy taka prognoza jest dokładna.

Przewidywane według obserwowanych

Przedstawia on wykres rozrzutu z kategoryzacją przewidywanych wartości na osi pionowej przez obserwowane wartości na osi poziomej. W idealnym przypadku punkty te powinny leżeć na prostej nachylonej pod kątem 45 stopni; widok ten może stwierdzić, czy którekolwiek z wyników zostały przewidziane przez model w sposób oczywisty.

Uwaga: W przypadku modeli LSVM i SVM obliczanie ważności predyktorów może zająć więcej czasu niż w innych modelach. Wybranie tej opcji może spowodować spowolnienie działania, szczególnie w przypadku dużych zbiorów danych.

Macierz pomyłek

Macierz pomyłek, nazywana niekiedy tabelą podsumowań, zawiera liczbę obserwacji prawidłowo i nieprawidłowo przypisanych do każdej grupy na podstawie analizy LSVM.

Ustawienia modelu LSVM

Karta Ustawienia modelu użytkowego LSVM umożliwia określenie opcji surowej skłonności oraz generowania kodu SQL podczas oceny modelu. Ta karta jest dostępna tylko po dodaniu modelu użytkowego do strumienia.

Wylicz surowe oceny skłonności W przypadku modeli zawierających zmienne przewidywane typu flaga można zażądać surowych ocen skłonności, które wskazują wiarygodność wyniku prawda określonego dla zmiennej przewidywanej. Te oceny są stosowane dodatkowo obok standardowych wartości predykcji i ufności. Skorygowane oceny skłonności są niedostępne.

Generuj kod SQL dla tego modelu Korzystając z danych z bazy danych, kod SQL może zostać skierowany do bazy danych w celu wykonania, zapewniając lepszą wydajność dla wielu operacji.

Aby określić sposób generowania kodu SQL, wybierz jedną z następujących opcji.

- **Domyślne: Przeprowadź ocenę, używając składnika Server Scoring Adapter (o ile jest zainstalowany), w przeciwnym wypadku w trakcie przetwarzania.** Jeśli dostępne jest połączenie z bazą danych i jest zainstalowany składnik Scoring Adapter, wówczas ta opcja powoduje wygenerowanie kodu SQL z użyciem tego składnika oraz powiązanych funkcji zdefiniowanych przez użytkownika, a następnie ocenia model użytkownika w bazie danych. Jeśli składnik Scoring Adapter nie jest dostępny, ta opcja pobiera dane z bazy danych i ocenia je w programie SPSS Modeler.
- **Przeprowadź ocenę poza bazą danych.** Jeśli ta opcja jest wybrana, dane użytkownika pobierane są z bazy danych i oceniane w produkcie SPSS Modeler.

Rozdział 16. Modele najbliższego sąsiedztwa

Węzeł KNN

Analiza najbliższego sąsiedztwa jest metodą klasyfikacji obserwacji na podstawie ich podobieństwa do innych obserwacji. Zostało to opracowane w nauczaniu maszynowym jako sposób rozpoznawania wzorców danych bez konieczności zapewnienia dokładnej zgodności z jakimikolwiek zapamiętanymi wzorcami lub obserwacjami. Podobne obserwacje znajdują się blisko siebie, a niepodobne — daleko. Zatem odległość między dwoma obserwacjami stanowi miarę ich niepodobieństwa.

Obserwacje znajdujące się blisko siebie nazywają się „sąsiedztwem". Podczas prezentacji nowej (wstrzymanej) obserwacji, obliczana jest odległość od każdej obserwacji modelu. Zostaje określona klasyfikacja najbardziej podobnych obserwacji najbliższego sąsiedztwa, a nowa obserwacja zostaje umieszczona w kategorii, która zawiera największą liczbę obserwacji najbliższego sąsiedztwa.

Można określić liczbę najbliższych elementów sąsiednich do analizowania; ta wartość to *k*. Rysunki przedstawiają, jak nowa obserwacja będzie sklasyfikowana za pomocą dwóch różnych wartości *k*. Jeśli $k = 5$, nowa obserwacja jest umieszczana w kategorii *1*, ponieważ większość najbliższych elementów sąsiednich należy do kategorii *1*. Jeśli jednak $k = 9$, nowa obserwacja jest umieszczana w kategorii *0*, ponieważ większość najbliższych elementów sąsiednich należy do kategorii *0*.

Analiza najbliższego sąsiedztwa może być również użyta do obliczania docelowych wartości ilościowych. W tej sytuacji do uzyskania przewidywanej wartości dla nowej obserwacji stosowana jest docelowa wartość średniej lub mediany najbliższych sąsiadów.

Opcje celów węzła KNN

Karta Cele to miejsce, w którym można wybrać, czy budowany model będzie przewidywał wartość zmiennej przewidywanej w danych wejściowych na podstawie wartości najbliższego sąsiedztwa, albo można po prostu znaleźć najbliższe sąsiedztwo dla konkretnej interesującej obserwacji.

Jakiego typu analizę zamierzasz przeprowadzić?

Przewidywanie zmiennej przewidywanej. Wybierz tę opcję, jeśli zamierzasz przewidzieć wartość zmiennej przewidywanej na podstawie wartości jej najbliższego sąsiedztwa.

Tylko zidentyfikuj najbliższych sąsiadów. Wybierz tę opcję, jeśli zamierzasz tylko sprawdzić najbliższe sąsiedztwo dla konkretnej zmiennej wejściowej.

Jeśli wybierzesz identyfikację tylko najbliższego sąsiedztwa, pozostałe opcje na tej karcie dotyczące zgodności i szybkości będą wyłączone, ponieważ są istotne tylko dla przewidywania zmiennych przewidywanych.

Jaki jest cel?

Podczas przewidywania zmiennej przewidywanej ta grupa opcji umożliwia podjęcie decyzji na temat tego, czy szybkość, czy dokładność (czy oba te parametry) jest najważniejszym czynnikiem przy przewidywaniu zmiennej przewidywanej. Można również samodzielnie dostosować ustawienia.

Jeśli wybierzesz opcję Zrównoważenie, Szybkość lub Dokładność, wówczas algorytm wybierze wstępnie najbardziej odpowiednią kombinację ustawień dla tej opcji. Użytkownicy zaawansowani mogą zastąpić te opcje — można to zrobić w wielu panelach na karcie Ustawienia.

Zrównoważenie szybkości i dokładności. Umożliwia wybór najlepszej liczby sąsiadów w niewielkim przedziale.

Szybkość. Umożliwia znalezienie stałej liczby sąsiadów.

Dokładność. Umożliwia wybór najlepszej liczby sąsiadów w większym przedziale i używa ważności predyktora podczas obliczania odległości.

Analiza użytkownika. Wybierz tę opcję, aby precyzyjnie dostosować algorytm na karcie Ustawienia.

Uwaga: Rozmiar wynikowego modelu KNN — w przeciwieństwie do większości innych modeli — zwiększa się liniowo wraz z ilością danych uczących. Jeśli przy próbie utworzenia modelu KNN pojawi się błąd informujący „brak pamięci", spróbuj zwiększyć maksymalną ilość pamięci systemowej używanej przez produkt IBM SPSS Modeler. W tym celu wybierz opcje

Narzędzia > Opcje > Opcje systemowe

i wprowadź nowy rozmiar do pola **Maksimum pamięci**. Zmiany wprowadzone w oknie dialogowym Opcje systemowe obowiązują dopiero po restarcie produktu IBM SPSS Modeler.

Ustawienia węzła KNN

Karta Ustawienia to miejsce, w którym można określać opcje specyficzne dla analizy najbliższego sąsiedztwa. Pasek boczny po lewej stronie ekranu zawiera listę paneli, których można używać w celu określania opcji.

Model

Panel Model udostępnia opcje, które określają sposób budowania modelu — na przykład określają to, czy używane są modele podzielone na podzbiory, czy rozdzielone, czy możliwe jest przekształcanie liczbowych zmiennych wejściowych w taki sposób, aby należały do tego samego przedziału, a dodatkowo określają sposób zarządzania interesującymi obserwacjami. Dodatkowo dla modelu można wybrać niestandardową nazwę.

Uwaga: Opcje **Użyj danych podzielonych na podzbiory** i **Użyj opisu obserwacji** nie mogą dotyczyć tej samej zmiennej.

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Utwórz modele rozdzielone. Tworzy osobny model dla każdej możliwej wartości zmiennych wejściowych, jakie są określone jako zmienne podziału. Więcej informacji można znaleźć w "Budowanie modeli rozdzielonych" na stronie 28.

Aby wybrać zmienne ręcznie... Domyślnie węzeł korzysta z ustawień zmiennych dzielących na podzbiory i zmiennych podziału (o ile są dostępne) z węzła Typ, ale tutaj można zastąpić te ustawienia. Aby aktywować pola **Podział** i **Rozdzielone**, wybierz zakładkę **Zmienne**, a następnie wybierz opcję **Użyj ustawień niestandardowych** i wróć tutaj.

- **Podział.** To pole umożliwia określenie zmiennej używanej do podziału danych na osobne próby do uczenia, testowania i walidacji podczas budowania modelu. Korzystając z jednej próby do generowania modelu oraz innej do testowania go, można uzyskać wiarygodne informacje o tym, jak dobrze model pozwala uogólnić większe zbiory danych, podobne do danych bieżących. Jeśli korzystając z węzłów Typ lub Partycja, zdefiniowano wiele zmiennych dzielących na podzbiory, na karcie Zmienne każdego węzła modelowania korzystającego z tego podziału na podzbiory należy wybrać jedną zmienną dzielącą na podzbiory. (Jeśli obecna jest tylko jedna zmienna dzieląca na podzbiory, jest ona automatycznie używana po aktywowaniu dzielenia). Należy również pamiętać, że aby zastosować wybrany podział w analizie, dzielenie musi być również włączone na karcie Opcje modelu danego węzła. (Usunięcie zaznaczenia tej opcji umożliwia wyłączenie podziału bez zmiany ustawień zmiennych).

- **Rozdzielone.** W przypadku modeli rozdzielonych należy wybrać zmienne lub zmienną podziału. Działanie jest podobne, jak w przypadku ustawienia roli zmiennej na wartość *Rozdzielone* w węźle Typ. Jako zmienne podziału można wyznaczyć tylko zmienne typu **Flaga**, **Nominalna** lub **Porządkowa**. Zmienne wybrane jako zmienne podziału nie mogą być używane jako zmienne przewidywane, wejściowe, zmienne dzielące na podzbiory, zmienne częstości ani zmienne ważące. Więcej informacji można znaleźć w temacie “Budowanie modeli rozdzielonych” na stronie 28.

Normalizuj wejściowe zmienne ilościowe. Zaznacz to pole wyboru, aby normalizować wartości dla ciągłych zmiennych wejściowych. Znormalizowane funkcje mają ten sam zakres wartości, co może poprawić wydajność algorytmu estymacji. Używana jest normalizacja skorygowana [2*(x−min)/(max−min)]−1. Skorygowane wartości znormalizowane zawierają się w zakresie od -1 do 1.

Użyj opisu obserwacji. Zaznacz to pole wyboru, aby włączyć listę rozwijaną, z której można będzie wybrać zmienną, której wartości będą używane jako etykiety w celu identyfikacji interesujących obserwacji w wykresie przestrzennym predyktora, wykresie elementów równorzędnych, a także w mapie kwadratowej w przeglądarce modelu. Na zmienną etykiet można wybrać dowolną zmienną o poziomie pomiaru *Nominalna*, *Porządkowa* lub *Flaga*. Jeśli w tym przypadku nie została wybrana zmienna, rekordy będą wyświetlane w wykresach przeglądarki modelu, a ich najbliżsi sąsiedzi będą identyfikowani na podstawie numeru wiersza w danych źródłowych. Jeśli po zbudowaniu modelu dane będą manipulowane, użyj etykiet obserwacji, aby uniknąć konieczności ponownego odwoływania się do danych źródłowych każdorazowo w celu zidentyfikowania obserwacji w widoku.

Wskaż obserwację centralną. Zaznacz to pole, aby włączyć listę rozwijaną, która umożliwi zaznaczenie szczególnie interesującej zmiennej wejściowej (to dotyczy tylko zmiennych flagi). Jeśli na tej liście zostanie zaznaczona zmienna, punkty reprezentujące tę zmienną będą początkowo wybrane w przeglądarce modelu podczas budowania modelu. Wybranie obserwacji centralnej jest opcjonalne; dowolny punkt może stać się obserwacją centralną, gdy zostanie wybrany ręcznie w przeglądarce modelu.

Sąsiedzi

Panel Sąsiedzi zawiera zestaw opcji, które umożliwiają kontrolowanie sposobu obliczania liczby najbliższych sąsiadów.

Liczba najbliższych sąsiadów (k). Określ liczbę najbliższych sąsiadów dla konkretnej obserwacji. Należy pamiętać, że większa liczba obserwacji najbliższego sąsiedztwa nie zawsze oznacza dokładniejszy model.

Jeśli celem jest przewidzenie zmiennej przewidywanej, masz dwie opcje do wyboru:

- **Określ stałą k.** Użyj tej opcji, jeśli zamierzasz określić stałą liczbę najbliższych sąsiadów do znalezienia.
- **Automatycznie wybierz wartość k** Można również użyć zmiennych **Minimum** i **Maksimum**, aby określić przedział wartości i pozwolić procedurze na wybór „najlepszej” liczby sąsiadów w danym przedziale. Metoda określenia liczby obserwacji najbliższego sąsiedztwa zależy od tego, czy zaznaczono wybór predyktora na panelu Dobór predyktorów:

  Jeżeli wybór predyktora jest włączony, wówczas jest wykonywany dla każdej wartości *k* w żądanym przedziale i wybierane jest *k* oraz towarzyszący zestaw predyktorów z najniższym poziomem błędów (lub najniższą sumą kwadratów błędów, jeśli zmienna przewidywana jest ilościowa).

  Jeżeli wybór funkcji nie jest włączony, *V*-krotna walidacja krzyżowa jest używana w celu wybrania najlepszej liczby obserwacji najbliższego sąsiedztwa. Informacje na temat kontrolowania przypisań krotności zawiera panel Walidacja krzyżowa.

Obliczanie odległości. Jest to miara używana do określenia metryki odległości używanej do pomiaru podobieństwa obserwacji.

- **Metryka euklidesowa.** Odległość pomiędzy dwiema obserwacjami, x i y, jest pierwiastkiem kwadratowym sumy, we wszystkich wymiarach, różnic pomiędzy wartościami obserwacji podniesionymi do kwadratu.
- **Metryka miejska.** Odległość pomiędzy dwoma obserwacjami jest sumą, we wszystkich wymiarach, bezwzględnych różnic pomiędzy wartościami tych obserwacji. Miara ta jest również nazywana odległością manhattańską.

Opcjonalnie jeśli cel ma przewidzieć zmienną przewidywaną, wówczas można wybrać ważenie predyktorów według ich znormalizowanej ważności podczas obliczania odległości. Ważność predyktora jest obliczana jako współczynnik poziomu błędu lub błąd sumy kwadratów modelu z predyktorem usuniętym z modelu do poziomu błędu lub błędu sumy kwadratów pełnego modelu. Znormalizowana ważność jest obliczana przez zmianę wag wartości ważności właściwości, tak aby dawały w sumie 1.

Ważenie zmiennych według ważności przy obliczaniu odległości. (Wyświetlane tylko wówczas, gdy cel ma przewidzieć zmienną przewidywaną). Zaznacz to pole wyboru, aby spowodować użycie ważności predyktora podczas obliczania odległości między sąsiadami. Ważność predyktora będzie wyświetlana w modelu użytkowym i będzie używana w predykcjach (i z tego względu będzie wpływać na ocenianie). Więcej informacji można znaleźć w temacie “Ważność predyktorów” na stronie 43.

Predykcje dla zmiennej ilościowej. (Wyświetlane tylko wówczas, gdy cel ma przewidzieć zmienną przewidywaną). Jeśli określona jest zmienna przewidywana ilościowa (zakres wartości numerycznych), wówczas ta opcja określa, czy przewidywana wartość jest obliczana na podstawie zmiennej, czy na podstawie wartości mediany najbliższych sąsiadów.

Dobór predyktorów

Ten panel jest aktywowany tylko wówczas, gdy cel ma przewidzieć zmienną przewidywaną. Umożliwia żądanie i określanie opcji doboru predyktorów. Domyślnie wszystkie funkcje są uwzględniane przy wyborze funkcji, ale można wybrać zestaw funkcji, które zostaną wymuszone w modelu.

Dokonaj wyboru predyktora. Zaznacz to pole wyboru, aby włączyć opcje doboru predyktorów.

- **Wymuszone wprowadzanie.** Kliknij przycisk selektora zmiennych obok tego pola, a następnie wybierz jedną lub większą liczbę zmiennych, które zostaną wprowadzone do modelu w sposób wymuszony.

Kryterium zatrzymywania. Na każdym etapie predyktor, którego dodanie do modelu powoduje najmniejszy błąd (obliczony jako poziom błędu jakościowej zmiennej przewidywanej i błąd sumy kwadratów ilościowej zmiennej przewidywanej), jest uwzględniany w celu dołączenia do zestawu modelu. Selekcja postępująca jest kontynuowana do osiągnięcia określonego warunku.

- **Zatrzymaj po selekcji określonej liczby zmiennych.** Algorytm dodaje stałą liczbę funkcji oprócz tych, które są wymuszone w modelu. Określ dodatnią liczbę całkowitą. Zmniejszenie wartości liczby do wyboru tworzy skromniejszy model, z ryzykiem pominięcia istotnych funkcji. Zwiększenie wartości liczby do wyboru spowoduje ujęcie wszystkich istotnych funkcji, z ryzykiem dodania funkcji, które w rzeczywistości zwiększają błąd modelu.
- **Zatrzymaj, jeśli zmiana w ilorazie błędu bezwzględnego jest mniejsza lub równa minimum.** Algorytm zostaje zatrzymany, gdy zmiana bezwzględnego współczynnika błędu wskazuje, że model nie może być dalej udoskonalony przez dodanie dalszych funkcji. Określ liczbę dodatnią. Zmniejszenie wartości minimalnej zmiany spowoduje uwzględnienie większej liczby predyktorów przy ryzyku uwzględnienia predyktorów, które nie dodają dużej wartości do modelu. Zwiększenie wartości minimalnej zmiany spowoduje wyłączenie większej ilości funkcji przy ryzyku utraty funkcji, które są istotne w modelu. Optymalna wartość minimalnej zmiany zależy od danych i ich zastosowania. Dziennik błędów wyboru funkcji wyników pomaga ocenić, które funkcje są najbardziej istotne. Więcej informacji można znaleźć w temacie “Dziennik błędów wyboru predyktorów” na stronie 350.

Walidacja krzyżowa

Ten panel jest aktywowany tylko wówczas, gdy cel ma przewidzieć zmienną przewidywaną. Opcje dostępne na tym panelu kontrolują, czy podczas obliczania najbliższych sąsiadów stosowana jest walidacja krzyżowa.

Walidacja krzyżowa dzieli próbę na kilka podprób (**krotności**). Następnie generowane są modele najbliższego sąsiedztwa, wyłączając kolejno dane z każdej podpróby. Pierwszy model jest oparty na wszystkich obserwacjach z wyjątkiem tych w pierwszej krotności próby; drugi model jest oparty na wszystkich obserwacjach z wyjątkiem drugiej krotności próby itd. Dla każdego modelu szacowany jest błąd z zastosowaniem modelu na podpróbie wyłączonej podczas generowania modelu. „Najlepsza” liczba obserwacji najbliższego sąsiedztwa to ta, która powoduje najniższy błąd we wszystkich krotnościach.

Krotności walidacji krzyżowej. *V*-krotność walidacji krzyżowej jest używana do określenia „najlepszej” liczby obserwacji najbliższego sąsiedztwa. Nie jest dostępna w połączeniu z wyborem funkcji z powodów wydajności.

- **Losowo przydziel obserwacje do krotności.** Określ liczbę krotności, które powinny być użyte do walidacji krzyżowej. Procedura losowo przypisuje obserwacje do krotności, ponumerowanych od 1 do *V*, liczby krotności.
- **Ustaw wartość początkową generatora liczb losowych.** W przypadku szacowania dokładności modelu w oparciu o losową wartość procentową opcja ta pozwala na zduplikowanie tych samych wyników w innej sesji. Określenie wartości początkowej używanej przez generator liczb pseudolosowych zapewni, że podczas każdego wykonywania węzła przypisywane będą te same rekordy. Wprowadź żądaną wartość startową generatora. Jeśli nie wybrano tej opcji, przy każdej próbie wykonania węzła wygenerowana zostanie inna próba.
- **Użyj zmiennej przypisującej obserwacje do grup.** Określ zmienną numeryczną, która przydziela każdą obserwację w aktywnym zbiorze danych do krotności. Zmienna musi być numeryczna i przyjmować wartości od 1 do *V*. Jeżeli brakuje jakiejkolwiek wartości z tego przedziału, spowoduje to błąd, podobnie jak wszelkie zmienne dzielące w przypadku stosowania modeli rozdzielonych.

Analiza

Panel Analiza jest aktywowany tylko wówczas, gdy cel ma przewidzieć zmienną przewidywaną. Można go używać w celu określenia, czy model ma zawierać dodatkowe zmienne, które mogą być następujące:

- prawdopodobieństwa dla każdej możliwej wartość zmiennej przewidywanej
- odległości między obserwacją a jej najbliższymi sąsiadami
- surowe i skorygowane oceny skłonności (tylko dla zmiennych przewidywanych typu flaga)

Dołącz wszystkie prawdopodobieństwa. Jeśli ta opcja jest zaznaczona, prawdopodobieństwa poszczególnych możliwych wartości zmiennej nominalnej lub zmiennej przewidywanej typu flaga są wyświetlane dla każdego rekordu przetworzonego przez węzeł. Jeśli ta opcja nie jest zaznaczona, wówczas dla zmiennych nominalnych i przewidywanych typu flaga wyświetlana będzie tylko wartość przewidywana i jej prawdopodobieństwo.

Zapisz odległości pomiędzy obserwacjami i k najbliższymi sąsiadami. Dla każdej obserwacji centralnej tworzona jest osobna zmienna dla każdego z *k* najbliższych sąsiadów tej obserwacji (z próby uczącej) oraz odpowiadających im *k* najbliższych odległości.

Oceny skłonności

Oceny skłonności można aktywować w węźle modelowania oraz na karcie Ustawienia w modelu użytkowym. Ta funkcja jest dostępna tylko wówczas, gdy wybrana zmienna przewidywana jest zmienną typu flaga. Więcej informacji można znaleźć w temacie “Oceny skłonności” na stronie 35.

Wylicz surowe oceny skłonności. Surowe oceny skłonności są wyznaczane z modelu wyłącznie w oparciu o dane uczące. Jeśli model przewiduje wartość *true* (udzieli odpowiedzi), wówczas skłonność jest taka sama jak P, gdzie P to prawdopodobieństwo predykcji. Jeśli model przewidzi wartość typu false, wówczas skłonność jest obliczana jako (1 – P).

- W przypadku wybrania tej opcji podczas budowania modelu oceny skłonności będą domyślnie aktywowane w modelu użytkowym. Surowe oceny skłonności można jednak aktywować w modelu użytkowym w dowolnym czasie, niezależnie od tego, czy zostały wybrane w węźle modelowania.
- Podczas oceniania modelu surowe oceny skłonności zostaną dodane do zmiennej z literami *RP* dodanymi do standardowego przedrostka. Przykładowo, jeśli predykcje znajdują się w zmiennej o nazwie *$R-churn*, wprowadzona nazwa zmiennej oceny skłonności będzie następująca: *$RRP-churn*.

Wylicz skorygowane oceny skłonności. Surowe skłonności są wyznaczane wyłącznie w oparciu o oszacowania udostępnione przez model, które mogą być nadmiernie dopasowane, co może doprowadzić do zbyt optymistycznych oszacowań skłonności. Skorygowane skłonności spróbują przeprowadzić wyrównanie, sprawdzając, jak model działa w podzbiorze testowym lub walidacyjnym i korygując skłonności, tak aby uzyskać lepsze oszacowanie.

- To ustawienie wymaga, aby w strumieniu obecna była poprawna zmienna dzieląca na podzbiory.
- W przeciwieństwie do surowych ocen ufności skorygowane oceny skłonności muszą być obliczone podczas budowania modelu; w przeciwnym razie nie będą dostępne podczas oceniania modelu użytkowego.

- Podczas oceniania modelu skorygowane oceny skłonności zostaną dodane do zmiennej z literami *AP* dodanymi do standardowego przedrostka. Przykładowo, jeśli predykcje znajdują się w zmiennej o nazwie *$R-churn*, wprowadzona nazwa zmiennej oceny skłonności będzie następująca: *$RAP-churn*. Skorygowane oceny skłonności są niedostępne dla modeli regresji logistycznej.
- Podczas obliczania skorygowanych ocen skłonności podzbiór testowy lub walidacyjny używany do obliczeń nie może być zrównoważony. Aby tego uniknąć, należy sprawdzić, czy opcja **Równoważ tylko dane uczące** jest zaznaczona w którymkolwiek poprzedzającym węźle ważenia. Ponadto, jeśli w poprzedzającej części strumienia przeprowadzona została złożona próba, spowoduje to unieważnienie skorygowanych ocen skłonności.
- Skorygowane oceny skłonności są niedostępne w przypadku modeli drzewa wzmacnianego i zestawu reguł. Więcej informacji można znaleźć w temacie “Wzmacniane modele C5.0” na stronie 124.

Model użytkowy KNN

Model KNN tworzy konkretną liczbę nowych zmiennych, co przedstawia poniższa tabela. W celu wyświetlenia tych zmiennych i ich wartości należy dodać węzeł Tabela do modelu użytkowego KNN i wykonać węzeł Tabela lub kliknąć przycisk Podgląd w modelu użytkowym.

Tabela 32. Zmienne modelu KNN

Nazwa nowej zmiennej	Opis
$KNN-*fieldname*	Predykowana wartość zmiennej przewidywanej.
$KNNP-*fieldname*	Prawdopodobieństwo wartości przewidywanej.
$KNNP-*value*	Prawdopodobieństwa wystąpienia każdej możliwej wartości zmiennej nominalnej lub przewidywanej typu flaga. Uwzględniana tylko wówczas, gdy pole wyboru **Dołącz wszystkie prawdopodobieństwa** jest zaznaczone na karcie Ustawienia modelu użytkowego.
$KNN-neighbor-*n*	Nazwa *n*-tego najbliższego sąsiada obserwacji centralnej. Uwzględniana tylko wówczas, gdy opcja **Wyświetlaj najbliższe** na karcie Ustawienia modelu użytkowego jest ustawiona na wartość różną od zera.
$KNN-distance-*n*	Odległość względna od obserwacji centralnej *n*-tego najbliższego sąsiada do obserwacji centralnej. Uwzględniana tylko wówczas, gdy opcja **Wyświetlaj najbliższe** na karcie Ustawienia modelu użytkowego jest ustawiona na wartość różną od zera.

Widok modelu najbliższego sąsiedztwa

Widok modelu

Widok modelu zawiera okno z 2 panelami:

- Pierwszy panel wyświetla przegląd modelu nazywany widokiem głównym.
- Drugi panel wyświetla jeden z dwóch rodzajów widoków:

  Pomocniczy widok modelu przedstawia więcej informacji o modelu, ale nie koncentruje się na samym modelu.

  Połączony widok jest widokiem przedstawiającym szczegółowe informacje o modelu, gdy użytkownik rozwinie część widoku głównego.

Domyślnie pierwszy panel przedstawia przestrzeń predyktorów, a drugi – wykres ważności predyktorów. Jeśli wykres ważności predyktorów jest niedostępny; to znaczy, że nie wybrano opcji **Ważenie zmiennych według ważności** na karcie w panelu Sąsiedzi karty Ustawienia, wówczas przedstawiany jest pierwszy dostępny widok z listy rozwijanej Widok.

Jeśli w konkretnym widoku nie są dostępne żadne informacje, jest on usuwany z listy rozwijanej Widok.

Przestrzeń predyktorów: Wykres przestrzeni predyktorów jest interaktywną grafiką przestrzeni predyktorów (lub podprzestrzenią, jeśli istnieją więcej niż 3 predyktory). Każda oś reprezentuje predyktor w modelu, a lokalizacja punktów na wykresie przestawia wartości tych predyktorów dla obserwacji w podzbiorze uczącym i testującym.

Klucze. Dodatkowo obok wartości predyktorów punkty wykresu prezentują również inne informacje.

- Kształt określa podział, do którego należy punkt; jest to podział szkoleniowy lub wstrzymany.
- Kolor/deseń punktu oznacza wartość wartości docelowej tej obserwacji; wartości o wyraźnych kolorach są równe kategoriom jakościowych wartości docelowych, a cienie oznaczają zakres wartości docelowych wartości ilościowych. Wskazana wartość podziału szkoleniowego jest wartością obserwowaną; w przypadku podziału wstrzymanego jest to wartość przewidywana. Jeżeli nie określono wartości docelowej, ten klucz nie jest wyświetlany.
- Grubszy obrys oznacza, że obserwacja jest centralna. Obserwacje centralne są połączone z *k* najbliższymi sąsiadami.

Elementy sterujące i interaktywność. Kilka elementów sterujących na wykresie umożliwia eksplorację przestrzeni predyktorów.

- Możliwy jest wybór zestawu predyktorów wyświetlanych na wykresie i zmiana predyktorów reprezentowanych w wymiarach.
- „Obserwacje centralne" są punktami wybranymi na wykresie przestrzeni predyktorów. Jeżeli zostanie określona zmienna obserwacji centralnej, punkty reprezentujące obserwacje centralne zostaną wstępnie zaznaczone. Każdy punkt może jednak zostać obserwacją centralną, jeśli zostanie wybrany przez użytkownika. Obowiązują „standardowe" elementy sterujące wyboru punktu; kliknięcie punktu powoduje zaznaczenie tego punktu i odznaczenie innych punktów; kliknięcie punktu z naciśniętym klawiszem Ctrl dodaje punkt do grupy wybranych punktów. Połączone widoki, takie jak np. Wykres elementów równorzędnych, zostaną automatycznie aktualizowane na podstawie obserwacji wybranych w przestrzeni predyktorów.
- Możliwa jest zmiana liczby najbliższych sąsiadów (*k*) wyświetlanych dla obserwacji centralnych.
- Najechanie kursorem nad punkt na wykresie powoduje wyświetlenie informacji z wartością opisu obserwacji lub numerem obserwacji, jeśli opisy obserwacji nie są zdefiniowane, oraz z obserwowanymi i przewidywanymi wartościami docelowymi.
- Przycisk „Resetuj" umożliwia przywrócenie przestrzeni predyktorów do oryginalnego stanu.

Zmiana osi na wykresie przestrzeni predyktorów: Użytkownik może kontrolować, które predyktory są wyświetlane na osiach wykresu przestrzeni predyktorów.

Aby zmienić ustawienia osi:

1. Kliknij przycisk Tryb edycji (ikona pędzla) w panelu po lewej stronie, aby wybrać tryb edycji dla przestrzeni predyktorów.
2. Zmień widok (na dowolny) w panelu po prawej stronie. Między dwoma głównymi panelami pojawi się panel **Pokaż strefy**.
3. Kliknij pole wyboru **Pokaż strefy**.
4. Kliknij dowolny punkt danych w przestrzeni predyktorów.
5. Aby zastąpić oś predyktorem tego samego typu danych:
   - Przeciągnij nowy predyktor nad etykietę strefy (z małym przyciskiem X) predyktora wybranego do zastąpienia.
6. Aby zastąpić oś predyktorem innego typu danych:
   - Na etykiecie strefy predyktora, którego chcesz zastąpić, kliknij mały przycisk X. Przestrzeń predyktorów zmieni się w widok dwuwymiarowy.
   - Przeciągnij nowy predyktor nad etykietę strefy **Dodaj wymiar**.
7. Kliknij przycisk Tryb eksploracji (ikona grotu strzałki) w panelu po lewej stronie, aby zamknąć tryb edycji.

Ważność predyktorów: Zazwyczaj działania modelujące mają koncentrować się na zmiennych predyktorów, które są najważniejsze, a opuszczane lub ignorowane mają być te zmienne, które są najmniej ważne. Wykres ważności predyktorów pomaga osiągnąć ten cel przez wskazanie względnej ważności każdego predyktora przy szacowaniu modelu. Ponieważ wartości są względne, suma wartości wszystkich wyświetlanych predyktorów wynosi 1,0. Ważność predyktora nie jest powiązana z dokładnością modelu. Jest powiązana z ważnością każdego predyktora przy prognozach, a nie z tym, czy taka prognoza jest dokładna.

Odległości najbliższego sąsiedztwa: Tabela przestawia *k* obserwacji najbliższego sąsiedztwa oraz odległości wyłącznie do obserwacji centralnych. Jest dostępna tylko wówczas, gdy w węźle modelowania na karcie określono identyfikator obserwacji kluczowej, i przedstawia tylko obserwacje centralne wskazywane przez tę zmienną.

Każdy wiersz:

- Kolumna **Obserwacja centralna** zawiera wartość zmiennej opisu obserwacji dla obserwacji centralnej. Jeżeli opis obserwacji nie jest zdefiniowany, kolumna zawiera numer obserwacji centralnej.
- *i*-ta kolumna grupy **Najbliższe sąsiedztwo** zawiera wartość zmiennej opisu obserwacji dla *i*-tej obserwacji najbliższego sąsiedztwa obserwacji centralnej; jeśli opis obserwacji nie jest zdefiniowany, kolumna zawiera numer *i*-tej obserwacji najbliższego sąsiedztwa obserwacji centralnej.
- *i*-ta kolumna grupy **Najbliższe sąsiedztwo** zawiera odległość *i*-tej obserwacji najbliższego sąsiedztwa do obserwacji centralnej

Elementy o zbliżonych wartościach: Wykres przedstawia obserwacje centralne oraz *k* najbliższych sąsiadów dla każdego predyktora i zmiennej przewidywanej. Wykres jest dostępny, jeśli w przestrzeni predyktorów zaznaczono obserwację centralną.

Wykres elementów równorzędnych jest połączony z przestrzenią predyktorów na dwa sposoby.

- Obserwacje wybrane (centralne) w przestrzeni predyktorów są wyświetlane na wykresie elementów równorzędnych razem z *k* najbliższymi sąsiadami.
- Wartość *k* wybrana w przestrzeni predyktorów jest używana na wykresie elementów równorzędnych.

Wybór predyktorów. Umożliwia wybranie predyktorów do wyświetlenia na wykresie elementów równorzędnych.

Mapa kwadratowa: Wykres przedstawia obserwacje kluczowe oraz *k* najbliższych sąsiadów na wykresie rozrzutu (lub wykresie punktowym, w zależności od poziomu pomiaru wartości zmiennej przewidywanej), ze zmienną przewidywaną na osi *y* i predyktorem skalowym na osi *x*, ograniczone predyktorami. Wykres jest dostępny, jeśli istnieje zmienna przewidywana i w przestrzeni predyktorów zaznaczono obserwację centralną.

- Linie referencyjne są rysowane dla zmiennych ilościowych, przy średnich zmiennej w podziale szkoleniowym.

Wybór predyktorów. Umożliwia wybranie predyktorów do wyświetlenia w mapie kwadratowej.

Dziennik błędów wyboru predyktorów: Punkty na wykresie przedstawiają błąd (współczynnik poziomu błędu lub błąd sumy kwadratów, w zależności od poziomu pomiaru zmiennej przewidywanej) na osi *y* dla modelu z predyktorem przedstawionym na osi *x* (plus wszystkie zmienne po lewej na osi *x*). Wykres jest dostępny, jeśli istnieje wartość docelowa i działa wybór funkcji.

Tabela klasyfikacji: Ta tabela przedstawia klasyfikację krzyżową wartości obserwowanych i przewidywanych wartości docelowej, według podziału. Ta tabela jest dostępna, jeśli istnieje zmienna przewidywana i jest ona jakościowa (flaga, nominalna lub porządkowa).

- Wiersz **(Brakujące)** w podziale wstrzymanym zawiera obserwacje wstrzymane z brakującymi wartościami w wartości docelowej. Obserwacje te mają wpływ na próbę wstrzymaną: wartości Ogólnie procent, ale nie wartości Procent poprawny.

Podsumowanie błędów: Tabela jest dostępna, jeśli istnieje zmienna docelowa. W tabeli jest wyświetlany błąd powiązany z modelem; suma kwadratów dla docelowych wartości ilościowych oraz poziom błędu (100% - ogólnie procent poprawnie) dla docelowych wartości jakościowych.

Ustawienia modelu KNN

Karta Ustawienia umożliwia określenie dodatkowych zmiennych, jakie będą wyświetlane podczas przeglądania listy wyników (np. w wyniku wykonania węzła Tabela dołączonego do modelu użytkowego). Można zobaczyć efekt działania poszczególnych opcji; w tym celu należy je zaznaczyć i kliknąć przycisk Podgląd — aby zobaczyć dodatkowe zmienne należy przewinąć w prawo wyniki podglądu.

Dołącz wszystkie prawdopodobieństwa (tylko dla przewidywanych zmiennych jakościowych). Jeśli ta opcja jest zaznaczona, prawdopodobieństwa poszczególnych możliwych wartości zmiennej nominalnej lub zmiennej przewidywanej typu flaga są wyświetlane dla każdego rekordu przetworzonego przez węzeł. Jeśli ta opcja nie jest zaznaczona, wówczas dla zmiennych nominalnych i przewidywanych typu flaga wyświetlana będzie tylko wartość przewidywana i jej prawdopodobieństwo.

Ustawienie domyślne dla tego pola wyboru jest określane przez odpowiednie pole w modelu użytkowym.

Wylicz surowe oceny skłonności. W przypadku modeli z przewidywaną zmienną typu flaga (zwracających predykcje tak lub nie) można wyliczyć oceny skłonności definiujące prawdopodobieństwo prawdziwego wyniku określonego dla zmiennej przewidywanej. Stanowią one uzupełnienie pozostałych współczynników ufności i wartości predykcyjnych, jakie mogą zostać wygenerowane podczas oceniania.

Wylicz skorygowane oceny skłonności. Surowe oceny skłonności bazują na danych uczących i mogą być zbyt optymistyczne z uwagi na tendencję wielu modeli do przeuczania tych danych. Opcja wyliczania skłonności skorygowanych próbuje skompensować tę tendencję na drodze oceny wydajności modelu w teście lub w podzbiorze walidacyjnym. Opcja ta wymaga, aby zmienna dzieląca na podzbiory była zdefiniowana w strumieniu oraz aby skorygowane oceny skłonności były włączone w węźle modelowania przed przystąpieniem do generowania modelu.

Wyświetlaj najbliższe. Jeśli ta wartość zostanie ustawiona na *n*, gdzie *n* jest niezerową dodatnią liczbą całkowitą, wówczas w modelu zostanie uwzględnionych *n* najbliższych sąsiadów obserwacji centralnej, razem z ich względnymi odległościami od obserwacji centralnej.

Rozdział 17. Węzły Python

SPSS Modeler oferuje węzły umożliwiające korzystanie z algorytmów zapisanych bezpośrednio w języku Python. Karta **Python** na paleta węzłów zawiera następujące węzły umożliwiające uruchamianie algorytmów w języku Python. Węzły te są obsługiwane na platformach Windows 64, Linux64 i Mac.

Węzeł SMOTE (Synthetic Minority Over-sampling Technique — generowanie próbek syntetycznych z klasy mniejszościowej) realizuje algorytm nadpróbkowania przydatny w pracy z niezrównoważonymi zbiorami danych. Udostępnia on zaawansowaną metodę równoważenia danych. Węzeł procesowy SMOTE w programie SPSS Modeler jest zaimplementowany w języku Python i wymaga biblioteki Python imbalanced-learn©.

XGBoost Linear© to zaawansowana implementacja algorytmu wzmacniania gradientowego, który jako model bazowy wykorzystuje model liniowy. Algorytmy wzmacniania iteracyjnie ucząc się, wyznaczają słabe klasyfikatory i dodają je do ostatecznego silnego klasyfikatora. Węzeł Liniowy XGBoost w programie SPSS Modeler jest zaimplementowany w języku Python.

XGBoost Tree© to zaawansowana implementacja algorytmu wzmacniania gradientowego, który jako model bazowy wykorzystuje model drzewa. Algorytmy wzmacniania iteracyjnie ucząc się, wyznaczają słabe klasyfikatory i dodają je do ostatecznego silnego klasyfikatora. XGBoost Tree jest algorytmem bardzo elastycznym i oferuje liczne parametry, które mogą być trudne do praktycznego wykorzystania przez użytkowników. Dlatego węzeł Drzewo XGBoost w programie SPSS Modeler eksponuje tylko funkcje podstawowe i najczęściej używane parametry. Węzeł jest zaimplementowany w języku Python.

Stochastyczne włączanie sąsiadów o rozkładzie t (t-SNE — t-Distributed Stochastic Neighbor Embedding) to narzędzie do wizualizacji danych wysokowymiarowych. Przekształca ono powinowactwa punktów danych w prawdopodobieństwa. Węzeł t-SNE w programie SPSS Modeler został zaimplementowany w języku Python i wymaga biblioteki Python scikit-learn©.

Model mieszanin rozkładów Gaussa — Gaussian Mixture© — to model probabilistyczny, w którym zakłada się, że punkty danych generowane są na podstawie mieszaniny skończonej liczby rozkładów Gaussa o nieznanych parametrach. Modele mieszanin można opisać jako uogólnienie grupowania metodą k-średnich z uwzględnieniem informacji o strukturze kowariancji danych oraz środkach ukrytych rozkładów Gaussa. Węzeł Mieszanina rozkładów Gaussa w produkcie SPSS Modeler eksponuje podstawowe funkcje i często używane parametry biblioteki Gaussian Mixtore. Węzeł jest zaimplementowany w języku Python.

Jądrowy estymator gęstości — Kernel Density Estimation (KDE)© — używa algorytmów Ball Tree lub KD Tree do efektywnej obsługi zapytań i integruje techniki uczenia nienadzorowanego, generowania cech (feature enginnering) i modelowania danych. Do najpopularniejszych i najbardziej użytecznych technik estymacji gęstości należą metody oparte na analizie sąsiedztwa, takie jak KDE. Węzły Modelowanie KDE i Symulacja KDE w produkcie SPSS Modeler eksponują podstawowe funkcje i często używane parametry biblioteki KDE. Węzły są zaimplementowane w języku Python.

Węzeł Las losowy korzysta z zaawansowanej implementacji algorytmu agregacji (bagging), która jako model bazowy wykorzystuje model drzewa. Węzeł modelowania Las losowy w programie SPSS Modeler został zaimplementowany w języku Python i wymaga biblioteki Python scikit-learn©.

Węzeł Hierarchical Density-Based Spatial Clustering (HDBSCAN)© korzysta z algorytmu uczenia nienadzorowanego, aby wyszukać skupienia lub regiony o dużej gęstości w zbiorze danych. Węzeł HDBSCAN w programie SPSS Modeler eksponuje tylko funkcje podstawowe i najczęściej używane parametry biblioteki HDBSCAN. Węzeł jest zaimplementowany w języku Python i można go użyć do skupiania zbioru danych w osobne grupy, jeśli nie wiemy z góry, co to są za grupy.

Węzeł SVM z jedną klasą korzysta z algorytmu uczenia nienadzorowanego. Węzeł ten można wykorzystać do wykrywania nowości. Wykryje on miękką granicę danego zbioru próbek, a następnie sklasyfikuje nowe punkty jako należące do tego zbioru albo do niego nienależące. Węzeł modelowania SVM z jedną klasą w programie SPSS Modeler został zaimplementowany w języku Python i wymaga biblioteki Python scikit-learn©.

Węzeł SMOTE

Węzeł SMOTE (Synthetic Minority Over-sampling Technique — generowanie próbek syntetycznych z klasy mniejszościowej) realizuje algorytm nadpróbkowania przydatny w pracy z niezrównoważonymi zbiorami danych. Udostępnia on zaawansowaną metodę równoważenia danych. Węzeł procesowy SMOTE jest zaimplementowany w języku Python i wymaga biblioteki Python imbalanced-learn©. Szczegółowe informacje o bibliotece imbalanced-learn można znaleźć na stronie http://contrib.scikit-learn.org/imbalanced-learn/about.html[1].

Karta Python na palecie węzłów zawiera węzeł SMOTE i inne węzły Python.

[1]Lemaître, Nogueira, Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research*, vol. 18, no. 17, 2017, pp. 1-5. (http://jmlr.org/papers/v18/16-365.html)

Ustawienia węzła SMOTE

Na karcie **Ustawienia** węzła SMOTE zdefiniuj następujące ustawienia.

Ustawienia zmiennej przewidywanej

Zmienna przewidywana. Wybierz zmienną przewidywaną. Obsługiwane są wszystkie typy pomiaru: flaga, nominalne, porządkowe i dyskretne. Jeżeli w sekcji Podział zaznaczona jest opcja **Użyj danych podzielonych na podzbiory**, to nadpróbkowane będą tylko dane uczące.

Współczynnik próbek syntetycznych

Wybierz opcję **Automatycznie**, aby automatycznie wybrać współczynnik próbek syntetycznych, albo wybierz opcję **Ustaw współczynnik (mniejszość do większości)**, aby określić niestandardowy współczynnik. Współczynnik to liczba próbek w klasie mniejszościowej podzielona przez liczbę próbek w klasie większościowej. Wartość współczynnika musi być większa od **0** i mniejsza lub równa **1**.

Losowa wartość początkowa

Ustaw wartość początkową generatora liczb losowych. Wybierz tę opcję i kliknij przycisk **Utwórz**, aby wygenerować wartość początkową dla generatora liczb losowych.

Metody

Rodzaj algorytmu. Wybierz typ algorytmu SMOTE, który ma być używany.

Reguły próbek

K-sąsiedzi. Określ liczbę najbliższych sąsiadów, która ma być używana do tworzenia próbek syntetycznych.

M-sąsiedzi. Określ liczbę najbliższych sąsiadów, która ma być używana do określenia, czy próbka mniejszościowa jest zagrożona. Ten parametr jest używany tylko w przypadku wybrania typu algorytmu SMOTE **Borderline1** lub **Borderline2**.

Podział

Użyj danych podzielonych na podzbiory. Wybierz tę opcję, jeśli tylko dane uczące mają być nadpróbkowane.

Węzeł SMOTE wymaga biblioteki Python imbalanced-learn©. W poniższej tabeli przedstawiono relacje między ustawieniami w oknie dialogowym węzła SMOTE w programie SPSS Modeler a parametrami algorytmu w języku Python.

Tabela 33. Właściwości węzła odwzorowane na parametry biblioteki Python

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Nazwa parametru w interfejsie API środowiska Python
Współczynnik próbek syntetycznych (pole do wprowadzania liczby)	sample_ratio_value	ratio
Wartość początkowa	random_seed	random_state
K-sąsiedzi	k_neighbours	k
M-sąsiedzi	m_neighbours	m
Rodzaj algorytmu	algorithm_kind	kind

Węzeł Liniowy XGBoost

XGBoost Linear© to zaawansowana implementacja algorytmu wzmacniania gradientowego, który jako model bazowy wykorzystuje model liniowy. Algorytmy wzmacniania iteracyjnie ucząc się, wyznaczają słabe klasyfikatory i dodają je do ostatecznego silnego klasyfikatora. Węzeł Liniowy XGBoost w programie SPSS Modeler jest zaimplementowany w języku Python.

Więcej informacji o algorytmach wzmacniania zawierają kursy XGBoost dostępne na stronie http://xgboost.readthedocs.io/en/latest/tutorials/index.html. [1]

Uwaga: w programie SPSS Modeler nie jest obsługiwana funkcja walidacji krzyżowej XBoost. Ten sam cel można zrealizować za pomocą węzła Podział w programie SPSS Modeler. Ponadto algorytm XGBoost w SPSS Modeler automatycznie wykonuje kodowanie one-hot (kodowanie z gorącą jedynką) dla zmiennych kategorialnych.

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. WWW. © 2015-2016 DMLC.

Zmienne węzła Liniowy XGBoost

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Wybierz tę opcję, aby ręcznie przypisać zmienną przewidywaną.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli. Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji.

Predyktory. Można wybrać jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Opcje budowania węzła Liniowy XGBoost

Karta Opcje budowania umożliwia określenie opcji tworzenia węzła Liniowy XGBoost, w tym **opcji podstawowych**, takich jak parametry wzmacniania liniowego i budowania modelu, a także **opcje zadania uczenia** dotyczące celów. Więcej informacji na temat tych opcji można znaleźć w następujących źródłach internetowych:

- Skorowidz parametrów XGBoost[1]
- Interfejs API XGBoost w środowisku Python[2]
- Strona główna XGBoost[3]

Podstawowe

Optymalizacja hiperparametrów (wg Rbfopt). Wybranie tej opcji włącza optymalizację hiperparametrów opartą na Rbfopt, która automatycznie wykrywa optymalną kombinację parametrów, przy której model osiągnie oczekiwany lub niższy od oczekiwanego wskaźnika błędów dla prób. Aby uzyskać szczegółowe informacje na temat biblioteki Rbfopt, patrz http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html.

Alfa. Składnik regularyzacji L1 wag. Zwiększenie tej wartości powoduje, że model jest bardziej konserwatywny.

Lambda. Składnik regularyzacji L2 wag. Zwiększenie tej wartości powoduje, że model jest bardziej konserwatywny.

Obciążenie Lambda. Składnik regularyzacji L2 obciążenia. (Nie ma składnika regularyzacji L1 obciążenia, ponieważ jest nieistotny).

Liczba rund wzmocnienia. Liczba iteracji wzmacniania.

Zadanie uczenia

Cel. Wybierz jeden z następujących typów celu zadania: **reg:linear**, **reg:logistic**, **reg:gamma**, **reg:tweedie**, **count:poisson**, **rank:pairwise**, **binary:logistic**, **multi**.

Losowa wartość początkowa. Można kliknąć przycisk **Utwórz**, aby wygenerować wartość początkową dla generatora liczb losowych.

W następującej tabeli przedstawiono relacje między ustawieniami w oknie dialogowym węzła Liniowy XGBoost w programie SPSS Modeler a parametrami biblioteki Python XGBoost.

Tabela 34. Właściwości węzła odwzorowane na parametry biblioteki Python

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr XGBoost
Zmienna przewidywana	TargetField	
Predyktory	InputFields	
Lambda	lambda	lambda
Alfa	alpha	alpha
Obciążenie Lambda	lambdaBias	lambda_bias

Tabela 34. Właściwości węzła odwzorowane na parametry biblioteki Python (kontynuacja)

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr XGBoost
Liczba rund wzmocnienia	numBoostRound	num_boost_round
Cel	objectiveType	objective
Losowa wartość początkowa	random_seed	seed

[1] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. WWW. © 2015-2016 DMLC.

[2] "Plotting API" *Scalable and Flexible Gradient Boosting*. WWW. © 2015-2016 DMLC.

[3] "Scalable and Flexible Gradient Boosting." WWW. © 2015-2016 DMLC.

Opcje modelu węzła Liniowy XGBoost

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Węzeł Drzewo XGBoost

XGBoost Tree© to zaawansowana implementacja algorytmu wzmacniania gradientowego, który jako model bazowy wykorzystuje model drzewa. Algorytmy wzmacniania iteracyjnie ucząc się, wyznaczają słabe klasyfikatory i dodają je do ostatecznego silnego klasyfikatora. XGBoost Tree jest algorytmem bardzo elastycznym i oferuje liczne parametry, które mogą być trudne do praktycznego wykorzystania przez użytkowników. Dlatego węzeł Drzewo XGBoost w programie SPSS Modeler eksponuje tylko funkcje podstawowe i najczęściej używane parametry. Węzeł jest zaimplementowany w języku Python.

Więcej informacji o algorytmach wzmacniania zawierają kursy XGBoost dostępne na stronie http://xgboost.readthedocs.io/en/latest/tutorials/index.html. [1]

Uwaga: w programie SPSS Modeler nie jest obsługiwana funkcja walidacji krzyżowej XBoost. Ten sam cel można zrealizować za pomocą węzła Podział w programie SPSS Modeler. Ponadto algorytm XGBoost w SPSS Modeler automatycznie wykonuje kodowanie one-hot (kodowanie z gorącą jedynką) dla zmiennych kategorialnych.

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. WWW. © 2015-2016 DMLC.

Zmienne węzła Drzewo XGBoost

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Wybierz tę opcję, aby ręcznie przypisać zmienną przewidywaną.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli. Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji.

Predyktory. Można wybrać jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Opcje budowania węzła Drzewo XGBoost

Na karcie Opcje budowania można określić opcje tworzenia dla węzła Drzewo XGBoost, w tym **opcje podstawowe** dotyczące tworzenia modelu i wzrostu drzewa, **opcje zadania uczenia** dotyczące celów i **opcje zaawansowane** do zapobiegania przeuczeniu i obsługi niezrównoważonych zbiorów danych. Więcej informacji na temat tych opcji można znaleźć w następujących źródłach internetowych:

- Skorowidz parametrów XGBoost[1]
- Interfejs API XGBoost w środowisku Python[2]
- Strona główna XGBoost[3]

Podstawowe

Optymalizacja hiperparametrów (wg Rbfopt). Wybranie tej opcji włącza optymalizację hiperparametrów opartą na Rbfopt, która automatycznie wykrywa optymalną kombinację parametrów, przy której model osiągnie oczekiwany lub niższy od oczekiwanego wskaźnika błędów dla prób. Aby uzyskać szczegółowe informacje na temat biblioteki Rbfopt, patrz http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html.

Metoda drzewa. Wybierz algorytm tworzenia drzewa XGBoost.

Liczba rund wzmacniania. Określ liczbę iteracji wzmacniania.

Maks. głębokość. Określ maksymalną głębokość drzew. Zwiększenie tej wartości zwiększa złożoność modelu i ryzyko przeuczenia.

Min. waga elementu podrzędnego. Określ minimalną sumę wagi wystąpień (Hessiana) wymaganą w elemencie podrzędnym. Gdy krok podziału drzewa doprowadzi do powstania węzła-liścia z sumą wag wystąpień mniejszą od wartości **Min. waga elementu podrzędnego**, podczas tworzenia drzewa nie będą już wprowadzane dalsze podziały. W trybie regresji liniowej wartość ta odpowiada po prostu minimalnej liczbie wystąpień wymaganej w każdym węźle. Im większa waga, tym bardziej konserwatywnie działa algorytm.

Maks. krok zmiany. Określ maksymalny krok zmiany umożliwiający oszacowanie wag drzewa. Wartość **0** oznacza brak ograniczenia. Wartość dodatnia powoduje, że krok aktualizacji może być realizowany bardziej konserwatywnie. Zwykle ten parametr nie jest potrzebny, ale może pomóc w przypadku regresji logistycznej, gdy klasa jest skrajnie niezrównoważona.

Zadanie uczenia

Cel. Wybierz jeden z następujących typów celu zadania: **reg:linear**, **reg:logistic**, **reg:gamma**, **reg:tweedie**, **count:poisson**, **rank:pairwise**, **binary:logistic**, **multi**.

Przedwczesne zatrzymanie. Wybierz tę opcję, aby skorzystać z funkcji przedwczesnego zatrzymywania. Uczenie jest kontynuowane tylko wtedy, gdy przez liczbę **rund zatrzymania** błędy walidacji muszą zmniejszać się co najmniej w każdej rundzie. **Współczynnik danych ewaluacji** to współczynnik danych wejściowych używanych dla błędów walidacji.

Losowa wartość początkowa. Można kliknąć przycisk **Utwórz**, aby wygenerować wartość początkową dla generatora liczb losowych.

Zaawansowane

Podpróba. Podpróba określa współczynnik wystąpień używanych do uczenia. Na przykład przy wartości **0.5** algorytm XGBoost losowo wybierze połowę wystąpień danych do wzrostu drzewa. Takie ograniczenie zapobiega przeuczeniu.

Eta. Redukcja wielkości kroku używana podczas aktualizacji w celu zapobiegania przeuczeniu. Po każdym kroku wzmacniania wagi nowych predyktorów mogą być uzyskane bezpośrednio. Eta redukuje także wagi predyktorów, aby proces wzmacniania działał bardziej konserwatywnie.

Gamma. Minimalna redukcja straty wymagana do dalszego podziału węzła-liścia w drzewie. Im większa wartość gamma, tym bardziej konserwatywnie działa algorytm.

Próba z kolumn na każde drzewo. Współczynnik podpróbkowania kolumn podczas tworzenia każdego drzewa.

Próba z kolumn na każdy poziom. Współczynnik podpróbkowania kolumn dla każdego podziału na każdym poziomie.

Lambda. Składnik regularyzacji L2 wag. Zwiększenie tej wartości powoduje, że model jest bardziej konserwatywny.

Alfa. Składnik regularyzacji L1 wag. Zwiększenie tej wartości powoduje, że model jest bardziej konserwatywny.

Skaluj wagi dodatnie. Steruje równoważeniem wag dodatnich i ujemnych. Parametr ten jest przydatny w przypadku klas niezrównoważonych.

W następującej tabeli przedstawiono relacje między ustawieniami w oknie dialogowym węzła Drzewo XGBoost w programie SPSS Modeler a parametrami biblioteki Python XGBoost.

Tabela 35. Właściwości węzła odwzorowane na parametry biblioteki Python

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr XGBoost
Zmienna przewidywana	TargetField	
Predyktory	InputFields	
Metoda drzewa	treeMethod	tree_method
Liczba rund wzmocnienia	numBoostRound	num_boost_round
Maks. głębokość	maxDepth	max_depth
Maks. waga elementu podrzędnego	minChildWeight	min_child_weight
Maks. krok zmiany	maxDeltaStep	max_delta_step
Cel	objectiveType	objective
Przedwczesne zatrzymanie	earlyStopping	early_stopping_rounds
Rundy zatrzymania	stoppingRounds	
Współczynnik danych ewaluacji	evaluationDataRatio	
Losowa wartość początkowa	random_seed	seed
Podpróba	sampleSize	subsample
Eta	eta	eta
Gamma	gamma	gamma
Próba z kolumn na każde drzewo	colsSampleRatio	colsample_bytree
Próba z kolumn na każdy poziom	colsSampleLevel	colsample_bylevel
Lambda	lambda	lambda
Alfa	alpha	alpha
Skaluj wagi dodatnie	scalePosWeight	scale_pos_weight

[1] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. WWW. © 2015-2016 DMLC.

[2] "Plotting API" *Scalable and Flexible Gradient Boosting*. WWW. © 2015-2016 DMLC.

[3] "Scalable and Flexible Gradient Boosting." WWW. © 2015-2016 DMLC.

Opcje modelu węzła Drzewo XGBoost

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Węzeł t-SNE

Stochastyczna metoda porządkowania sąsiadów w oparciu o rozkład t (t-SNE — t-Distributed Stochastic Neighbor Embedding©) to narzędzie do wizualizacji danych wysokowymiarowych. Przekształca ono powinowactwa punktów danych w prawdopodobieństwa. Powinowactwa w przestrzeni pierwotnej są reprezentowane przez gaussowskie prawdopodobieństwa łączne, a powinowactwa w przestrzeni włączanej są reprezentowane przez rozkłady t Studenta. Dzięki temu algorytm t-SNE jest szczególnie czuły na struktury lokalne i ma kilka innych przewag nad wcześniej stosowanymi technikami: [1]

- Ujawnianie struktur w wielu skalach na jednej mapie
- Ujawnianie danych leżących w wielu różnych rozgałęzieniach lub grupach
- Ograniczenie tendencji do skupiania punktów w środku

Węzeł t-SNE w programie SPSS Modeler został zaimplementowany w języku Python i wymaga biblioteki Python scikit-learn©. Aby uzyskać szczegółowe informacje o algorytmie t-SNE i bibliotece scikit-learn, patrz:

- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE
- https://scikit-learn.org/stable/modules/manifold.html#t-sne

Karta Python na palecie węzłów zawiera ten i inne węzły Python. Węzeł t-SNE jest także dostępny na karcie Wykresy.

[1] Piśmiennictwo:

van der Maaten, L.J.P.; Hinton, G. „Visualizing High-Dimensional Data using t-SNE”. Journal of Machine Learning Research. 9:2579-2605, 2008.

van der Maaten, L.J.P. „t-Distributed Stochastic Neighbor Embedding”.

van der Maaten, L.J.P. „Accelerating t-SNE using Tree-Based Algorithms”. Journal of Machine Learning Research. 15(Oct):3221-3245, 2014.

Opcje zaawansowane węzła t-SNE

Wybierz tryb **Prosty** albo **Zaawansowany** w zależności od tego, z których opcji węzła t-SNE chce się korzystać.

Typ wizualizacji. Wybierz **2W** albo **3W**, aby określić, czy wykres ma być rysowany jako dwuwymiarowy czy trójwymiarowy.

Metoda. Wybierz metodę **Barnes Hut** albo **Dokładnie**. Domyślnie algorytm obliczania gradientu używa aproksymacji Barnesa-Huta, która jest znacznie szybsza niż metoda dokładna. Aproksymacja Barnesa-Huta umożliwia stosowanie techniki t-SNE do dużych zbiorów danych spotykanych w świecie rzeczywistym. Algorytm dokładny zapewni skuteczniejsze unikanie błędów najbliższego sąsiedztwa.

Rozpoczęcie Wybierz metodę inicjowania włączania: **Losowe** albo **PCA**.

Zmienna przewidywana. Wybierz zmienną przewidywaną, która ma zostać przedstawiona jako mapa kolorów na wynikowym wykresie. Jeśli zmienna przewidywana nie zostanie tutaj określona, wykres będzie jednokolorowy.

Optymalizacja

Zagmatwanie. Stopień zagmatwania związany z liczbą najbliższych sąsiadów używanych w innych algorytmach typu Manifold Learning. Większe zbiory danych zwykle wymagają większego zagmatwania. Należy rozważyć wybór wartości z przedziału od **5** do **50**. Wartość domyślna to **30**, a zakres wynosi **2 - 9999999**.

Wstępne wyolbrzymienie. To ustawienie określa, jak ciasno upakowane będą grupy naturalne w przestrzeni włączanej i ile miejsca pozostanie między nimi. Wartość domyślna to **12**, a zakres wynosi **2 - 9999999**.

Współczynnik uczenia. Jeśli współczynnik uczenia będzie za wysoki, dane mogą przypominać „piłkę", a każdy z punktów będzie w przybliżeniu równoodległy od najbliższych sąsiadów. Jeśli Współczynnik uczenia będzie za niski, większość punktów może skupić się w gęstą chmurę z niewielką liczbą punktów odstających. Jeśli funkcja kosztu utknie w nieprawidłowym minimum lokalnym, rozwiązaniem może być zwiększenie współczynnika uczenia. Wartość domyślna to **200**, a zakres wynosi **0 - 9999999**.

Maks. liczba iteracji Maksymalna liczba iteracji optymalizacji. Wartość domyślna to **1000**, a zakres wynosi **250 - 9999999**.

Wielkość kątowa. Wielkość kątowa odległego węzła zmierzona z punktu. Wprowadź wartość z zakresu od **0** do **1**. Wartością domyślną jest **0,5**.

Wartość początkowa

Ustaw wartość początkową generatora liczb losowych. Wybierz tę opcję i kliknij przycisk **Utwórz**, aby wygenerować wartość początkową dla generatora liczb losowych.

Warunek zatrzymania optymalizacji

Maks. liczba iteracji bez postępu. Maksymalna liczba iteracji bez postępu, po której optymalizacja ma zostać zatrzymana. Zaczyna obowiązywać po 250 iteracjach początkowych w przypadku użycia wstępnej nadmiarowości. Należy zwrócić uwagę, że postęp jest sprawdzany co 50 iteracji, zatem ta wartość zostanie zaokrąglona do następnej wielokrotności 50. Wartość domyślna to **300**, a zakres wynosi **0 - 9999999**.

Min. norma gradientu. Jeśli norma gradientu będzie niższa od tego progu, optymalizacja zostanie przerwana. Wartość domyślna to **1.0E-7**.

Metric. Metryka, która ma być stosowana przy obliczaniu odległości między wystąpieniami w macierzy predyktorów. Jeśli metryka jest łańcuchem, musi być jedną z opcji dozwolonych jako parametr metric funkcji scipy.spatial.distance.pdist lub metryką wymienioną w wyliczeniu pairwise.PAIRWISE_DISTANCE_FUNCTIONS. Wybierz jeden z dostępnych typów metryk. Wartością domyślną jest **euclidean**.

Jeśli liczba rekordów jest większa od. Podaj metodę tworzenia większych zbiorów danych. Można określić maksymalną wielkość zbioru danych lub użyć domyślnej liczby punków, wynoszącej 2000. Wydajność dla dużych zbiorów danych ulega poprawie po wybraniu opcji **Kategoria** lub **Próba**. Alternatywnie można wybrać wykreślanie wszystkich punktów danych przez wybór opcji **Użyj wszystkich danych**; należy jednak zwrócić uwagę, że może to drastycznie obniżyć wydajność oprogramowania.

- **Kategoria.** Tę opcję należy wybrać, aby aktywować kategoryzację, jeśli zbiór danych zawiera więcej rekordów niż określona liczba. Kategoryzacja dzieli wykres na mniejsze siatki przed rzeczywistym wykreśleniem go i zlicza liczbę połączeń, jakie zostaną wyświetlone w każdej komórce siatki. Na końcowym wykresie w środku ciężkości kategorii wykreślane jest jedno połączenie dla każdej komórki (średnia wszystkich punktów połączeń w kategorii).
- **Przykład.** Tę opcję należy wybrać, aby w przeprowadzić próbę losową danych dla określonej liczby rekordów.

W poniższej tabeli przedstawiono relację między ustawieniami na karcie Zaawansowane okna dialogowego t-SNE w programie SPSS Modeler a parametrami biblioteki t-SNE w języku Python.

Tabela 36. Właściwości węzła odwzorowane na parametry biblioteki Python

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr t-SNE w języku Python
Dominanta	mode_type	
Typ wizualizacji	n_components	n_components
Metoda	method	method
Inicjowanie włączania	init	init
Zmienna przewidywana	target_field	target_field
Zagmatwanie	perplexity	perplexity
Wstępne wyolbrzymienie	early_exaggeration	early_exaggeration
Współczynnik uczenia	learning_rate	learning_rate
Maks. liczba iteracji	n_iter	n_iter
Wielkość kątowa	angle	angle
Ustaw wartość startową generatora liczb losowych	enable_random_seed	
Wartość początkowa	random_seed	random_state
Maks. liczba iteracji bez postępu	n_iter_without_progress	n_iter_without_progress
Min. norma gradientu	min_grad_norm	min_grad_norm
Wykonaj t-SNE z wieloma poziomami zagmatwania	isGridSearch	

Opcje wyników węzła t-SNE

Określ opcje generowania wyników węzła t-SNE na karcie **Wynik**.

Nazwa wyniku. Określa nazwę wyniku uzyskanego po wykonaniu węzła. Wybranie opcji **Automatycznie** powoduje, że nazwa pliku wynikowego będzie określana automatycznie.

Wynik na ekran. Wybierz tę opcję, aby wygenerować i wyświetlić wynik w nowym oknie. Wynik jest także dodawany do Menedżera wyników.

Wynik do pliku. Ta opcja umożliwia zapisanie wyniku w pliku. Wybranie jej powoduje uaktywnienie pól **Nazwa pliku** i **Typ pliku**. Węzeł t-SNE musi mieć dostęp do tego pliku wyników, jeśli w celach porównawczych wykresy mają być tworzone przy użyciu innych zmiennych — lub jeśli wyniki mają być używane jako predyktory w modelach klasyfikacji lub regresji. Model t-SNE tworzy plik wynikowy zmiennych x, y (i z), do którego najłatwiej uzyskać dostęp za pośrednictwem węzła Plik kolumnowy.

Modele użytkowe t-SNE

Modele użytkowe t-SNE zawierają wszystkie informacje zgromadzone przez model t-SNE. Dostępne są następujące karty.

Wykres

Na karcie **Wykres** wyświetlany jest wynik węzła t-SNE w formie graficznej. Wykres pyplot rozrzutu przedstawia wynik niskowymiarowy. Jeśli na karcie Zaawansowany węzła t-SNE nie zostanie zaznaczona opcja **Wykonaj t-SNE z wieloma poziomami zagmatwania**, to uwzględniony będzie tylko jeden wykres, a nie sześć z różnymi poziomami zagmatwania.

Wyniki tekstowe

Na karcie **Wynik tekstowy** wyświetlane są wyniki działania algorytmu t-SNE. Jeśli na karcie Zaawansowany węzła t-SNE wybrano typ wizualizacji **2W**, to wynik będzie przedstawiony jako wartości punktów w dwóch wymiarach. W przypadku wybrania opcji **3W** wynik będzie przedstawiony jako wartości punktów w trzech wymiarach.

Węzeł mieszaniny rozkładów Gaussa

Model mieszanin rozkładów Gaussa — Gaussian Mixture© — to model probabilistyczny, w którym zakłada się, że punkty danych generowane są na podstawie mieszaniny skończonej liczby rozkładów Gaussa o nieznanych parametrach. Modele mieszanin można opisać jako uogólnienie grupowania metodą k-średnich z uwzględnieniem informacji o strukturze kowariancji danych oraz środkach ukrytych rozkładów Gaussa[1]

Węzeł Mieszanina rozkładów Gaussa w produkcie SPSS Modeler eksponuje podstawowe funkcje i często używane parametry biblioteki Gaussian Mixtore. Węzeł jest zaimplementowany w języku Python.

Więcej informacji o algorytmach modelowania mieszanin rozkładów Gaussa i ich parametrach zawiera dokumentacja dostępna na stronach http://scikit-learn.org/stable/modules/mixture.html i https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html. [2]

[1] "Podręcznik użytkownika." *Gaussian mixture models*. WWW. © 2007 - 2017. scikit-learn developers.

[2] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Węzeł Mieszanina rozkładów Gaussa — Zmienne

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Użyj wstępnie zdefiniowanych ról. Ta opcja wykorzystuje ustawienia danych wejściowych z poprzedzającego węzła Typ (lub z karty Typ poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Wybierz tę opcję, aby ręcznie przypisać dane wejściowe.

Zmienne. Za pomocą przycisków strzałek przypisz elementy z listy ręcznie na listę predyktorów po prawej stronie ekranu. Ikony wskazują prawidłowe poziomy pomiaru dla każdej zmiennej. Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Predyktory. Wybierz jedną lub więcej zmiennych jako predyktory.

Węzeł Mieszanina rozkładów Gaussa — Opcje budowania

Karta Opcje budowania służy do określania opcji budowania dla węzła Mieszanina rozkładów Gaussa, w tym **opcji podstawowych** i **opcji zaawansowanych**. Więcej informacji na temat opcji nieomówionych w tej sekcji można znaleźć w następujących źródłach internetowych:

- Skorowidz parametrów węzła Mieszanina rozkładów Gaussa[1]
- Podręcznik użytkownika węzła Mieszanina rozkładów Gaussa[2]

Podstawowe

Typ kowariancji. Wybierz jedną z następujących macierzy kowariancji:

- **Pełna.** Każdy komponent ma własną ogólną macierz kowariancji.
- **Powiązana.** Wszystkie komponenty mają tę samą ogólną macierz kowariancji.
- **Diagonalna.** Każdy komponent ma własną diagonalną macierz kowariancji.
- **Sferyczna.** Każdy komponent ma własną jedną wariancję.

Liczba komponentów. Określ liczbę komponentów mieszaniny, które mają być używane przy budowaniu modelu.

Etykieta skupienia. Określ, czy etykieta skupienia jest liczbą, czy łańcuchem. W przypadku wybrania opcji **Łańcuch** należy określić przedrostek dla etykiety skupienia (na przykład domyślny przedrostek to skupienie, co powoduje utworzenie etykiet, takich jak **skupienie-1**, **skupienie-2** itd.).

Losowa wartość początkowa. Wybierz tę opcję i kliknij przycisk **Utwórz**, aby wygenerować wartość początkową dla generatora liczb losowych.

Zaawansowane

Tolerancja. Określić próg zbieżności. Wartość domyślna to **0,001**.

Liczba iteracji. Określ maksymalną liczbę wykonywanych iteracji. Wartość domyślna to **100**.

Parametry początkowe. Wybierz parametr początkowy **Kmeans** (odpowiedzialności są inicjowane przy użyciu k-średnich) albo **Losowe** (odpowiedzialności są inicjowane losowo).

Gorący start. Wybranie opcji **Prawda** spowoduje, że zainicjowania następnego dopasowania zostanie użyte rozwiązanie ostatniego dopasowania. Może to przyspieszyć uzyskanie zbieżności, gdy dopasowanie wywoływane jest kilka razy w odniesieniu do tych samych problemów.

W poniższej tabeli przedstawiono relację między ustawieniami w oknie dialogowym węzła Mieszanina rozkładów Gaussa w programie SPSS Modeler a parametrami biblioteki Gaussian Mixture w języku Python.

Tabela 37. Właściwości węzła odwzorowane na parametry biblioteki Python

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr mieszaniny rozkładów Gaussa
Użyj wstępnie zdefiniowanych ról/Użyj niestandardowych przypisań zmiennych	role_use	
Dane wejściowe	predyktory	
Użyj danych podzielonych na podzbiory	use_partition	
Typ kowariancji	covariance_type	covariance_type
Liczba komponentów	number_component	n_components
Etykieta skupienia	component_lable	
Przedrostek etykiety	label_prefix	
Ustaw wartość startową generatora liczb losowych	enable_random_seed	
Losowa wartość początkowa	random_seed	random_state
Tolerancja	tol	tol
Liczba iteracji	max_iter	max_iter
Parametry początkowe	init_params	init_params
Gorący start	warm_start	warm_start

[1] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[2] "Podręcznik użytkownika." *Gaussian mixture models*. WWW. © 2007 - 2017. scikit-learn developers.

Węzeł Mieszanina rozkładów Gaussa — Opcje modelu

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Węzły KDE

Jądrowy estymator gęstości — Kernel Density Estimation (KDE)© — używa algorytmów Ball Tree lub KD Tree do efektywnej obsługi zapytań i działa na pograniczu między uczeniem nienadzorowanym, generowaniem cech (feature enginnering) i modelowaniem danych. Do najpopularniejszych i najbardziej użytecznych technik estymacji gęstości należą metody oparte na analizie sąsiedztwa, takie jak KDE. Algorytm KDE może być realizowany w dowolnej liczbie wymiarów, jednak w praktyce duża liczba wymiarów powoduje pogorszenie wydajności. Węzły Modelowanie KDE i Symulacja KDE w produkcie SPSS Modeler eksponują podstawowe funkcje i często używane parametry biblioteki KDE. Węzły są zaimplementowane w języku Python. [1]

Aby użyć węzła KDE, należy skonfigurować poprzedzający węzeł Typ. Węzeł KDE odczyta wprowadzane wartości z węzła Typ (lub karty Typy poprzedzającego węzła źródłowego).

Węzeł **Modelowanie KDE** jest dostępny na kartach Modelowanie i Python w programie SPSS Modeler. Węzeł KDE Modeling generuje model użytkowy, a wartości oceniane przez model użytkowy są gęstościami jądra z danych wejściowych.

Węzeł **Symulacja KDE** jest dostępny na karcie wyników i karcie Python. Węzeł Symulacja KDE generuje węzeł źródłowy Gen. KDE, który może utworzyć rekordy o tym samym rozkładzie, co dane wejściowe. Węzeł Gen. KDE zawiera kartę Ustawienia, na której można określić liczbę utworzonych rekordów (domyślnie 1) i wygenerować wartość startową generatora liczb losowych.

Więcej informacji na temat algorytmów KDE, wraz z przykładami, znajduje się w dokumentacji algorytmów KDE dostępnej pod adresem http://scikit-learn.org/stable/modules/density.html#kernel-density-estimation. [1]

[1] "User Guide." *Kernel Density Estimation*. WWW. © 2007-2018, scikit-learn developers.

Węzeł Modelowanie KDE węzeł Symulacja KDE — Zmienne

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Użyj wstępnie zdefiniowanych ról. Ta opcja wykorzystuje ustawienia danych wejściowych z poprzedzającego węzła Typ (lub z karty Typ poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Wybierz tę opcję, aby ręcznie przypisać dane wejściowe.

Zmienne. Za pomocą przycisków strzałek przypisz elementy z listy ręcznie na listę danych wejściowych po prawej stronie ekranu. Ikony wskazują prawidłowe poziomy pomiaru dla każdej zmiennej. Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Zmienne wejściowe. Można wybrać jedną lub więcej zmiennych jako dane wejściowe dla grupowania. Węzeł KDE może działać tylko na zmiennych ciągłych.

Węzły KDE — Opcje budowania

Karta Opcje budowania umożliwia określenie opcji budowania dla węzłów KDE, w tym **opcji podstawowych** dotyczących parametrów gęstości jądra oraz etykiet skupień, a także **opcji zaawansowanych** takich jak tolerancja, wielkość liścia i stosowanie metody „najpierw szerokość”. Więcej informacji na temat tych opcji można znaleźć w następujących źródłach internetowych:

- Skorowidz parametrów węzła jądrowej estymacji gęstości w interfejsie API środowiska Python[1]

- Podręcznik użytkownika jądrowej estymacji gęstości[2]

Podstawowe

Przepustowość. Określ przepustowość jądra.

Jądro. Wybierz jądro (algorytm domyślny), które ma być używane. Jądra dostępne dla węzła Modelowanie KDE to: **Gaussian**, **Tophat**, **Epanechnikov**, **Wykładniczy**, **Liniowy** i **Cosinus**. Jądra dostępne dla węzła Symulacja KDE to: **Gaussian** i **Tophat**. Szczegółowe informacje o dostępnych jądrach zawiera Podręcznik użytkownika jądrowej estymacji gęstości.[2]

Algorytm. Jako algorytm drzewa wybierz **Automatyczny**, **Ball Tree** lub **Drzewo KD**. Aby uzyskać więcej informacji — patrz Ball Tree[3] i KD Tree.[4]

Metric. Wybierz metrykę odległości. Dostępne są metryki: **Euclidean**, **Braycurtis**, **Chebyshev**, **Canberra**, **Cityblock**, **Dice**, **Hamming**, **Infinity**, **Jaccard**, **L1**, **L2**, **Matching**, **Manhattan**, **P**, **Rogerstanimoto**, **Russellrao**, **Sokalmichener**, **Sokalsneath**, **Kulsinski** i **Minkowski**. W przypadku wybrania miary **Minkowski** należy ustawić żądaną wartość w polu **Wartość P**.

To, które metryki dostępne będą w tym menu rozwijanym, zależy od wybranego algorytmu. Należy także zwrócić uwagę, że normalizacja wynikowej gęstości jest prawidłowa tylko dla metryki Euclidean.

Zaawansowane

Tolerancja bezwzględna. Określ żądaną tolerancję bezwzględną wyniku. Większa tolerancja z reguły przyspiesza wykonanie algorytmu. Wartością domyślną jest **0,0**.

Tolerancja względna. Określ żądaną tolerancję względną wyniku. Większa tolerancja z reguły przyspiesza wykonanie algorytmu. Wartością domyślną jest **1E-8**.

Wielkość liścia. Określ wielkość liścia podstawowego drzewa. Wartością domyślną jest **40**. Zmiana wielkości liścia może istotnie wpłynąć na wydajność oraz na zapotrzebowanie na pamięć. Aby uzyskać więcej informacji o algorytmach Ball Tree i KD Tree — patrz Ball Tree[3] i KD Tree.[4]

Najpierw szerokość. Wybierz opcję **Prawda**, jeśli ma być stosowana metoda „najpierw szerokość”, a **Fałsz**, jeśli ma być stosowana metoda „najpierw głębokość”.

W poniższej tabeli przedstawiono relację między ustawieniami w oknach dialogowych węzłów KDE w programie SPSS Modeler a parametrami biblioteki KDE w środowisku Spark.

Tabela 38. Właściwości węzła odwzorowane na parametry biblioteki Python

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr KDE
Dane wejściowe	inputs	
Przepustowość	bandwidth	bandwidth
Jądro	kernel	kernel
Algorytm	algorithm	algorithm
Metryka	metric	metric
Wartość P	pValue	pValue
Tolerancja bezwzględna	atol	atol
Tolerancja względna	rtol	Rtol
Wielkość liścia	leafSize	leafSize

Tabela 38. Właściwości węzła odwzorowane na parametry biblioteki Python (kontynuacja)

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr KDE
Najpierw szerokość	breadthFirst	breadthFirst

[1] „Informacje o API.” *sklearn.neighbors.KernelDensity*. WWW. © 2007-2018, scikit-learn developers.

[2] "User Guide." *Kernel Density Estimation*. WWW. © 2007-2018, scikit-learn developers.

[3] "Ball Tree." *Five balltree construction algorithms*. © 1989, Omohundro, S.M., International Computer Science Institute Technical Report.

[4] "K-D Tree." *Multidimensional binary search trees used for associative searching*. © 1975, Bentley, J.L., Communications of the ACM.

Węzeł Modelowanie KDE i węzeł Symulacja KDE — Opcje modelu

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Węzeł Las losowy

Las losowy (Random Forest©) to zaawansowana implementacja algorytmu agregacji (bagging), która jako model bazowy wykorzystuje model drzewa. W lasach losowych każde drzewo w zespole jest zbudowane z prób losowanych ze zwracaniem (np. próby bootstrapowej) ze zbioru uczącego. Podczas dzielenia węzła w trakcie tworzenia drzewa wybrany podział nie jest już najlepszym spośród wszystkich predyktorów. Zamiast tego wybierany jest najlepszy podział z losowego podzbioru predyktorów. Ze względu na tę losowość odchylenie lasu zwykle nieznacznie wzrasta (względem odchylenia jednego drzewa nielosowego), jednak w wyniku uśredniania zmniejsza się także jego zmienność — zwykle w stopniu z nawiązką kompensującym wzrost odchylenia. W rezultacie ogólna jakość modelu jest wyższa.[1]

Węzeł Las losowy w programie SPSS Modeler jest zaimplementowany w języku Python. Karta Python na palecie węzłów zawiera ten i inne węzły Python.

Aby uzyskać więcej informacji o algorytmach lasu losowego, patrz https://scikit-learn.org/stable/modules/ensemble.html#forest.

[1]L. Breiman, "Random Forests," Machine Learning, 45(1), 5-32, 2001.

Węzeł Las losowy — Zmienne

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Wybierz tę opcję, aby ręcznie przypisać zmienną przewidywaną.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli. Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji.

Predyktory. Można wybrać jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Węzeł Las losowy — Opcje budowania

Karta Opcje budowania służy do określania opcji budowania dla węzła Las losowy, w tym **opcji podstawowych** i **opcji zaawansowanych**. Aby uzyskać więcej informacji o tych opcjach, patrz https://scikit-learn.org/stable/modules/ensemble.html#forest

Podstawowe

Liczba drzew do utworzenia. Wybierz liczbę drzew w lesie.

Określ maksymalną głębokość. Gdy ta opcja nie jest zaznaczona, węzły są rozbudowywane do czasu, aż wszystkie węzły będą puste lub wszystkie węzły będą zawierać mniej niż min_samples_split prób.

Maks. głębokość. Maksymalna głębokość drzewa.

Minimalna wielkość węzła-liścia. Minimalna liczba prób, z jakiej musi składać się węzeł-liść.

Liczba predyktorów, jaka ma być używana do podziału. Liczba predyktorów, jaka ma być brana pod uwagę przy poszukiwaniu najlepszego podziału:

- Przy ustawieniu auto przyjmuje się max_features=sqrt(n_features) dla klasyfikatora i max_features=sqrt(n_features) dla regresji.
- Przy ustawieniu sqrt przyjmuje się max_features=sqrt(n_features).
- Przy ustawieniu log2 przyjmuje się max_features=log2(n_features).

Zaawansowane

Użycie prób bootstrapowych przy budowaniu drzew. Gdy ta opcja jest wybrana, przy budowaniu drzew używane są próby bootstrapowe.

Używaj prób OOB do szacowania dokładności uogólnienia. Powoduje użycie prób spoza zbioru uczącego (out-of-bag, OOB) do oszacowania dokładności uogólnienia.

Używaj ekstremalnie randomizowanych drzew. Powoduje użycie skrajnie randomizowanych drzew zamiast ogólnych lasów losowych. W przypadku skrajnej randomizacji przy obliczaniu podziałów obowiązuje większa losowość. Podobnie jak w lasach losowych używany jest losowy podzbiór potencjalnych predyktorów, ale zamiast wyszukiwania najbardziej rozróżniających progów stosuje się progi losowane dla każdego potencjalnego predyktora i jako regułę podziału przyjmuje się najlepsze z tych losowo wygenerowanych progów. Umożliwia to zwykle dalsze nieznaczne ograniczenie zmienności modelu kosztem nieznacznego zwiększenia odchylenia. [1]

Replikacja wyników. Gdy ta opcja jest wybrana, proces budowania modelu jest replikowany w celu uzyskania tych samych wyników oceny.

Wartość początkowa. Można kliknąć przycisk **Utwórz**, aby wygenerować wartość początkową dla generatora liczb losowych.

Optymalizacja hiperparametrów (wg Rbfopt). Wybranie tej opcji włącza optymalizację hiperparametrów opartą na Rbfopt, która automatycznie wykrywa optymalną kombinację parametrów, przy której model osiągnie oczekiwany lub niższy od oczekiwanego wskaźnika błędów dla prób. Aby uzyskać szczegółowe informacje na temat biblioteki Rbfopt, patrz http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html.

Zmienna przewidywana. Wartość funkcji celu (wskaźnik błędu modelu dla prób), którą chcemy osiągnąć (tj. wartość nieznanego optimum). Należy podać wartość akceptowalną, na przykład 0,01.

Maks. liczba iteracji Maksymalna liczba iteracji na modelu. Wartość domyślna to 1000.

Maks. liczba ewaluacji. Określa, ile razy ma być wyznaczana wartość funkcji w trybie dokładnym. Wartość domyślna to 300.

W poniższej tabeli przedstawiono relację między ustawieniami w oknie dialogowym węzła Las losowy w programie SPSS Modeler a parametrami biblioteki Las losowy w języku Python.

Tabela 39. Właściwości węzła odwzorowane na parametry biblioteki Python

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr biblioteki Las losowy
Zmienna przewidywana	target	
Predyktory	inputs	
Liczba drzew do utworzenia	n_estimators	n_estimators
Określ maksymalną głębokość	specify_max_depth	specify_max_depth
Maks. głębokość	max_depth	max_depth
Minimalna wielkość węzła-liścia	min_samples_leaf	min_samples_leaf
Liczba predyktorów, jaka ma być używana do podziału	max_features	max_features
Używaj prób bootstrapowych przy budowaniu drzew	bootstrap	bootstrap
Używaj prób OOB do szacowania dokładności uogólnienia	oob_score	oob_score
Używaj ekstremalnie randomizowanych drzew	extreme	
Replikuj wyniki	use_random_seed	
Wartość początkowa	random_seed	random_seed
Optymalizacja hiperparametrów (wg Rbfopt)	enable_hpo	
Zmienna przewidywana (dla optymalizacji HPO)	target_objval	
Maks. liczba iteracji (dla optymalizacji HPO)	max_iterations	
Maks. liczba ewaluacji (dla optymalizacji HPO)	max_evaluations	

[1]L. Breiman, "Random Forests," Machine Learning, 45(1), 5-32, 2001.

Węzeł Las losowy — Opcje modelu

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Modele użytkowe Las losowy

Modele użytkowe Las losowy zawierają wszystkie informacje zgromadzone przez model lasu losowego. Dostępne są następujące sekcje.

Informacje o modelu

Ten widok zawiera najważniejsze informacje o modelu, w tym zmienne wejściowe, wartości kodowania One-Hot i parametry modelu.

Ważność predyktorów

Ten widok przedstawia ważność względną każdego predyktora w procesie estymacji modelu. Aby uzyskać więcej informacji, zobacz “Ważność predyktorów” na stronie 43.

Węzeł HDBSCAN

Węzeł Hierarchical Density-Based Spatial Clustering (HDBSCAN)© korzysta z algorytmu uczenia nienadzorowanego, aby wyszukać skupienia lub regiony o dużej gęstości w zbiorze danych. Węzeł HDBSCAN w programie SPSS Modeler eksponuje tylko funkcje podstawowe i najczęściej używane parametry biblioteki HDBSCAN. Węzeł jest zaimplementowany w języku Python i można go użyć do skupiania zbioru danych w osobne grupy, jeśli nie wiemy z góry, co to są za grupy. W przeciwieństwie do większości metod uczenia w produkcie SPSS Modeler w modelach HDBSCAN *nie* są stosowane zmienne przewidywane. Sposób uczenia bez zmiennej przewidywanej jest nazywany *uczeniem nienadzorowanym*. Węzeł HDBSCAN nie próbuje przewidzieć wyniku, lecz ujawnia istniejące wzorce w zbiorze zmiennych wejściowych. Rekordy są grupowane w taki sposób, aby rekordy w ramach grupy lub skupienia były do siebie podobne, zaś rekordy z różnych grup były do siebie niepodobne. Algorytm HDBSCAN wyświetla skupienia jako obszary o dużej gęstości oddzielone obszarami o niskiej gęstości. Ze względu na ten dość ogólny widok, skupienia odnalezione przez węzeł HDBSCAN mogą mieć dowolny kształt, w przeciwieństwie do metody k-średnich, która zakłada, że skupienia mają kształt wypukły. Punkty wartości odstających, które leżą w regionach o niskiej gęstości, są również oznaczone. Węzeł HDBSCAN obsługuje również ocenianie nowych próbek.[1]

Aby użyć węzła HDBSCAN, należy skonfigurować poprzedzający węzeł Typ. Węzeł HDBSCAN odczyta wprowadzane wartości z węzła Typ (lub karty Typy poprzedzającego węzła źródłowego).

Więcej informacji na temat algorytmów grupowania HDBSCAN znajduje się w dokumentacji węzła HDBSCAN dostępnej pod adresem http://hdbscan.readthedocs.io/en/latest/. [1]

[1] „Podręcznik użytkownika / Samouczek.” *Biblioteka skupień hdbscan*. WWW. © 2016, Leland McInnes, John Healy, Steve Astels.

Zmienne węzła HDBSCAN

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Ważne: Aby uczyć model HDBSCAN, należy użyć co najmniej jednej zmiennej z rolą ustawioną na **Dane wejściowe**. Zmienne z rolą ustawioną na wartość **Wynik**, **Łącznie** lub **Brak** są ignorowane.

Użyj wstępnie zdefiniowanych ról. Ta opcja wykorzystuje ustawienia danych wejściowych z poprzedzającego węzła Typ (lub z karty Typ poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Wybierz tę opcję, aby ręcznie przypisać dane wejściowe.

Zmienne. Za pomocą przycisków strzałek przypisz elementy z listy ręcznie na listę danych wejściowych po prawej stronie ekranu. Ikony wskazują prawidłowe poziomy pomiaru dla każdej zmiennej. Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Zmienne wejściowe. Można wybrać jedną lub więcej zmiennych jako dane wejściowe dla grupowania.

Opcje budowania węzła HDBSCAN

Karta Opcje budowania umożliwia określenie opcji budowania dla węzła HDBSCAN, w tym **opcji ogólnych** dotyczących parametrów skupienia i **opcji zaawansowanych** dotyczących parametrów zaawansowanych i opcji wykresu wynikowego. Więcej informacji na temat tych opcji można znaleźć w następujących źródłach internetowych:

- Skorowidz parametrów węzła HDBSCAN w interfejsie API środowiska Python[1]
- Strona główna HDBSCAN[2]

Podstawowe

Optymalizacja hiperparametrów (wg Rbfopt). Wybranie tej opcji włącza optymalizację hiperparametrów opartą na Rbfopt, która automatycznie wykrywa optymalną kombinację parametrów, przy której model osiągnie oczekiwany lub niższy od oczekiwanego wskaźnika błędów dla prób. Aby uzyskać szczegółowe informacje na temat biblioteki Rbfopt, patrz http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html.

Minimalny rozmiar skupienia. Określ minimalny rozmiar skupień. Podziały pojedynczego wiązania, które zawierają mniej punktów niż podana tutaj wartość będą uznawane za punkty „wykraczające poza" skupienie, a nie za skupienie dzielące się na dwa nowe skupienia.

Min. liczba próbek. Określ minimalną liczbę próbek w sąsiedztwie dla punktu, aby punkt był uznawany za punkt główny. Jeśli wartość zostanie ustawiona na **0**, przyjęta zostanie domyślnie minimalna wielkość skupienia.

Algorytm. Wybierz algorytm, który ma być stosowany. Węzeł HDBSCAN ma warianty, które są wyspecjalizowane dla różnych charakterystyk danych. Domyślnie używany jest algorytm **BEST**, który automatycznie wybiera najlepszy algorytm na podstawie charakteru danych. Szczegóły na temat tych typów algorytmów znajdują się w dokumentacji HDBSCAN.[1] Należy pamiętać, że wybór algorytmu wpływa na wydajność. Na przykład do obszernych danych radzimy stosować algorytm Boruvka KDTree lub Boruvka BallTree.

Metryka odległości. Wybierz metrykę, która ma być stosowana przy obliczaniu odległości między wystąpieniami w macierzy predyktorów.

Etykieta skupienia. Określ, czy etykieta skupienia jest liczbą, czy łańcuchem. W przypadku wybrania opcji **Łańcuch** należy określić przedrostek dla etykiety skupienia (na przykład domyślny przedrostek to skupienie, co powoduje utworzenie etykiet, takich jak **skupienie-1**, **skupienie-2** itd.).

Zaawansowane

Przybliżone minimalne drzewo rozpinające. Wybierz opcję **Prawda**, jeśli przybliżone minimalne drzewo rozpinające ma zostać zaakceptowane. W przypadku niektórych algorytmów może to zwiększyć wydajność, ale wygenerowane skupienia mogą mieć nieznacznie niższą jakość. Jeśli skupienia mają być generowane wolniej, ale z większą dokładnością, wybierz opcję **Fałsz**. W większości przypadków zalecane jest użycie opcji **Prawda**.

Metoda wyboru skupienia. Wybierz metodę, która ma być używana do wyboru skupień z drzewa skondensowanego. Standardowe podejście dla węzła HDBSCAN to użycie algorytmu Excess of Mass (**EOM**) do odnalezienia najbardziej trwałych skupień. Można również wybrać skupienia w liściach drzewa, co zapewnia najbardziej precyzyjne i jednorodne skupienia.

Akceptuj jedno skupienie. Zmień to ustawienie na wartość **Prawda**, aby zezwolić wyłącznie na wyniki jednego skupienia, jeśli jest to prawidłowy wynik dla zbioru danych.

Wartość P. W przypadku stosowania metryki odległości Minkowskiego (pod **podstawowymi** opcjami budowania) w razie potrzeby można zmienić tę wartość p.

Wielkość liścia. W przypadku stosowania algorytmu drzewa przestrzeni (KDTree Boruvki lub BallTree Boruvki) jest to liczba punktów węzła-liścia drzewa. To ustawienie nie ma wpływu na wygenerowane skupienia, ale może niekorzystnie wpływać na czas wykonywania algorytmu.

Wskaźnik ważności. Wybierz tę opcję, aby uwzględnić wykres wskaźnika ważności w wyniku modelu użytkowego.

Drzewo skondensowane. Wybierz tę opcję, aby uwzględnić drzewo skondensowane w wyniku modelu użytkowego.

Drzewo pojedynczego wiązania. Wybierz tę opcję, aby uwzględnić drzewo pojedynczego wiązania w wyniku modelu użytkowego.

Minimalne drzewo rozpinające. Wybierz tę opcję, aby uwzględnić minimalne drzewo rozpinające w wyniku modelu użytkowego.

W następującej tabeli przedstawiono relacje między ustawieniami w oknie dialogowym węzła HDBSCAN w programie SPSS Modeler a parametrami biblioteki Python HDBSCAN.

Tabela 40. Właściwości węzła odwzorowane na parametry biblioteki Python

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr HDBSCAN
Dane wejściowe	inputs	inputs
Optymalizacja hiperparametrów	useHPO	
Minimalny rozmiar skupienia	min_cluster_size	min_cluster_size
Min. liczba próbek	min_samples	min_samples
Algorytm	algorithm	algorithm
Metryka odległości	metric	metric
Etykieta skupienia	useStringLabel	
Przedrostek etykiety	stringLabelPrefix	
Przybliżone minimalne drzewo rozpinające	approx_min_span_tree	approx_min_span_tree
Metoda wyboru skupienia	cluster_selection_method	cluster_selection_method
Akceptuj jedno skupienie	allow_single_cluster	allow_single_cluster
Wartość P	p_value	p_value
Wielkość liścia	leaf_size	leaf_size
Wskaźnik ważności	outputValidity	
Drzewo skondensowane	outputCondensed	
Drzewo pojedynczego wiązania	outputSingleLinkage	
Minimalne drzewo rozpinające	outputMinSpan	

[1] „Informacje o API.” *Biblioteka skupień hdbscan*. WWW. © 2016, Leland McInnes, John Healy, Steve Astels.

[2] „Podręcznik użytkownika / Samouczek.” *Biblioteka skupień hdbscan*. WWW. © 2016, Leland McInnes, John Healy, Steve Astels.

Opcje modelu węzła HDBSCAN

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Węzeł SVM z jedną klasą

Węzeł SVM z jedną klasą (One-Class SVM©) korzysta z algorytmu uczenia nienadzorowanego. Węzeł ten można wykorzystać do wykrywania nowości. Wykryje on miękką granicę danego zbioru próbek, a następnie sklasyfikuje nowe punkty jako należące do tego zbioru albo do niego nienależące. Węzeł modelowania SVM z jedną klasą został zaimplementowany w języku Python i wymaga biblioteki Python scikit-learn©. Szczegółowe informacje można znaleźć na stronie http://contrib.scikit-learn.org/imbalanced-learn/about.html[1].

Karta Python na palecie węzłów zawiera węzeł SVM z jedną klasą i inne węzły Python.

Uwaga: Węzeł SVM z jedną klasą służy do nienadzorowanego wykrywania wartości odstających i nowości. W większości przypadków zalecamy użycie do zbudowania modelu znanego, „normalnego" zbioru danych, który umożliwi algorytmowi wyznaczenie prawidłowej granicy dla danych próbek. Parametry modelu – takie jak nu, gamma i kernel – znacząco wpływają na wyniki. Konieczne może być eksperymentalny dobór tych opcji w celu znalezienia ustawień optymalnych w danej sytuacji.

[1]Smola, Schölkopf. "A Tutorial on Support Vector Regression." *Statistics and Computing Archive*, vol. 14, no. 3, August 2004, pp. 199-222. (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.114.4288)

Zmienne węzła SVM z jedną klasą

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Użyj wstępnie zdefiniowanych ról. Wybierz tę opcję, aby przypisać wszystkim zmiennym zdefiniowaną rolę wejściowych.

Użyj niestandardowych przypisań. Aby ręcznie wybrać zmienne, zaznacz tę opcję i wybierz zmienne wejściowe oraz zmienne podziału:

Zmienne wejściowe. Wybierz zmienne wejściowe, które mają być używane w analizie. Obsługiwane są wszystkie typy składowania i pomiaru, z wyjątkiem zmiennych bez typu i nieznanego typu. Jeśli zmienna ma typ składowania Łańcuch, jej wartości zostaną zbinaryzowane na zasadzie jedna-wszystkie przy użyciu algorytmu kodowania one-hot (z gorącą jedynką).

Podział. Wybierz, które zmienne mają być używane jako zmienne podziału. Obsługiwane są wszystkie typy pomiaru: flaga, nominalne, porządkowe i dyskretne.

Użyj danych podzielonych na podzbiory Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Opcje zaawansowane węzła SVM z jedną klasą

Na karcie ustawień zaawansowanych węzła SVM z jedną klasą można wybrać tryb **Prosty** albo **Zaawansowany**. W przypadku wybrania trybu **Prosty** wszystkim parametrom zostaną nadane wartości domyślne przedstawione poniżej. Wybranie trybu **Zaawansowany** umożliwi określenie niestandardowych wartości tych parametrów. Aby uzyskać więcej informacji o tych parametrach, patrz http://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html#sklearn.svm.OneClassSVM.

Kryteria zatrzymania. Określ tolerancję jako kryterium zatrzymania. Wartość domyślna to **1.0E-3** (0.001).

Precyzja regresji (nu). Związana z ułamkiem błędów uczenia i wektorów pokrycia. Wartość domyślna to **0.1**.

Typ jądra. Typ jądra algorytmu. Dostępne są jądra: **RBF**, **Wielomianowe**, **Sigmoidalne**, **Liniowe** i **Wstępnie obliczone**. Wartość domyślna to **RBF**.

Określ Gamma. Wybierz tę opcję, aby określić wartość Gamma. W przeciwnym razie użyta zostanie automatycznie określona wartość gamma.

Gamma. Ustawienie Gamma jest dostępne tylko dla algorytmów RBF, Wielomianowe i Sigmoidalne.

Coef0. Ustawienie Coef0 jest dostępne tylko dla algorytmów Wielomianowe i Sigmoidalne.

Stopnia. Ustawienie Stopień jest dostępne tylko dla algorytmu Wielomianowe.

Używaj heurystyki z redukcją. Wybierz tę opcję, aby używać heurystyki z redukcją. Ta opcja jest domyślnie niewybrana.

Ustaw wartość początkową generatora liczb losowych. Wybierz tę opcję, aby określić początkową wartość dla generatora liczb losowych używanego przy losowej zmianie kolejności danych na potrzeby oszacowania prawdopodobieństwa. Ta opcja jest domyślnie niewybrana.

Określ wielkość pamięci podręcznej jądra (w MB). Wybierz tę opcję, aby określić wielkość pamięci podręcznej jądra. Ta opcja jest domyślnie niewybrana. Gdy jest wybrana, wartość domyślna wynosi **200** MB.

Optymalizacja hiperparametrów (wg Rbfopt). Wybranie tej opcji włącza optymalizację hiperparametrów opartą na Rbfopt, która automatycznie wykrywa optymalną kombinację parametrów, przy której model osiągnie oczekiwany lub niższy od oczekiwanego wskaźnika błędów dla prób. Aby uzyskać szczegółowe informacje na temat biblioteki Rbfopt, patrz http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html.

Zmienna przewidywana. Wartość funkcji celu (wskaźnik błędu modelu dla prób), którą chcemy osiągnąć (na przykład wartość nieznanego optimum). Należy podać wartość akceptowalną, na przykład **0,01**.

Maks. liczba iteracji Maksymalna liczba iteracji na modelu. Wartość domyślna to **1000**.

Maks. liczba ewaluacji. Określa, ile razy maksymalnie zostanie wyznaczona wartość funkcji na modelu, w sytuacji gdy ważniejsza od szybkości jest dokładność. Wartość domyślna to **300**.

Węzeł SVM z jedną klasą wymaga biblioteki Python scikit-learn©. W poniższej tabeli przedstawiono relacje między ustawieniami w oknie dialogowym węzła SMOTE w programie SPSS Modeler a parametrami algorytmu w języku Python.

Tabela 41. Właściwości węzła odwzorowane na parametry biblioteki Python

Nazwa parametru	Nazwa w skryptach (nazwa właściwości)	Nazwa parametru w interfejsie API środowiska Python
Kryteria zatrzymania	stopping_criteria	tol
Precyzja regresji	precision	nu
Typ jądra	kernel	kernel
Gamma	gamma	gamma
Coef0	coef0	coef0
Stopień	degree	degree
Używaj heurystyki z redukcją	shrinking	shrinking
Określ wielkość pamięci podręcznej jądra (pole do wprowadzania liczby)	cache_size	cache_size
Wartość początkowa	random_seed	random_state

Opcje węzła SVM z jedną klasą

Na karcie Opcje węzła SVM z jedną klasą można określić następujące opcje.

Typ grafiki współrzędnych równoległych. SPSS Modeler rysuje grafikę współrzędnych równoległych, aby zaprezentować zbudowany model. Niekiedy wartości niektórych kolumn danych/elementów będą znacznie większe od innych, co utrudni obserwację pozostałych części wykresu. W takich sytuacjach można wybrać opcję **Niezależne osie pionowe**, aby każda oś pionowa miała swoją skalę, albo wybrać opcję **Ogólne osie pionowe**, aby wymusić tę samą skalę na wszystkich osiach pionowych.

Maksymalna liczba linii na grafice. Określ maksymalną liczbę wierszy danych (linii) do umieszczenia na wynikowym wykresie. Domyślną wartością jest 100. Ze względów wydajnościowych wyświetlonych zostanie maksymalnie 20 zmiennych.

Wykreśl wszystkie zmienne wejściowe w grafice. Wybierz tę opcję, aby wszystkie zmienne wejściowe były uwzględnione na wynikowym wykresie. Domyślnie każda zmienna danych będzie miała swoją oś pionową. Ze względów wydajnościowych wyświetlonych zostanie maksymalnie 30 zmiennych.

Zmienne niestandardowe do wykreślenia. Zamiast prezentować wszystkie zmienne wejściowe na wynikowym wykresie, można wybrać tę opcję i podzbiór zmiennych. Takie rozwiązanie może korzystnie wpłynąć na wydajność. Ze względów wydajnościowych wyświetlonych zostanie maksymalnie 20 zmiennych.

Rozdział 18. Węzły spark

SPSS Modeler udostępnia węzły umożliwiające bezpośrednie korzystanie z algorytmów środowiska Spark. Karta **Spark** na Palecie węzłów zawiera następujące węzły, których można używać do uruchamiania algorytmów Spark. Węzły te są obsługiwane na platformach Windows 64, Mac 64 i Linux 64. Należy pamiętać, że te węzły nie obsługują określania kolumny liczba całkowita/liczba typu double jako flaga/nominalna do budowania modelu. W tym celu należy przekształcić wartość kolumny na 0/1 lub 0,1,2,3,4...

Regresja izotoniczna należy do rodziny algorytmów regresji. Węzeł Izotoniczna-AS w programie SPSS Modeler jest zaimplementowany w środowisku Spark. Aby uzyskać szczegółowe informacje na temat algorytmów regresji izotonicznej, patrz https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html.

XGBoost© to zaawansowana implementacja algorytmu wzmacniania gradientowego. Algorytmy wzmacniania iteracyjnie ucząc się, wyznaczają słabe klasyfikatory i dodają je do ostatecznego silnego klasyfikatora. XGBoost jest algorytmem bardzo elastycznym i oferuje liczne parametry, które mogą być trudne do praktycznego wykorzystania przez większość użytkowników. Dlatego węzeł XGBoost-AS w programie SPSS Modeler eksponuje tylko funkcje podstawowe i najczęściej używane parametry. Węzeł XGBoost-AS jest zaimplementowany w środowisku Spark.

K-średnie to jeden z najpowszechniej używanych algorytmów grupowania. Grupuje on punkty danych w określoną z góry liczbę skupień. Węzeł K-średnie-AS w programie SPSS Modeler jest zaimplementowany w środowisku Spark. Aby uzyskać szczegółowe informacje na temat algorytmów K-średnich, patrz https://spark.apache.org/docs/2.2.0/ml-clustering.html. Należy zwrócić uwagę, że węzeł K-średnie-AS automatycznie wykonuje kodowanie one-hot (kodowanie z gorącą jedynką) dla zmiennych kategorialnych.

Węzeł Izotoniczna-AS

Regresja izotoniczna należy do rodziny algorytmów regresji. Węzeł Izotoniczna-AS w programie SPSS Modeler jest zaimplementowany w środowisku Spark.

Aby uzyskać szczegółowe informacje na temat algorytmów regresji izotonicznej, patrz https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html. [1]

[1] "Regression - RDD-based API." *Apache Spark*. MLlib: Main Guide. WWW. 3 października 2017 r.

Węzeł Izotoniczna-AS — Zmienne

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Zmienne. Lista wszystkich zmiennych w źródle danych. Aby ręcznie przypisać pozycje z tej listy do roli przewidywanej, wejściowej i wagi po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli. Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną.

Zmienna wejściowa. Umożliwia wybór zmiennej wejściowej lub kilku zmiennych.

Waga. Wybierz zmienną wagi wykładniczej. Jeśli zmienna nie zostanie określona, przyjęta zostanie domyślna wartość **1**.

Węzeł Izotoniczna-AS — Opcje budowania

Karta Opcje budowania służy do określania opcji budowania dla węzła Izotoniczna-AS, w tym indeksu predyktorów i typu izotonicznego. Aby uzyskać więcej informacji, patrz http://spark.apache.org/docs/latest/api/java/org/apache/spark/ml/regression/IsotonicRegression.html.[1]

Indeks zmiennych wejściowych. Określ indeks zmiennych wejściowych. Wartością domyślną jest **0**.

Typ izotoniczny. To ustawienie określa, czy kolejność wyników powinna być izotoniczna/rosnąca, czy antytoniczna/malejąca. Domyślnie wybrany jest typ **Izotoniczny**.

[1] "Class IsotonicRegression." *Apache Spark*. JavaDoc. WWW. 3 października 2017 r.

Modele użytkowe Izotoniczna-AS

Modele użytkowe Izotoniczna-AS zawierają wszystkie informacje zgromadzone przez model Izotoniczna-AS. Dostępne są następujące sekcje.

Podsumowanie modelu

Ten widok zawiera najważniejsze informacje o modelu, w tym zmienne wejściowe, zmienną przewidywaną i opcje budowania modelu.

Wykres modelu

Ten widok przedstawia wykres rozrzutu.

Węzeł XGBoost-AS

XGBoost© to zaawansowana implementacja algorytmu wzmacniania gradientowego. Algorytmy wzmacniania iteracyjnie ucząc się, wyznaczają słabe klasyfikatory i dodają je do ostatecznego silnego klasyfikatora. XGBoost jest algorytmem bardzo elastycznym i oferuje liczne parametry, które mogą być trudne do praktycznego wykorzystania przez większość użytkowników. Dlatego węzeł XGBoost-AS w programie SPSS Modeler eksponuje tylko funkcje podstawowe i najczęściej używane parametry. Węzeł XGBoost-AS jest zaimplementowany w środowisku Spark.

Więcej informacji o algorytmach wzmacniania zawierają kursy XGBoost dostępne na stronie http://xgboost.readthedocs.io/en/latest/tutorials/index.html. [1]

Uwaga: w programie SPSS Modeler nie jest obsługiwana funkcja walidacji krzyżowej XBoost. Ten sam cel można zrealizować za pomocą węzła Podział w programie SPSS Modeler. Ponadto algorytm XGBoost w SPSS Modeler automatycznie wykonuje kodowanie one-hot (kodowanie z gorącą jedynką) dla zmiennych kategorialnych.

Uwaga: Do budowania modeli XGBoost-AS na platformie Mac wymagana jest wersja 10.12.3 lub wyższa.

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. WWW. © 2015-2016 DMLC.

Węzeł XGBoost-AS — Zmienne

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Użyj wstępnie zdefiniowanych ról. Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Wybierz tę opcję, aby ręcznie przypisać zmienną przewidywaną.

Zmienne. Aby ręcznie przypisać pozycje z tej listy do zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli. Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Zmienna przewidywana. Wybierz jedną zmienną jako zmienną przewidywaną dla predykcji.

Predyktory. Można wybrać jedną lub więcej zmiennych jako dane wejściowe dla predykcji.

Węzeł XGBoost-AS — Opcje budowania

Karta Opcje budowania umożliwia określenie opcji budowania dla węzła XGBoost-AS, w tym **opcji ogólnych** dotyczących tworzenia modelu i postępowania z niezrównoważonymi zbiorami danych, **opcji uczenia** dotyczących celów i metryk oceny, a także **parametrów wzmocnienia**. Więcej informacji na temat tych opcji zawierają następujące zasoby internetowe:

- Strona główna XGBoost[1]
- Skorowidz parametrów XGBoost[2]
- Interfejs API XGBoost środowiska Spark[3]

Opcje ogólne

Liczba procesów roboczych. Liczba procesów roboczych używanych do uczenia modelu XGBoost.

Liczba wątków. Liczba wątków przypadająca na jeden proces roboczy.

Używaj pamięci zewnętrznej. Określa, czy używać pamięci zewnętrznej jako podręcznej.

Typ wzmocnienia. Typ wzmocnienia: **gbtree**, **gblinear** albo **dart**.

Liczba rund wzmocnienia. Liczba rund wzmocnienia.

Skaluj wagi dodatnie. To ustawienie wpływa na równowagę między wagami dodatnimi i ujemnymi. Jest użyteczne w przypadku klas niezrównoważonych.

Losowa wartość początkowa. Kliknij przycisk **Utwórz**, aby wygenerować wartość startową dla generatora liczb losowych.

Zadanie uczenia

Cel. Wybierz jeden z następujących typów celu zadania: **reg:linear**, **reg:logistic**, **reg:gamma**, **reg:tweedie**, **rank:pairwise**, **binary:logistic** lub **multi**.

Metryka ewaluacyjna. Metryka ewaluacyjna dla danych walidacji. Metryka domyślna zostanie przypisana zgodnie z celem (**rmse** dla regresji, **error** dla klasyfikacji albo **mean average precision** dla rankowania). Dostępne opcje: **rmse**, **mae**, **logloss**, **error**, **merror**, **mlogloss**, **uac**, **ndcg**, **map** i **gamma-deviance** (domyślnie **rmse**).

Parametry wzmocnienia

Lambda. Składnik regularyzacji L2 wag. Zwiększenie tej wartości powoduje, że model jest bardziej konserwatywny.

Alfa. Składnik regularyzacji L1 wag. Zwiększenie tej wartości powoduje, że model jest bardziej konserwatywny.

Obciążenie Lambda. Składnik regularyzacji L2 obciążenia. (Nie ma składnika regularyzacji L1 obciążenia, ponieważ jest nieistotny).

Metoda drzewa. Wybierz algorytm tworzenia drzewa XGBoost.

Maks. głębokość. Określ maksymalną głębokość drzew. Zwiększenie tej wartości zwiększa złożoność modelu i ryzyko przeuczenia.

Min. waga elementu podrzędnego. Określ minimalną sumę wagi wystąpień (Hessiana) wymaganą w elemencie podrzędnym. Gdy krok podziału drzewa doprowadzi do powstania węzła-liścia z sumą wag wystąpień mniejszą od wartości **Min. waga elementu podrzędnego**, podczas tworzenia drzewa nie będą już wprowadzane dalsze podziały. W trybie regresji liniowej wartość ta odpowiada po prostu minimalnej liczbie wystąpień wymaganej w każdym węźle. Im większa waga, tym bardziej konserwatywnie działa algorytm.

Maks. krok zmiany. Określ maksymalny krok zmiany umożliwiający oszacowanie wag drzewa. Wartość **0** oznacza brak ograniczenia. Wartość dodatnia powoduje, że krok aktualizacji może być realizowany bardziej konserwatywnie. Zwykle ten parametr nie jest potrzebny, ale może pomóc w przypadku regresji logistycznej, gdy klasa jest skrajnie niezrównoważona.

Podpróba. Podpróba określa współczynnik wystąpień używanych do uczenia. Na przykład przy wartości **0.5** algorytm XGBoost losowo wybierze połowę wystąpień danych do wzrostu drzewa. Takie ograniczenie zapobiega przeuczeniu.

Eta. Redukcja wielkości kroku używana podczas aktualizacji w celu zapobiegania przeuczeniu. Po każdym kroku wzmacniania wagi nowych predyktorów mogą być uzyskane bezpośrednio. Eta redukuje także wagi predyktorów, aby proces wzmacniania działał bardziej konserwatywnie.

Gamma. Minimalna redukcja straty wymagana do dalszego podziału węzła-liścia w drzewie. Im większa wartość gamma, tym bardziej konserwatywnie działa algorytm.

Próba z kolumn na każde drzewo. Współczynnik podpróbkowania kolumn podczas tworzenia każdego drzewa.

Próba z kolumn na każdy poziom. Współczynnik podpróbkowania kolumn dla każdego podziału na każdym poziomie.

Algorytm normalizacji. Algorytm normalizacji, który ma być używany w przypadku wybrania wzmocnienia dart w opcjach ogólnych. Dostępne opcje: **tree** i **forest** (domyślnie **tree**).

Algorytm próbkowania. Algorytm próbkowania, który ma być używany w przypadku wybrania wzmocnienia dart w opcjach ogólnych. Algorytm **uniform** jednorodnie wybiera drzewa do wypadnięcia. Algorytm **weighted** wybiera drzewa proporcjonalnie do wagi. Domyślnie używany jest algorytm **uniform**.

Współczynnik wypadania. Współczynnik wypadania, który ma być używany w przypadku wybrania wzmocnienia dart w opcjach ogólnych.

Prawdopodobieństwo pominiętego wypadnięcia. Prawdopodobieństwo pominiętego wypadnięcia, które ma być używane w przypadku wybrania wzmocnienia dart w opcjach ogólnych. W przypadku pominięcia wypadnięcia nowe drzewa dodawane są tak samo, jak po wybraniu opcji **gbtree**.

W poniższej tabeli przedstawiono relację między ustawieniami w oknie dialogowym węzła XGBoost-AS w programie SPSS Modeler a parametrami algorytmu XGBoost w środowisku Spark.

Tabela 42. Właściwości węzła odwzorowane na parametry Spark

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr XGBoost w środowisku Spark
Zmienna przewidywana	target_fields	
Predyktory	input_fields	
Lambda	lambda	lambda
Liczba procesów roboczych	nWorkers	nWorkers
Liczba wątków	numThreadPerTask	numThreadPerTask

Tabela 42. Właściwości węzła odwzorowane na parametry Spark (kontynuacja)

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr XGBoost w środowisku Spark
Używaj pamięci zewnętrznej	useExternalMemory	useExternalMemory
Typ wzmocnienia	boosterType	boosterType
Liczba rund boostingu	numBoostRound	round
Skala wag dodatnich	scalePosWeight	scalePosWeight
Cel	objectiveType	objective
Metryka ewaluacyjna	evalMetric	evalMetric
Lambda	lambda	lambda
Alfa	alpha	alpha
Obciążenie Lambda	lambdaBias	lambdaBias
Metoda drzewa	treeMethod	treeMethod
Maksymalna głębokość	maxDepth	maxDepth
Maks. waga elementu podrzędnego	minChildWeight	minChildWeight
Maks. krok zmiany	maxDeltaStep	maxDeltaStep
Podpróba	sampleSize	sampleSize
Eta	eta	eta
Gamma	gamma	gamma
Próba z kolumn na każde drzewo	colsSampleRation	colSampleByTree
Próba z kolumn na każdy poziom	colsSampleLevel	colsSampleLevel
Algorytm normalizacji	normalizeType	normalizeType
Algorytm próbkowania	sampleType	sampleType
Współczynnik wypadania	rateDrop	rateDrop
Prawdopodobieństwo pominiętego wypadnięcia	skipDrop	skipDrop

[1] "Scalable and Flexible Gradient Boosting." WWW. © 2015-2016 DMLC.

[2] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. WWW. © 2015-2016 DMLC.

[3] "ml.dmlc.xgboost4j.scala.spark Params." *DMLC for Scalable and Reliable Machine Learning*. WWW. 3 października 2017 r.

Węzeł XGBoost-AS — Opcje modelu

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Węzeł K-średnie-AS

K-średnie to jeden z najpowszechniej używanych algorytmów grupowania. Grupuje on punkty danych w określoną z góry liczbę grup.[1] Węzeł K-średnie-AS w programie SPSS Modeler jest zaimplementowany w środowisku Spark.

Aby uzyskać szczegółowe informacje na temat algorytmów K-średnich, patrz https://spark.apache.org/docs/2.2.0/ml-clustering.html.

Należy zwrócić uwagę, że węzeł K-średnie-AS automatycznie wykonuje kodowanie one-hot (kodowanie z gorącą jedynką) dla zmiennych kategorialnych.

[1] "Clustering." *Apache Spark*. MLlib: Main Guide. WWW. 3 października 2017 r.

Węzeł K-średnie-AS — Zmienne

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Użyj wstępnie zdefiniowanych ról. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej z wcześniejszego węzła Typ. Ta opcja jest wybrana domyślnie.

Użyj niestandardowych przypisań. Aby ręcznie przypisać zmienne wejściowe, wybierz tę opcję, a następnie zmienną lub zmienne wejściowe. Działanie tej opcji jest podobne, jak ustawienie roli zmiennej na **Zmienna wejściowa** w węźle Typy.

Węzeł K-średnie-AS — Opcje budowania

Na karcie Opcje budowania można określić opcje budowania dla węzła K-średnie-AS, w tym zwykłe opcje budowania modelu, opcje inicjowania środków grup i opcje zaawansowane dotyczące iteracji obliczeń i wartości startowej generatora liczb losowych. Aby uzyskać więcej informacji, patrz JavaDoc for K-Means on SparkML.[1]

Regular

Nazwa modelu. Nazwa zmiennej generowanej po ocenie z przypisaniem do określonej grupy. Wybierz opcję **Automatycznie** (ustawienie domyślne) lub wybierz opcję **Użytkownika** i wpisz nazwę.

Liczba grup. Określ liczbę skupień do wygenerowania. Wartość domyślna to **5**, a wartość minimalna to **2**.

Inicjowanie

Tryb inicjowania. Określ metodę inicjowania środków grup. Ustawienie domyślne to **K-średnie||**. Aby uzyskać szczegółowe informacje o obu metodach, patrz Scalable K-Means++. [2]

Kroki inicjowania. Jeśli wybrany jest tryb inicjowania **K-średnie||**, określ liczbę kroków inicjowania. Wartość domyślna to **2**.

Zaawansowane

Ustawienia zaawansowane. Wybierz tę opcję, jeśli chcesz ustawić poniższe opcje zaawansowane.

Maks. liczba iteracji. Określ, maksymalną liczbę iteracji, jaka ma być wykonywana podczas poszukiwania środków grup. Wartość domyślna to **20**.

Tolerancja. Określ tolerancję zbieżności dla algorytmów iteracyjnych. Wartość domyślna to **1.0E-4**.

Ustaw wartość startową generatora liczb losowych. Wybierz tę opcję i kliknij przycisk **Utwórz**, aby wygenerować wartość początkową dla generatora liczb losowych.

Wyświetl

Wyświetl wykres. Wybierz tę opcję, jeśli wyniki mają zawierać wykres.

W poniższej tabeli przedstawiono relację między ustawieniami w oknie dialogowym węzła K-średnie-AS w programie SPSS Modeler a parametrami algorytmu K-średnie w środowisku Spark.

Tabela 43. Właściwości węzła odwzorowane na parametry Spark

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr SparkML algorytmu K-średnie
Zmienne wejściowe	właściwości	
Liczba skupień	clustersNum	k
Tryb inicjowania	initMode	initMode
Kroki inicjowania	initSteps	initSteps
Maks. liczba iteracji	maxIter	maxIter
Tolerancja	toleration	tol
Losowa wartość początkowa	randomSeed	seed

[1] "Class KMeans." *Apache Spark*. JavaDoc. WWW. 3 października 2017 r.

[2] Bahmani, Moseley, et al. "Scalable K-Means++." 28 września 2012 r. http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf.

Uwagi

Niniejsza publikacja została przygotowana z myślą o produktach i usługach oferowanych w Stanach Zjednoczonych. Materiał ten jest również dostępny w IBM w innych językach. Jednakże w celu uzyskania dostępu do takiego materiału istnieje konieczność posiadania egzemplarza produktu w takim języku.

Produktów, usług lub opcji opisywanych w tym dokumencie IBM nie musi oferować we wszystkich krajach. Informacje o produktach i usługach dostępnych w danym kraju można uzyskać od lokalnego przedstawiciela IBM. Odwołanie do produktu, programu lub usługi IBM nie oznacza, że można użyć wyłącznie tego produktu, programu lub usługi IBM. Zamiast nich można zastosować ich odpowiednik funkcjonalny pod warunkiem, że nie narusza to praw własności intelektualnej IBM. Jednakże cała odpowiedzialność za ocenę przydatności i sprawdzenie działania produktu, programu lub usługi pochodzących od producenta innego niż IBM spoczywa na użytkowniku.

IBM może posiadać patenty lub złożone wnioski patentowe na towary i usługi, o których mowa w niniejszej publikacji. Przedstawienie niniejszej publikacji nie daje żadnych uprawnień licencyjnych do tychże patentów. Pisemne zapytania w sprawie licencji można przesyłać na adres:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

Zapytania dotyczące zestawów znaków dwubajtowych (DBCS) należy kierować do lokalnych działów własności intelektualnej IBM (IBM Intellectual Property Department) lub wysłać je na piśmie na adres:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan, Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokio 103-8510, Japonia

INTERNATIONAL BUSINESS MACHINES CORPORATION DOSTARCZA TĘ PUBLIKACJĘ W STANIE, W JAKIM SIĘ ZNAJDUJE ("AS IS") BEZ UDZIELANIA JAKICHKOLWIEK GWARANCJI (RĘKOJMIĘ RÓWNIEŻ WYŁĄCZA SIĘ), WYRAŹNYCH LUB DOMNIEMANYCH, A W SZCZEGÓLNOŚCI DOMNIEMANYCH GWARANCJI PRZYDATNOŚCI HANDLOWEJ, PRZYDATNOŚCI DO OKREŚLONEGO CELU ORAZ GWARANCJI, ŻE PUBLIKACJA TA NIE NARUSZA PRAW OSÓB TRZECICH. Ustawodawstwa niektórych krajów nie dopuszczają zastrzeżeń dotyczących gwarancji wyraźnych lub domniemanych w odniesieniu do pewnych transakcji; w takiej sytuacji powyższe zdanie nie ma zastosowania.

Informacje zawarte w niniejszej publikacji mogą zawierać nieścisłości techniczne lub błędy drukarskie. Informacje te są okresowo aktualizowane, a zmiany te zostaną uwzględnione w kolejnych wydaniach tej publikacji. IBM zastrzega sobie prawo do wprowadzania ulepszeń i/lub zmian w produktach i/lub programach opisanych w tej publikacji w dowolnym czasie, bez wcześniejszego powiadomienia.

Wszelkie wzmianki w tej publikacji na temat stron internetowych innych podmiotów zostały wprowadzone wyłącznie dla wygody użytkownika i w żadnym wypadku nie stanowią zachęty do ich odwiedzania. Materiały dostępne na tych stronach nie są częścią materiałów opracowanych dla tego produktu IBM, a użytkownik korzysta z nich na własną odpowiedzialność.

IBM ma prawo do używania i rozpowszechniania informacji przysłanych przez użytkownika w dowolny sposób, jaki uzna za właściwy, bez żadnych zobowiązań wobec ich autora.

Licencjobiorcy tego programu, którzy chcieliby uzyskać informacje na temat programu w celu: (i) wdrożenia wymiany informacji między niezależnie utworzonymi programami i innymi programami (łącznie z tym opisywanym) oraz (ii) wspólnego wykorzystywania wymienianych informacji, powinni skontaktować się z:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

Informacje takie mogą być udostępnione, o ile spełnione zostaną odpowiednie warunki, w tym, w niektórych przypadkach, zostanie uiszczona stosowna opłata.

Licencjonowany program opisany w niniejszej publikacji oraz wszystkie inne licencjonowane materiały dostępne dla tego programu są dostarczane przez IBM na warunkach określonych w Umowie IBM z Klientem, Międzynarodowej Umowie Licencyjnej IBM na Program lub w innych podobnych umowach zawartych między IBM i użytkownikami.

Dane dotyczące wydajności i cytowane przykłady zostały przedstawione jedynie w celu zobrazowania sytuacji. Faktyczne wyniki dotyczące wydajności mogą się różnić w zależności do konkretnych warunków konfiguracyjnych i operacyjnych.

Informacje dotyczące produktów innych podmiotów niż IBM zostały uzyskane od dostawców tych produktów, z ich publicznych ogłoszeń lub innych dostępnych publicznie źródeł. IBM nie testował tych produktów i nie może potwierdzić dokładności pomiarów wydajności, kompatybilności ani żadnych innych danych związanych z tymi produktami. Pytania dotyczące możliwości produktów innych podmiotów należy kierować do dostawców tych produktów.

Wszelkie stwierdzenia dotyczące przyszłych kierunków rozwoju i zamierzeń IBM mogą zostać zmienione lub wycofane bez powiadomienia.

Publikacja ta zawiera przykładowe dane i raporty używane w codziennej pracy. W celu kompleksowego ich zilustrowania, podane przykłady zawierają nazwiska osób prywatnych, nazwy przedsiębiorstw oraz nazwy produktów. Wszystkie te nazwy/nazwiska są fikcyjne i jakiekolwiek podobieństwo do istniejących nazw/nazwisk jest całkowicie przypadkowe.

Znaki towarowe

IBM, logo IBM i ibm.com są znakami towarowymi lub zastrzeżonymi znakami towarowymi International Business Machines Corp. zarejestrowanymi w wielu systemach prawnych na całym świecie. Pozostałe nazwy produktów i usług mogą być znakami towarowymi IBM lub innych przedsiębiorstw. Aktualna lista znaków towarowych IBM dostępna jest w serwisie WWW IBM, w sekcji "Copyright and trademark information" Informacje o prawach autorskich i znakach towarowych), pod adresem www.ibm.com/legal/copytrade.shtml.

Adobe, logo Adobe, PostScript oraz logo PostScript są znakami towarowymi lub zastrzeżonymi znakami towarowymi Adobe Systems Incorporated w Stanach Zjednoczonych i/lub w innych krajach.

Intel, logo Intel, Intel Inside, logo Intel Inside, Intel Centrino, logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium i Pentium są znakami towarowymi lub zastrzeżonymi znakami towarowymi Intel Corporation lub przedsiębiorstw podporządkowanych Intel Corporation w Stanach Zjednoczonych i w innych krajach.

Linux jest zastrzeżonym znakiem towarowym Linusa Torvaldsa w Stanach Zjednoczonych i/lub w innych krajach.

Microsoft, Windows, Windows NT oraz logo Windows są znakami towarowymi Microsoft Corporation w Stanach Zjednoczonych i/lub w innych krajach.

UNIX jest zastrzeżonym znakiem towarowym The Open Group w Stanach Zjednoczonych i/lub w innych krajach.

Java oraz wszystkie znaki towarowe i logo dotyczące języka Java są znakami towarowymi lub zastrzeżonymi znakami towarowymi Oracle i/lub przedsiębiorstw afiliowanych.

Warunki dotyczące dokumentacji produktu

Zezwolenie na korzystanie z tych publikacji jest przyznawane na poniższych warunkach.

Zakres stosowania

Niniejsze warunki stanowią uzupełnienie warunków używania serwisu WWW IBM.

Użytek osobisty

Użytkownik ma prawo kopiować te publikacje do własnego, niekomercyjnego użytku pod warunkiem zachowania wszelkich uwag dotyczących praw własności. Użytkownik nie ma prawa dystrybuować ani wyświetlać tych publikacji czy ich części, ani też wykonywać na ich podstawie prac pochodnych bez wyraźnej zgody IBM.

Użytek służbowy

Użytkownik ma prawo kopiować te publikacje, dystrybuować je i wyświetlać wyłącznie w ramach przedsiębiorstwa Użytkownika pod warunkiem zachowania wszelkich uwag dotyczących praw własności. Użytkownik nie ma prawa wykonywać na podstawie tych publikacji ani ich fragmentów prac pochodnych, kopiować ich, dystrybuować ani wyświetlać poza przedsiębiorstwem Użytkownika bez wyraźnej zgody IBM.

Prawa

Z wyjątkiem zezwoleń wyraźnie udzielonych w niniejszym dokumencie, nie udziela się jakichkolwiek innych zezwoleń, licencji ani praw, wyraźnych czy domniemanych, odnoszących się do tych publikacji czy jakichkolwiek informacji, danych, oprogramowania lub innej własności intelektualnej, o których mowa w niniejszym dokumencie.

IBM zastrzega sobie prawo do anulowania zezwolenia przyznanego w niniejszym dokumencie w każdej sytuacji, gdy, według uznania IBM, korzystanie z tych publikacji jest szkodliwe dla IBM lub jeśli IBM uzna, że warunki niniejszego dokumentu nie są przestrzegane.

Użytkownik ma prawo pobierać, eksportować lub reeksportować niniejsze informacje pod warunkiem zachowania bezwzględnej i pełnej zgodności z obowiązującym prawem i przepisami, w tym ze wszelkimi prawami i przepisami eksportowymi Stanów Zjednoczonych.

IBM NIE UDZIELA JAKICHKOLWIEK GWARANCJI, W TYM TAKŻE RĘKOJMI, DOTYCZĄCYCH TREŚCI TYCH PUBLIKACJI. PUBLIKACJE TE SĄ DOSTARCZANE W STANIE, W JAKIM SIĘ ZNAJDUJĄ ("AS-IS") BEZ UDZIELANIA JAKICHKOLWIEK GWARANCJI (RĘKOJMIĘ RÓWNIEŻ WYŁĄCZA SIĘ), WYRAŹNYCH CZY DOMNIEMANYCH, A W SZCZEGÓLNOŚCI DOMNIEMANYCH GWARANCJI PRZYDATNOŚCI HANDLOWEJ, PRZYDATNOŚCI DO OKREŚLONEGO CELU CZY NIENARUSZANIA PRAW OSÓB TRZECICH.

Glosariusz

A

AICC . Miara wybierania i porównywania modeli mieszanych tworzonych na podstawie -2 logarytmu wiarygodności (ograniczonego). Mniejsze wartości oznaczają lepszy model. Wartość AICC „poprawia" wartość AIC w przypadku małych prób. Przy wzroście wielkości próby wartość AICC zbiega do wartości AIC.

ANOVA dla każdej zmiennej . Dla każdej zmiennej niezależnej wykonuje test istotności różnic między średnimi grupowymi metodą jednoczynnikowej analizy wariancji.

B

Bayesowskie kryterium informacyjne (BIC) . Miara wybierania i porównywania modeli mieszanych tworzonych na podstawie -2 logarytmu wiarygodności. Mniejsze wartości oznaczają lepszy model. Wartość BIC „karze" także modele przeparametryzowane (na przykład złożone modele z dużą liczbą danych wejściowych), jednak silniej niż miara AIC.

Błąd standardowy . Miara tego, jak bardzo wartość statystyki testowej (sprawdzianu testu) zmienia się pomiędzy próbami. Jest to odchylenie standardowe rozkładu wartości danej statystyki dla poszczególnych prób. Na przykład błąd standardowy średniej to odchylenie standardowe średnich z prób.

Błąd standardowy kurtozy . Iloraz kurtozy i jej błędu standardowego; może być wykorzystywany jako test normalności (tzn. można odrzucić założenie normalności, jeśli wartość ilorazu jest mniejsza od -2 lub większa od +2). Wysoka dodatnia wartość dla kurtozy wskazuje na to, iż krańce rozkładu są dłuższe niż te dla rozkładu normalnego; ujemna wartość dla kurtozy wskazuje na krótsze ogony (podobnie jak w rozkładach prostokątnych).

Błąd standardowy skośności . Iloraz skośności i jej błędu standardowego; może być wykorzystywany jako test normalności (tzn. można odrzucić założenie o normalności, jeśli wartość ilorazu jest mniejsza od -2 lub większa od +2). Wysoka dodatnia wartość dla skośności wskazuje na długi prawy kraniec; skrajnie ujemna wartość wskazuje na długi lewy kraniec.

Błąd standardowy średniej . Miara tego, jak bardzo wartość średniej może się zmieniać dla różnych prób losowanych z tego samego rozkładu. Może być wykorzystywana do pobieżnego porównania rzeczywistej wartości średniej z wartością hipotetyczną (tj. można sądzić, że te dwie wartości są różne, jeśli iloraz różnicy i błędu standardowego jest mniejszy od -2 lub większy od +2).

D

Dominanta . Wartość występująca najczęściej. Jeśli więcej, niż jedna wartość występuje z taką samą, największą częstością, każda z nich jest dominantą (wartością modalną).

F

Fishera . Wyświetla współczynniki funkcji klasyfikacyjnej Fishera, które mogą być bezpośrednio używane do klasyfikowania. Dla każdej grupy otrzymywany jest oddzielny zestaw współczynników funkcji klasyfikacji, a przypadek klasyfikuje się do tej grupy, dla której ma najwyższą ocenę dyskryminacyjną (wartość funkcji klasyfikacji).

J

Jeden minus przeżycie . Umożliwia wykreślenie funkcji Jeden minus przeżycie na skali liniowej.

K

Klasyfikacja typu pozostaw-jedną-poza . Każda analizowana obserwacja jest klasyfikowana przez funkcję wyprowadzoną w oparciu o wszystkie pozostałe obserwacje z wyłączeniem tej jednej. Znana również jako „metoda U".

Korelacja wewnątrzgrupowa . Wyświetla macierz sumarycznych (połączonych) korelacji wewnątrzgrupowych, uzyskiwaną przez uśrednienie macierzy kowariancji dla wszystkich grup przed obliczeniem korelacji.

Kowariancja . Nieustandaryzowana miara powiązania dwóch zmiennych, równa sumie iloczynów wektorowych odchyleń wartości tych zmiennych od ich średnich podzielonej przez N-1.

Kowariancja całkowita . Wyświetla macierz kowariancji obliczanych na podstawie wszystkich obserwacji, tak jakby pochodziły z jednej próby.

Kowariancja wewnątrzgrupowa . Wyświetla macierz sumarycznych (połączonych) kowariancji wewnątrzgrupowych, która może być różna od całkowitej macierzy kowariancji. Macierz jest uzyskiwana przez uśrednienie poszczególnych macierzy kowariancji dla wszystkich grup.

Kowariancje dla odrębnych grup . Wyświetla osobne macierze kowariancji dla każdej grupy.

Kurtoza . Miara ilości skrajnych wartości odstających. W przypadku rozkładu normalnego wartość statystyki kurtozy wynosi zero. Kurtoza dodatnia oznacza, że w danych jest więcej skrajnych wartości odstających niż w rozkładzie normalnym. Kurtoza ujemna oznacza, że w danych jest mniej skrajnych wartości odstających niż w rozkładzie normalnym.

M

MaksBB . Maksymalny błąd bezwzględny. Największy prognozowany błąd, wyrażony w jednostkach używanych przez szereg. Podobnie jak w przypadku MaksPBB, błąd ten jest pomocny w wyobrażaniu najgorszego możliwego scenariusza prognozy. Maksymalny błąd bezwzględny oraz maksymalny procentowy błąd bezwzględny mogą wystąpić w różnych miejscach szeregu, np. gdy błąd bezwzględny dużego szeregu jest w niewielkim stopniu większy od błędu bezwzględnego wartości małego szeregu. W takim wypadku maksymalny błąd bezwzględny wystąpi dla wartości dużego szeregu, a maksymalny procentowy błąd bezwzględny wystąpi dla wartości małego szeregu.

Maksimum . Największa wartość zmiennej numerycznej.

MaksPBB . Maksymalny procentowy błąd bezwzględny. Największy prognozowany błąd, wyrażony jako procent. Miara ta jest pomocna w wyobrażaniu najgorszego możliwego scenariusza prognozy.

Maksymalizacja najmniejszego ilorazu F — metoda wprowadzania . Metoda doboru zmiennych przy analizie metodą krokową, oparta na maksymalizacji ilorazu F, obliczanego na podstawie odległości Mahalanobisa pomiędzy grupami.

Mapa terytorialna . Oparty o wartości funkcji dyskryminacyjnej wykres granic, wykorzystany do klasyfikowania obserwacji do grup. Liczby odpowiadają grupom, do których zostały zaklasyfikowane poszczególne obserwacje. Średnie dla kolejnych grup są na wykresie oznaczone gwiazdkami, które znajdują się wewnątrz granic określonych dla tych grup. Mapa nie zostaje wyświetlona wtedy, gdy istnieje tylko jedna funkcja dyskryminacyjna.

MAPE . Bezwzględny procentowy błąd średniej. Mierzy, jak bardzo szereg zależny odbiega od poziomu przewidywanego przez model. Jest niezależny od używanych jednostek i tym samym może być używany do porównywania szeregów używających różnych jednostek.

Mediana . Jest to 50. percentyl, czyli taka wartość, że połowa obserwacji ma wartości mniejsze, a druga połowa ma wartości większe od niej. W sytuacji parzystej liczby obserwacji mediana jest średnią dwóch środkowych obserwacji w próbie posortowanej rosnąco lub malejąco. W przeciwieństwie do średniej, na którą wpływ może mieć nawet kilka ekstremalnie dużych lub małych wartości, mediana jest miarą tendencji centralnej niewrażliwą na wartości odstające).

Minimalizacja lambdy Wilksa . Metoda doboru zmiennych w krokowej analizie dyskryminacyjnej, przy której wybierane są takie zmienne, które po wprowadzeniu do równania najbardziej zmniejszą współczynnik lambda Wilksa. W każdym kolejnym kroku procedury wprowadzona zostaje ta zmienna, która minimalizuje wartość tego współczynnika.

Minimum . Najmniejsza wartość zmiennej numerycznej.

N

Niestandaryzowane . Wyświetla niestandaryzowane współczynniki funkcji dyskryminacyjnej.

O

Obserwacji . Dla każdej wyświetlane są kody rzeczywistej grupy, przewidywanej grupy, prawdopodobieństw a posteriori i ocen dyskryminacyjnych.

Odchylenie standardowe . Miara rozproszenia wokół wartości średniej, równa pierwiastkowi z wariancji. Odchylenie standardowe mierzy się w tych samych jednostkach co pierwotną wartość.

Odchylenie standardowe . Miara rozproszenia wokół średniej. W przypadku rozkładu normalnego, 68% obserwacji znajduje się w obszarze oddalonym o jedno odchylenie standardowe od średniej, zaś 95% — w przedziale oddalonym o dwa odchylenia standardowe. Na przykład, jeśli średnia wieku osób wynosi 45 lat, a odchylenie standardowe wynosi 10, wówczas 95% rozważanych osób znajduje się w przedziale wiekowym między 25 a 65 lat.

Odległość Mahalanobisa . Miara stopnia, w jakim wartości zmiennych niezależnych dla danej obserwacji różnią się od wartości przeciętnej dla wszystkich obserwacji. Duże wartości wskaźnika Mahalanobisa oznaczają, że obserwacja zawiera skrajne wartości jednej albo większej liczby zmiennych niezależnych.

Odrębne grupy . Wykorzystuje do klasyfikacji macierze kowariancji dla poszczególnych grup. Ponieważ klasyfikacja oparta jest na funkcjach dyskryminacyjnych a nie na pierwotnych zmiennych, opcja ta nie zawsze jest równoważna dyskryminacji kwadratowej.

P

PPOK . Pierwiastek z przeciętnego odchylenia kwadratowego. Pierwiastek kwadratowy obliczany z przeciętnego odchylenia kwadratowego. Mierzy, jak bardzo szereg zależny odbiega od poziomu przewidywanego przez model; miara wyrażona w jednostkach używanych przez szereg zależny.

Przedział . Różnica między największą a najmniejszą wartością zmiennej numerycznej; maksimum minus minimum.

R

R-kwadrat . Miara dobroci dopasowania modelu liniowego, czasami nazywana współczynnikiem determinacji. Jest to część zmienności w zmiennej zależnej wyjaśniona przez model regresji. Przyjmuje wartości z przedziału od 0 do 1. Małe wartości statystyki wskazują na słabe dopasowanie modelu do danych.

S

Sekwencyjna Bonferroniego . Jest to sekwencyjne zstępująca odrzucająca procedura Bonferroniego, która jest mniej konserwatywna w zakresie odrzucania indywidualnych hipotez, ale zachowuje identyczny całościowy poziom istotności.

Sekwencyjna Sidaka . Jest to sekwencyjnie zstępująca odrzucająca procedura Bonferroniego, która jest mniej konserwatywna w zakresie odrzucania indywidualnych hipotez, ale zachowuje identyczny całościowy poziom istotności.

Skośność . Miara asymetrii rozkładu. Rozkład normalny jest symetryczny, a jego wartość skośności wynosi 0. Rozkład o dużej skośności dodatniej posiada długi kraniec z prawej strony. Gdy zaś współczynnik skośności jest ujemny, rozkład ma długi kraniec z lewej strony. Jako wytyczna, wartość skośności przekraczająca dwukrotnie swój błąd standardowy na ogół oznacza odstępstwo od symetrii rozkładu.

Stacjonarność r-kwadrat . Miara porównująca stacjonarną część modelu z modelem średniej prostej. Miara ta jest używana zamiast zwykłego r-kwadrat, gdy istnieje trend lub wzorzec sezonowy. Stacjonarna miara r-kwadrat może mieć wartość ujemną z przedziałem od minus nieskończoności do 1. Wartości ujemne oznaczają, że brany pod uwagę model jest gorszy niż model bazowy. Wartości dodatnie oznaczają, że brany pod uwagę model jest lepszy niż model bazowy.

Suma . Suma wartości wszystkich obserwacji nieposiadających braków danych.

Swoista . Oszacowuje wielkość wszystkich efektów równocześnie, korygując każdy z nich ze względu na wszystkie inne efekty każdego typu.

Ś

ŚBB . Średni błąd bezwzględny. Mierzy, jak bardzo szereg odbiega od poziomu przewidywanego przez model. ŚBB jest zgłaszany w pierwotnych jednostkach szeregu.

Średnia . Miara tendencji centralnej. Średnia arytmetyczna; suma podzielona przez liczbę obserwacji.

Średnie . Wyświetla średnią ogólną i średnie w grupach oraz odchylenia standardowe dla zmiennych niezależnych.

T

Test M Boxa . Test równości macierzy kowariancji grupowych. Przy odpowiednio dużych wielkościach prób nieistotna wartość p oznacza, że dowód nierówności macierzy jest niewystarczający. Test jest wrażliwy na odstępstwa od normalności rozkładu wielowymiarowego.

U

Użyj wartości F . Zmienna zostaje wprowadzona do modelu, jeśli wartość F jest większa niż określona wartość kryterium wprowadzenia, a zostaje wyłączona, jeśli wartość F jest mniejsza niż wartość przyjęta jako kryterium usunięcia. Wartość wprowadzenia musi być większa od wartości usunięcia i obie muszą być dodatnie. Chcąc wprowadzić więcej zmiennych do modelu, należy obniżyć wartość wprowadzenia. Chcąc usunąć więcej zmiennych, należy zwiększyć wartość usunięcia.

V

V Rao (analiza dyskryminacyjna) . Miara różnic między średnimi grupowymi. Znana jest także pod nazwą śladu Lawleya-Hotellinga. W każdym kolejnym kroku procedury wprowadzona zostaje ta zmienna, która powoduje największy wzrost wskaźnika V. Po wybraniu tej opcji wprowadź minimalną wartość, którą zmienna musi posiadać, aby została wprowadzona do analizy.

W

Wariancja . Miara rozproszenia wokół średniej, równa sumie podniesionych do kwadratu odchyleń od średniej, podzielonej przez liczbę obserwacji minus jeden. Wariancja jest mierzona w jednostkach będących kwadratami jednostek miary dla zmiennej, do której wariancja się odnosi.

Wariancja niewyjaśniona . W każdym kolejnym kroku analizy do modelu wprowadzana jest zmienna, która minimalizuje sumę niewyjaśnionej zmienności między grupami.

Ważne . Poprawne obserwacje nie posiadające systemowych ani zdefiniowanych przez użytkownika braków danych.

Wewnątrzgrupowe . Do klasyfikacji obserwacji wykorzystywana jest połączona macierz kowariancji wewnątrzgrupowych.

Wykres hazardu . Umożliwia wyświetlenie funkcji skumulowanego przeżycia na skali liniowej.

Wykres przeżycia . Umożliwia wyświetlenie funkcji skumulowanego przeżycia na skali liniowej.

Wykresy Odrębne grupy . Tworzy wykresy rozrzutu oddzielne dla każdej z grup, z uwzględnieniem pierwszych dwu funkcji dyskryminacyjnych. Jeśli istnieje tylko jedna funkcja, wyświetlone zostaną histogramy.

Wykresy Połączone grupy . Po zaznaczeniu tej opcji tworzony jest wykres rozrzutu wartości dwóch pierwszych funkcji dyskryminacyjnych, obejmujący wszystkie grupy. Jeśli istnieje tylko jedna funkcja, wyświetlany jest histogram.

Wyniki klasyfikacji . Liczba obserwacji prawidłowo i nieprawidłowo przypisanych do każdej grupy na podstawie analizy dyskryminacyjnej. Czasem zwana „Macierzą nieporozumień”.

Z

Zastosuj prawdopodobieństwo F. Zmienna zostaje wprowadzona do modelu, jeśli oszacowany dla niej poziom istotności dla wartości F jest mniejszy niż określona wartość kryterium wprowadzenia, a zostaje wyłączona, jeśli poziom istotności jest większy niż wartość przyjęta jako kryterium usunięcia. Wartość wprowadzenia musi być mniejsza od wartości usunięcia i obie muszą być dodatnie. Chcąc wprowadzić więcej zmiennych do modelu, należy zwiększyć wartość wprowadzenia. Aby usunąć więcej zmiennych, należy zmniejszyć wartość usunięcia.

Znormalizowane BKI. Znormalizowane Bayesowskie kryterium informacyjne. Ogólna miara całkowitego dopasowania modelu, która ma na celu uwzględnienie złożoności modelu. Jest to ocena oparta na średnim błędzie kwadratowym, zawierająca karę za liczbę parametrów w modelu oraz długość szeregu. Kara powoduje usunięcie przewagi modeli z większą liczbą parametrów, ułatwiając porównanie statystyki w różnych modelach tego samego szeregu.

Indeks

K

L

Ł

M

N

O

P

R

S

T

U

V

W

Z

Drukowane w USA