IBM SPSS Modeler — podręcznik CRISP-DM

Uwaga

Przed skorzystaniem z niniejszych informacji oraz produktu, którego one dotyczą, należy zapoznać się z informacjami zamieszczonymi w sekcji “Uwagi” na stronie 39.

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

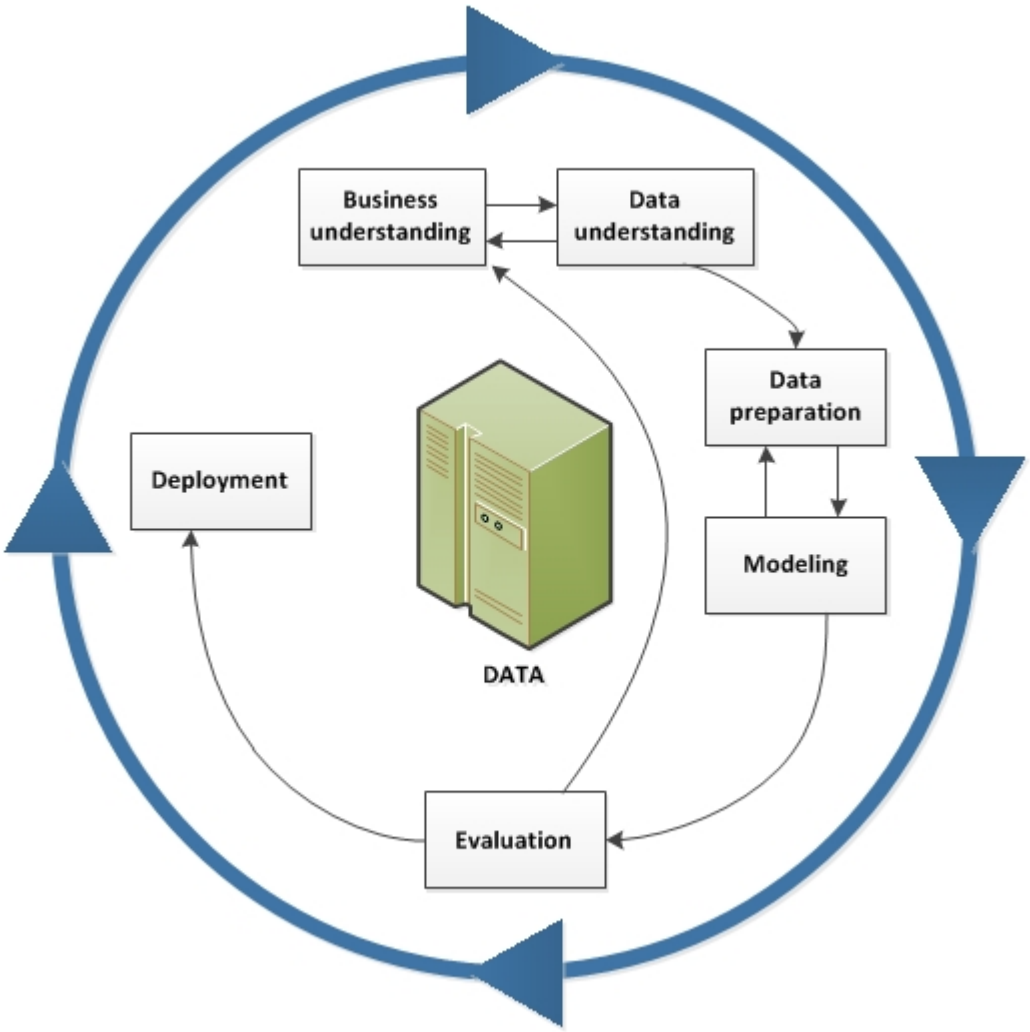
Wsparcie techniczne jest dostępne w celu zapewnienia klientom obsługi technicznej. Klienci mogą się kontaktować z działem Wsparcia technicznego w celu uzyskania pomocy dotyczącej korzystania z IBM Corp. produktów lub pomocy w instalacji dla jednego z obsługiwanych środowisk sprzętowych. Aby skontaktować 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. Wprowadzenie do metodologii CRISP-DM

Metodologia CRISP-DM — przegląd

CRISP-DM (Cross-Industry Standard Process for Data Mining) umożliwia eksplorację danych zgodnie ze sprawdzonymi procedurami branżowymi.

- CRISP-DM to **metodologia** udostępniająca opisy typowych faz projektu i zadań realizowanych w każdej fazie. Zawiera również wyjaśnienie związków pomiędzy tymi zadaniami.
- CRISP-DM to **model procesu** zawierający przegląd cyklu życia procesu eksploracji danych.



Rysunek 1. Cykl życia procesu eksploracji danych

Model cyklu życia składa się z sześciu faz ze strzałkami wskazującymi na najważniejsze i najczęściej występujące zależności pomiędzy fazami. Kolejność faz nie jest ściśle określona. W rzeczywistości w większości projektów następuje swobodne przechodzenie pomiędzy fazami w zależności od potrzeby.

Model CRISP-DM jest elastyczny i umożliwia łatwe dostosowywanie. Na przykład, jeśli w organizacji dąży się do wykrycia prania pieniędzy, wówczas przeprowadza się analizę dużych ilości danych bez konkretnego celu

modelowania. Działania skupiać się będą na eksploracji danych i wizualizacji zmierzającej do wykrycia podejrzanych schematów działań w zakresie danych finansowych, a nie na modelowaniu. CRISP-DM umożliwia tworzenie modelu eksploracji danych dostosowanego do określonych potrzeb.

W takiej sytuacji fazy modelowania, oceny i wdrożenia mogą mieć mniejsze znaczenie niż fazy interpretacji i przygotowania danych. Kwestie podniesione w tych późniejszych fazach należy jednak w dalszym ciągu uwzględniać w planowaniu długoterminowym oraz przyszłych celach eksploracji danych.

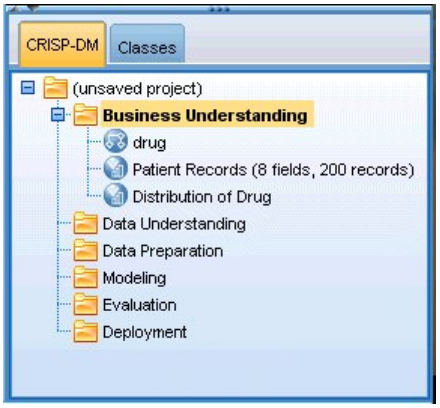
Metodologia CRISP-DM w programie IBM SPSS Modeler

Program IBM SPSS Modeler korzysta z metodologii CRISP-DM na dwa sposoby w celu zapewnienia unikalnego wsparcia służącego skutecznej eksploracji danych.

- Narzędzie projektowe CRISP-DM ułatwia organizowanie strumieni projektu, wyników i adnotacji zgodnie z fazami występującymi w typowym projekcie eksploracji danych. Raporty można tworzyć na dowolnym etapie projektu w oparciu o notatki dotyczące strumieni i faz CRISP-DM.
- Pomoc dotyczącą metodologii CRISP-DM ułatwia realizację projektu eksploracji danych. W systemie pomocy zawarte są listy zadań dla każdego kroku oraz przykłady działania CRISP-DM w warunkach rzeczywistych. Dostęp do pomocy CRISP-DM można uzyskać, wybierając opcję **Metodologia CRISP-DM** z okna głównego menu Pomoc.

Narzędzie projektowe CRISP-DM

Narzędzie projektowe CRISP-DM umożliwia zorganizowane podejście do eksploracji danych gwarantujące powodzenie projektu. Jest to zasadniczo rozszerzenie standardowego narzędzia projektowego IBM SPSS Modeler. Istnieje możliwość przełączania pomiędzy widokiem CRISP-DM a standardowym widokiem Klasy w celu wyświetlenia strumieni i wyników uporządkowanych według typu lub fazy CRISP-DM.



Rysunek 2. Narzędzie projektowe CRISP-DM

Korzystając z widoku CRISP-DM w narzędziu projektowym, można:

- Uporządkować strumienie i wyniki projektu zgodnie z fazami procesu eksploracji danych.
- Sporządzać notatki dotyczące celów organizacji dla każdej fazy.
- Tworzyć niestandardowe etykietki dla każdej fazy.
- Sporządzać notatki dotyczące wniosków wyciągniętych z określonego wykresu lub modelu.
- Generować lub aktualizować raport HTML przekazywany do zespołu projektowego.

Pomoc CRISP-DM

IBM SPSS Modeler udostępnia elektroniczny przewodnik po modelu procesów CRISP-DM. Model ten nie jest przypisany do konkretnego narzędzia. Struktura przewodnika odzwierciedla fazy projektu, a sam przewodnik zawiera następujące tematy:

- Przegląd i listę zadań w każdej fazie projektu CRISP-DM
- Pomoc w zakresie tworzenia raportów dotyczących różnych kamieni milowych

- Praktyczne przykłady przedstawiające sposób korzystania przez zespół projektowy z metodologii CRISP-DM w celu uproszczenia procesu eksploracji danych
- Łącza do dodatkowych zasobów dotyczących metodologii CRISP-DM

Dostęp do pomocy CRISP-DM można uzyskać, wybierając opcję **Metodologia CRISP-DM** z okna głównego menu Pomoc.

Dodatkowe zasoby

Oprócz pomocy dotyczącej CRISP-DM oferowanej w ramach programu IBM SPSS Modeler dostępnych jest kilka innych sposobów na lepsze zrozumienie procesów eksploracji danych.

- Podręcznik CRISP-DM opracowany przez konsorcjum CRISP-DM i dostarczony z tą wersją produktu.
- Podręcznik *Data Mining with Confidence*, copyright 2002 by SPSS Inc., ISBN 1-56827-287-1.

Rozdział 2. Zrozumienie biznesu

Zapoznanie z zadaniem — przegląd

Jeszcze przed rozpoczęciem pracy w programie IBM SPSS Modeler należy dokładnie przeanalizować oczekiwania organizacji wobec procesu eksploracji danych. Warto, aby w takiej analizie wzięło udział możliwie jak najwięcej kluczowych osób, a wyniki dyskusji należy dokładnie spisać. W końcowym etapie tej fazy CRISP-DM omówiony został sposób tworzenia planu projektu za pomocą zebranych w tym miejscu informacji.

Początkowo taka analiza może wydawać się zbędna. Jednak poznanie biznesowych przyczyn realizacji procesu eksploracji danych pozwala zapewnić zgodność w organizacji w zakresie wykorzystywania cennych zasobów.

Określenie celów biznesowych

W pierwszej kolejności należy zgromadzić możliwie jak najwięcej informacji dotyczących celów biznesowych, które zostaną wykorzystane w procesie eksploracji danych. Zadanie może okazać się znacznie trudniejsze, niż zakładano. Zdefiniowanie problemów, celów i zasobów może jednak zmniejszyć ryzyko występujące na późniejszych etapach projektu.

Metodologia CRISP-DM pozwala na realizację tego zadania w usystematyzowany sposób.

Lista zadań

- W pierwszej kolejności należy zgromadzić ogólne informacje dotyczące bieżącej sytuacji biznesowej.
- Należy udokumentować konkretne cele biznesowe określone przez główne osoby odpowiedzialne za podejmowanie decyzji.
- Należy uzgodnić kryteria stosowane do określenia powodzenia procesu eksploracji danych z biznesowego punktu widzenia.

Przykład z dziedziny handlu internetowego — określanie celów biznesowych

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Wraz ze zwiększającą się liczbą transakcji internetowych liczący się internetowy sprzedawca komputerów/elektroniki zmaga się z coraz większą konkurencją ze strony nowych serwisów. Zdając sobie sprawę, że sklepy internetowe powstają szybciej, niż następuje migracja klientów do strefy handlu elektronicznego, firma musi znaleźć sposoby na utrzymanie rentowności pomimo rosnących kosztów pozyskania klientów. Jednym z proponowanych rozwiązań jest wzmacnianie istniejących relacji z klientami, tak aby zwiększyć wartość, jaką wnoszą istniejący klienci do firmy.

Takie badanie ukierunkowane jest wówczas na:

- Zwiększenie sprzedaży wiązanej poprzez proponowanie trafniejszych produktów.
- Zwiększenie lojalności klientów dzięki bardziej spersonalizowanej obsłudze.

O pomyślności badania będą świadczyć następujące czynniki:

- Sprzedaży wiązana wzrosła o 10%.
- Klienci spędzają więcej czasu i przeglądają więcej stron podczas jednej wizyty.
- Badanie kończy się zgodnie z harmonogramem oraz mieści się w zakładanym budżecie.

Gromadzenie informacji ogólnych

Lepsze poznanie sytuacji biznesowej organizacji pozwala na skuteczniejsze określenie:

- Dostępnych zasobów (kadrowych i materiałowych)
- Problemów
- Celów

Aby uzyskać odpowiedzi na pytania mające wpływ na wyniki projektu eksploracji danych, należy przeprowadzić analizę bieżącej sytuacji biznesowej organizacji.

Zadanie 1 — Określanie struktury organizacyjnej

- Należy opracować schematy organizacyjne prezentujące piony, działy, departamenty oraz grupy projektowe w organizacji. Należy uwzględnić nazwiska kierowników oraz przypisane im obowiązki.
- Należy podać kluczowe osoby w organizacji.
- Należy określić wewnętrzny sektor odpowiedzialny za finansowanie i/lub wsparcie merytoryczne.
- Należy określić, czy będzie funkcjonować komitet sterujący, oraz podać listę jego członków.
- Należy zidentyfikować jednostki biznesowe, na które będzie miał wpływ projekt eksploracji danych.

Zadanie 2 — Opisywanie problematycznego obszaru

- Należy zidentyfikować problematyczny obszar, na przykład marketing, obsługa klienta lub rozwój działalności.
- Należy ogólnie opisać problem.
- Należy wyjaśnić wstępne wymagania projektu. Jakie są przesłanki do realizacji tego projektu? Czy piony biznesowe korzystają już z procesu eksploracji danych?
- Należy sprawdzić status projektu eksploracji danych w grupie biznesowej. Czy uzyskano ogólną aprobatę dla projektu? A może wystąpiła konieczność promowania eksploracji danych jako kluczowej technologii w grupie biznesowej?
- W razie konieczności należy przygotować i przeprowadzić w organizacji nieformalne prezentacje dotyczące eksploracji danych.

Zadanie 3 — Opisywanie bieżącego rozwiązania

- Należy opisać dowolne rozwiązania służące aktualnie do eliminowania problemów biznesowych.
- Należy opisać korzyści i wady bieżącego rozwiązania. Należy również określić poziom akceptacji tego rozwiązania w organizacji.

Definiowanie celów biznesowych

Jest to etap konkretnych działań. Po przeprowadzeniu analiz i spotkań należy określić konkretne cele. Muszą być one uzgodnione ze sponsorami projektu i innymi jednostkami biznesowymi, na które będą miały wpływ wyniki projektu. Tak określony cel, początkowo mało precyzyjny, np. "obniżenie poziomu odejścia", wkrótce zostanie przekuty na konkretny kierunek działań w zakresie eksploracji danych.

Lista zadań

Należy pamiętać o sporządzeniu notatek dotyczących następujących punktów. Takie informacje zostaną wykorzystanie na późniejszym etapie projektu. Cele powinny być realistyczne.

- Problem, który chcemy rozwiązać, należy opisać z wykorzystaniem procesu eksploracji danych.
- Wszystkie pytania biznesowe należy sprecyzować w największym możliwym stopniu.
- Należy określić inne wymogi biznesowe (np. utrzymanie istniejących klientów przy jednoczesnym wzroście możliwości sprzedaży wiązanej).
- Oczekiwane korzyści należy wyrazić językiem biznesowym (np. obniżenie poziomu odejścia najbardziej wartościowych klientów o 10%).

Kryteria sukcesu biznesowego

Wyznaczony cel może być bardzo precyzyjny, ale kiedy będzie można uznać, że został osiągnięty? Przed podjęciem dalszych kroków należy określić charakter sukcesu biznesowego, do którego dąży się w projekcie eksploracji danych. Istnieją dwie kategorie kryteriów sukcesu:

- **Obiektywne.** Kryteria te są bardzo przejrzyste, np. określony wzrost w zakresie rzetelności audytów lub zakładane obniżenie poziomu odejścia.
- **Subiektywne.** Trudno jest jednoznacznie stwierdzić występowanie kryteriów subiektywnych, np. „określenie serii skutecznych działań", ale można wyznaczyć osobę odpowiedzialną za podjęcie ostatecznej decyzji.

Lista zadań

- Kryteria sukcesu tego projektu należy dokumentować możliwie jak najdokładniej.
- Cel biznesowy powinien mieć skorelowane kryteria sukcesu.
- Należy przypisać arbitrów odpowiedzialnych za subiektywne pomiary sukcesu. W miarę możliwości należy zanotować ich oczekiwania.

Ocena sytuacji

Po precyzyjnym określeniu celu należy przeprowadzić ocenę bieżącej sytuacji. W tym kroku warto zadać następujące pytania:

- Jakie dane dostępne są do analizy?
- Czy organizacja dysponuje personelem wymaganym do realizacji projektu?
- Jakie są najważniejsze czynniki ryzyka?
- Czy dostępny jest plan interwencyjny dla każdego rodzaju ryzyka?

Przykład z dziedziny handlu internetowego — ocena sytuacji

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Internetowy sprzedawca elektroniki po raz pierwszy przystępuje do eksploracji sieci. Aby ułatwić sobie start, firma postanowiła skonsultować się ze specjalistą w zakresie eksploracji danych. Jednym z pierwszych zadań konsultanta jest ocena zasobów firmy wykorzystywanych do eksploracji danych.

Personel. Pracownicy dysponują specjalistyczną wiedzą w zakresie zarządzania dziennikami serwerów oraz bazami danych produktów i zakupów, jednak nie mają dużego doświadczenia w zakresie hurtowni danych oraz czyszczenia danych do analizy. Wskazane są zatem konsultacje ze specjalistą ds. bazy danych. Oczekuje się, że wyniki badania staną się częścią ciągłego procesu eksploracji sieci, dlatego też kierownictwo firmy musi rozważyć, czy stanowiska wypracowane podczas bieżących działań będą miały charakter trwały.

Dane. Firma działa już od dłuższego czasu, dlatego do analizy dostępna jest duża ilość danych z dzienników sieciowych i danych dotyczących sprzedaży. Na potrzeby tego badania firma ograniczy analizę wyłącznie do klientów zarejestrowanych w serwisie. W przypadku powodzenia projekt zostanie wdrożony w szerszym zakresie.

Ryzyka. Oprócz nakładów pieniężnych na konsultantów i czas poświęcony przez pracowników na badanie, przedsięwzięcie to nie stwarza poważnego bezpośredniego ryzyka. Czas jest jednak bardzo istotnym czynnikiem, dlatego początkowy projekt ma być realizowany w jednym kwartale finansowym.

Ponadto przepływy pieniężne nie są aktualnie imponujące, dlatego ważne jest, aby projekt nie przekroczył zakładanego budżetu. Jeśli którekolwiek z powyższych założeń zostanie zagrożone, zgodnie z sugestią kierowników zakres projektu powinien być ograniczony.

Inwentaryzacja zasobów

Przeprowadzenie dokładnej inwentaryzacji zasobów jest warunkiem koniecznym. Poprzez rzetelną analizę sprzętu, źródeł danych oraz zasobów ludzkich można zaoszczędzić wiele czasu i uniknąć potencjalnych problemów.

Zadanie 1 — Analiza zasobów sprzętowych

- Jakie jest wymagane oprogramowanie?

Zadanie 2 — Identyfikowanie źródeł danych i magazynów wiedzy

- Jakie źródła danych można wykorzystać w procesie eksploracji danych? Należy zanotować rodzaje i formaty danych.
- W jaki sposób przechowuje się dane? Czy mamy bezpośredni, dynamiczny dostęp do hurtowni danych i operacyjnych baz danych?
- Czy planowany jest zakup danych zewnętrznych, np. danych demograficznych?
- Czy istnieją kwestie bezpieczeństwa uniemożliwiające dostęp do wymaganych danych?

Zadanie 3 — Identyfikowanie zasobów kadrowych

- Czy można nawiązać kontakt ze specjalistami ds. biznesowych oraz w zakresie danych?
- Czy zidentyfikowano potrzebnych administratorów baz danych oraz inny personel wsparcia?

Po udzieleniu odpowiedzi na te pytania należy załączyć listę osób kontaktowych i pracowników, która zostanie wykorzystana w raporcie z tej fazy.

Wymogi, założenia i ograniczenia

Podejmowane wysiłki będą bardziej owocne, jeśli zostanie wykonana rzetelna ocena zobowiązań w projekcie. Wyraźnie zaznaczenie takich obaw pozwoli na uniknięcie problemów w przyszłości.

Zadanie 1 — Określanie wymagań

Podstawowym wymogiem jest omówiony wcześniej cel biznesowy. Należy jednak wziąć pod uwagę następujące zagadnienia:

- Czy w stosunku do wykorzystywanych danych bądź wyników projektu istnieją ograniczenia prawne lub związane z bezpieczeństwem?
- Czy wszystkie osoby znają wymogi dotyczące harmonogramu projektu?
- Czy istnieją wymogi dotyczące wdrażania wyników (np. publikowanie w serwisie lub wczytywanie wyników do bazy danych)?

Zadanie 2 — Wyjaśnianie założeń

- Czy istnieją czynniki ekonomiczne, które mogą mieć wpływ na projekt (np. opłaty z tytułu usług doradczych lub produkty konkurencyjne)?
- Czy istnieją założenia w zakresie jakości danych?
- W jaki sposób sponsor projektu/zespół kierowniczy będzie przeglądać wyniki? Innymi słowy, czy osoby te chcą zrozumieć cały model, czy wystarczy, że przejrzą wyniki?

Zadanie 3 — Weryfikacja ograniczeń

- Czy dostępne są wszystkie hasła wymagane, aby uzyskać dostęp do danych?
- Czy zostały zweryfikowane wszystkie ograniczenia prawne dotyczące wykorzystania danych?
- Czy w budżecie projektu zostały uwzględnione wszystkie ograniczenia finansowe?

Rodzaje ryzyka i nieprzewidziane zdarzenia

Podczas realizacji projektu warto rozważyć możliwe rodzaje ryzyka. Można wyróżnić następujące rodzaje ryzyka:

- Ryzyko planowania (Jakie będą konsekwencje przedłużającego się projektu?)
- Ryzyko finansowe (Jakie będą konsekwencje napotkania przez sponsorów trudności budżetowych?)
- Ryzyko związane z danymi (Jakie będą konsekwencje słabej jakości danych lub niewystarczającego zakresu danych?)

- Ryzyko związane z wynikami (Jakie będą konsekwencje uzyskania wyników gorszych od oczekiwanych?)

Po przeanalizowaniu różnych rodzajów ryzyka należy opracować plan awaryjny ułatwiający uniknięcie katastrofy.

Lista zadań
- Każde ewentualne ryzyko należy dobrze opisać.
- W planie awaryjnym należy opisać każdy rodzaj ryzyka.

Pojęcia

Aby zapewnić spójność terminologiczną w obszarze biznesu i eksploracji danych, należy opracować glosariusz terminów technicznych i haseł, które wymagają wyjaśnienia. Na przykład, jeśli termin "poziom odejścia" ma szczególne znaczenie w firmie, warto go jednoznacznie wyjaśnić. Będzie to z korzyścią dla całego zespołu. Pomocne może być również wyjaśnienie sposobu korzystania z wykresów korzyści.

Lista zadań
- Należy prowadzić listę terminów oraz żargonu stwarzającego problemy członkom zespołu. Glosariusz powinien uwzględniać zarówno terminy biznesowe, jak i z zakresu eksploracji danych.
- Glosariusz warto opublikować w intranecie lub uwzględnić w innej dokumentacji projektowej.

Analiza kosztów i korzyści

Ten etap pozwala na uzyskanie odpowiedzi na pytanie o **faktyczne korzyści projektu?** W ramach ostatecznej oceny kluczowe znaczenie na porównanie kosztów projektu z potencjalnymi korzyściami sukcesu.

Lista zadań

W analizie należy uwzględnić szacowane koszty:
- Gromadzenia danych oraz wykorzystania danych zewnętrznych
- Wdrożenia wyników
- Koszy operacyjne

Następnie należy uwzględnić korzyści:
- Zamierzonego głównego celu
- Dodatkowych spostrzeżeń zebranych podczas procesu eksploracji danych
- Możliwe korzyści wynikające z lepszego zrozumienia danych

Określanie celów eksploracji danych

Po określeniu celów biznesowych nadszedł czas na zinterpretowanie ich w kontekście eksploracji danych. Na przykład cel biznesowy "obniżenie poziomu odejścia" można zinterpretować jako cel eksploracji danych obejmujący:
- Identyfikowanie najbardziej wartościowych klientów w oparciu o ostatnie dane dotyczące zakupów
- Tworzenie modelu za pomocą dostępnych danych dotyczących klientów w celu przewidzenia prawdopodobieństwa odejścia każdego klienta
- Przypisanie każdemu klientowi punktacji w oparciu o skłonność do odejścia oraz wartość klienta

Wnioski z realizacji tych celów eksploracji danych można wykorzystać w działach biznesowych w celu obniżenia poziomu odejścia wśród najbardziej wartościowych klientów.

Okazuje się więc, że współpraca pomiędzy działami biznesu i technicznymi jest nieodzowna w skutecznej realizacji procesu eksploracji danych. W dalszej części znajdują się konkretne wskazówki dotyczące sposobu określania celów eksploracji danych.

Cele eksploracji danych

Współpracując z analitykami biznesowymi i analitykami danych w celu opracowania technicznego rozwiązania problemu biznesowego, należy pamiętać o konkretnym i precyzyjnym podejściu.

Lista zadań

- Należy opisać **rodzaj** problemu dotyczący eksploracji danych, np. grupowanie, predykcja lub klasyfikacja.
- Należy udokumentować cele techniczne za pomocą konkretnych jednostek czasu, np. predykcje z trzymiesięcznym okresem ważności.
- W miarę możliwości należy podać rzeczywiste liczby określające żądane cele, np. uzyskanie 80-procentowego poziomu odejścia wśród istniejących klientów.

Przykład z dziedziny handlu internetowego — cele eksploracji danych

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Korzystając z pomocy konsultanta ds. eksploracji danych, sprzedawca internetowy określił firmowe cele biznesowe za pomocą terminów z zakresu eksploracji danych. Cele początkowej analizy, które należy osiągnąć w tym kwartale, obejmują:

- Użycie informacji historycznych dotyczących wcześniejszych zakupów w celu utworzenia modelu łączącego pokrewne pozycje. Dla użytkowników korzystających z opisu pozycji należy podać łącza do innych pozycji w powiązanej grupie (**analiza koszyka zakupów**).
- Użycie dzienników sieciowych w celu określenia produktów, którymi klient jest zainteresowany na stronie, a następnie przeprojektowanie strony w oparciu o te pozycje. Strona główna serwisu będzie różniła się w zależności od rodzaju klienta (**profilowanie**).
- Użycie dzienników sieciowych w celu przewidzenia kolejnego kroku użytkownika, miejsca, z którego przeszedł do serwisu, lub ostatniej odwiedzanej strony w serwisie (**analiza kolejności**).

Kryteria sukcesu eksploracji danych

Sukces należy również zdefiniować w ujęciu technicznym, tak aby umożliwić sprawny przebieg procesu eksploracji danych. Aby zdefiniować wskaźniki sukcesu, warto skorzystać z określonego wcześniej celu eksploracji danych. IBM SPSS Modeler udostępnia takie narzędzia jak węzeł ewaluacji i węzeł analizy ułatwiające analizę dokładności i ważności uzyskanych wyników.

Lista zadań

- Należy opisać metody oceny jakości modelu (np. dokładność, wydajność itd.).
- Należy zdefiniować wskaźniki służące ocenie sukcesu. Należy podać konkretne liczby.
- Subiektywne pomiary należy zdefiniować w możliwie najlepszy sposób. Należy również wyznaczyć arbitra odpowiedzialnego za ocenę sukcesu.
- Należy rozważyć, czy pomyślne wdrożenie wyników modelu stanowi część sukcesu procesu eksploracji danych. Planowanie drożenia warto rozpocząć już na tym etapie.

Tworzenie planu projektu

Na tym etapie można płynnie przejść do tworzenia planu projektu eksploracji danych. Podstawą do opracowywania takiego planu będą zadane wcześniej pytania oraz sformułowane cele biznesowe i cele z zakresu eksploracji danych.

Opracowywanie planu projektu

Plan projektu to główny dokument dotyczący całości działań eksploracji danych. Prawidłowo opracowany plan projektu stanowi źródło informacji dotyczących projektu, np. celów, zasobów, rodzajów ryzyka czy harmonogramu etapów procesu eksploracji danych, dla wszystkich osób związanych z projektem. Plan wraz z dokumentacją zgromadzoną w tej fazie można opublikować w firmowym intranecie.

Lista zadań

Podczas tworzenia planu należy odpowiedzieć na następujące pytania:

- Czy zadania w projekcie i proponowany plan zostały omówiony ze wszystkimi zainteresowanymi osobami?
- Czy dla wszystkich faz lub zadań uwzględniono przewidywany czas trwania?
- Czy zostały uwzględnione wysiłki i zasoby wymagane do wdrożenia wyników lub rozwiązania biznesowego?
- Czy w planie uwzględniono punkty decyzyjne i wnioski o weryfikację/recenzję?
- Czy oznaczono fazy, w których zwykle pojawia się wiele iteracji, np. modelowanie?

Przykładowy plan projektu

Plan zawierający ogólny przegląd badania został przestawiony w poniższej tabeli.

Tabela 1. Przykładowy ogólny plan projektu

Faza	Czas	Zasoby	Ryzyka
Zapoznanie z zadaniem	1 tydzień	Wszyscy analitycy	Zmiana gospodarcza
Zrozumienie danych	3 tygodnie	Wszyscy analitycy	Problemy dotyczące danych, problemy technologiczne
Przygotowanie danych	5 tygodni	Konsultant ds. eksploracji danych, analityk baz danych w częściowym wymiarze	Problemy dotyczące danych, problemy technologiczne
Modelowanie	2 tygodnie	Konsultant ds. eksploracji danych, analityk baz danych w częściowym wymiarze	Problemy technologiczne, brak możliwości znalezienia odpowiedniego modelu
Ocena	1 tydzień	Wszyscy analitycy	Zmiana gospodarcza, brak możliwości wdrożenia wyników
Wdrożenie	1 tydzień	Konsultant ds. eksploracji danych, analityk baz danych w częściowym wymiarze	Zmiana gospodarcza, brak możliwości wdrożenia wyników

Ocena narzędzi i technik

Po wybraniu programu IBM SPSS Modeler jako narzędzia służącego do skutecznej eksploracji danych ten etap można wykorzystać w celu określenia, która technika eksploracji danych jest najlepiej dopasowana do potrzeb organizacji. Program IBM SPSS Modeler udostępnia całą gamę narzędzi wykorzystywanych podczas poszczególnych faz procesu eksploracji danych. Przed podjęciem decyzji o wyborze konkretnej techniki należy zapoznać się z sekcją dotyczącą modelowania w pomocy online.

Przed wykonaniem kolejnego kroku

Przed rozpoczęciem realizacji procesu eksploracji danych oraz korzystania z programu IBM SPSS Modeler należy odpowiedzieć na następujące pytania.

Z perspektywy biznesowej:

- Jakie korzyści organizacja zamierza osiągnąć poprzez realizację tego projektu?
- W jaki sposób zostanie zdefiniowany sukces podjętych działań?
- Czy dostępny jest budżet i zasoby wymagane do realizacji wyznaczonych celów?
- Czy istnieje dostęp do wszystkich danych wymaganych w tym projekcie?
- Czy w zespole zostały omówione ryzyka i nieprzewidziane zdarzenia związane z projektem?
- Czy wyniki analizy kosztów i korzyści potwierdzają opłacalność projektu?

Czy odpowiedzi na powyższe pytania zostały przekształcone na cele procesu eksploracji danych?

Z perspektywy eksploracji danych:

- W jaki sposób proces eksploracji danych może konkretnie wpłynąć na realizację celów biznesowych?
- Czy wiadomo w organizacji, które techniki eksploracji danych dają najlepsze wyniki?
- W jaki sposób zostanie wykonana ocena dokładności lub skuteczności wyników? (*Czy został ustalony sposób pomiaru sukcesu eksploracji danych?*)
- W jaki sposób zostaną wdrożone wyniki modelowania? Czy faza wdrożenia została uwzględniona w planie projektu?
- Czy plan projektu obejmuje wszystkie fazy metodologii CRISP-DM?
- Czy w planie zostały uwzględnione rodzaje ryzyka i nieoczekiwane zdarzenia?

Udzielenie odpowiedzi twierdzących na powyższe pytania oznacza gotowość do dokładniejszego zapoznania się z danymi.

Rozdział 3. Zrozumienie danych

Zrozumienie danych — przegląd

Faza zrozumienia danych w metodologii CRISP-DM obejmuje dokładną analizę danych podlegających eksploracji. Krok ten ma kluczowe znaczenie — pozwala zapobiec nieoczekiwanym problemom w kolejnej fazie, tj. przygotowywaniu danych, która jest zwykle najdłuższą częścią projektu.

Zrozumienie danych obejmuje uzyskanie dostępu do danych i ich eksplorację za pomocą tabel i elementów graficznych, które w programie IBM SPSS Modeler można uporządkować za pomocą narzędzia projektowego CRISP-DM. Takie podejście pozwoli na określenie jakości danych i opisanie wyników tych kroków w dokumentacji projektowej.

Gromadzenie danych początkowych

Na tym etapie metodologii CRISP-DM użytkownik może uzyskać dostęp do danych i przenieść je do programu IBM SPSS Modeler. Dane pochodzą z wielu źródeł, do których należą:

- **Istniejące dane.** Jest to cała gama danych, np. dane transakcyjne, dane ankietowe, dzienniki sieciowe itd. Należy ocenić, czy istniejące dane wystarczają do realizacji potrzeb organizacji.
- **Zakupione dane.** Czy w organizacji korzysta się z dodatkowych danych, np. danych demograficznych? Jeśli nie, należy zastanowić się, czy takie dane mogą być konieczne.
- **Dodatkowe dane.** Jeśli za pomocą danych z powyższych źródeł nie można zrealizować potrzeb organizacji, konieczne może być przeprowadzenie ankiet lub rozpoczęcie dodatkowych czynności śledzenia w celu uzupełnienia istniejących składnic danych.

Lista zadań

Czas, aby przyjrzeć się danym w programie IBM SPSS Modeler i odpowiedzieć na następujące pytania. Należy pamiętać o zanotowaniu wniosków. Więcej informacji można znaleźć w temacie “Tworzenie raportu z gromadzenia danych” na stronie 14.

- Które atrybuty (kolumny) z bazy danych wydają się najbardziej obiecujące?
- Które atrybuty wydają się nieistotne i można je wykluczyć?
- Czy dostępne dane wystarczają, aby wyciągnąć ogólne wnioski lub sformułować dokładne predykcje?
- Czy występuje zbyt dużo atrybutów dla wybranej metody modelowania?
- Czy łączone są różne źródła danych? Jeśli tak, czy występują obszary, które mogą stwarzać trudności podczas scalania?
- Czy sprawdzono, w jaki sposób braki danych obsługiwane są w poszczególnych źródłach danych?

Przykład z obszaru handlu internetowego — wstępne gromadzenie danych

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

W tym przykładzie sprzedawca internetowy korzysta z kilku ważnych źródeł danych uwzględniających:

Dzienniki sieciowe. Surowe dzienniki dostępu zawierają wszystkie informacje dotyczące sposobu poruszania się przez klientów w serwisie. W ramach procesu przygotowywania danych konieczne będzie usunięcie odniesień do plików graficznych oraz innych nieinformacyjnych pozycji w dziennikach sieciowych.

Dane dotyczące zakupów. Po przesłaniu zamówienia przez klienta zapisywane są wszystkie informacje związane z tym zamówieniem. Zamówienia w bazie danych muszą być odwzorowane na odpowiednią sesję w dziennikach WWW.

Produktowa baza danych. Atrybuty produktu mogą być użyteczne podczas identyfikowania powiązanych produktów. Informacje o produkcie należy zmapować do odpowiednich zamówień.

Baza danych klientów. Ta baza danych zawiera dodatkowe informacje zebrane od zarejestrowanych klientów. Rekordy nie są kompletne, ponieważ wielu klientów nie uzupełnia kwestionariuszy. Informacje o klientach należy zmapować do odpowiednich zakupów i sesji w dziennikach sieciowych.

Aktualnie firma nie planuje zakupu zewnętrznych baz danych ani wydania pieniędzy na przeprowadzenie ankiet, ponieważ firmowi analitycy są aktualnie zajęci zarządzaniem posiadanych danych. Niewykluczone, że w przyszłości firma będzie zainteresowana szerszym wdrożeniem wyników eksploracji danych. W takim przypadku przydatny będzie zakup dodatkowych danych demograficznych niezarejestrowanych klientów. Dane demograficzne mogą być również pomocne w analizie różnic pomiędzy danymi w firmowej bazie klientów a cechami przeciętnego klienta internetowego.

Tworzenie raportu z gromadzenia danych

Korzystając z informacji zebranych w poprzednim kroku, można przystąpić do tworzenia raportu z gromadzenia danych. Ukończony raport można dodać do serwisu projektu lub udostępnić zespołowi. Raport można również połączyć z raportami, które zostaną przygotowane w kolejnych krokach, tj. opisywanie, eksploracja i weryfikacja jakości danych. Raporty te nadają kierunek dalszym działaniom w fazie przygotowywania danych.

Opisywanie danych

Istnieje wiele sposobów opisywania danych. Większość opisów skupia się jednak na ilości i jakości danych (ile danych jest dostępnych) oraz na stanie danych. Poniższej zamieszczone zostały wybrane kluczowe cechy, o których należy pamiętać podczas opisywania danych.

- **Ilość danych.** W przypadku większości technik modelowania trzeba uwzględnić kompromisy związane z rozmiarem danych. Za pomocą dużych zbiorów danych można utworzyć dokładniejsze modele. Czas przetwarzania może być jednak w tym przypadku dłuższy. Warto zastanowić się wówczas nad użyciem podzbioru danych. Podczas robienia notatek do raportu końcowego należy pamiętać o uwzględnieniu statystyk rozmiaru dla wszystkich zbiorów danych. Ponadto na etapie opisywania danych należy wziąć pod uwagę liczbę rekordów oraz zmiennych (atrybutów).
- **Rodzaje wartości.** Dane mogą być prezentowane w wielu formatach, np. w formacie **numerycznym**, **jakościowym** (łańcuch) lub **logicznym** (prawda/fałsz). Zwrócenie uwagi na rodzaj wartości pomoże uniknąć problemów podczas dalszych faz modelowania.
- **Schematy kodowania.** Wartości w bazie danych często przedstawiają cechy charakterystyczne, np. płeć lub rodzaj produktu. Na przykład w jednym zbiorze danych symbole *M* i *F* mogą oznaczać kategorie *mężczyzna* i *kobieta*, natomiast w innym zbiorze mogą występować wartości numeryczne *1* i *2*. W raporcie dotyczącym danych należy zwrócić uwagę na wszystkie sprzeczne schematy.

Dotychczas zdobyta wiedza pozwoli na opracowanie raportu dotyczącego opisu danych i udostępnienie wniosków szerszemu kręgowi zainteresowanych.

Przykład z dziedziny handlu internetowego — opisywanie danych

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

W aplikacji do eksploracji sieci należy przetworzyć wiele rekordów i atrybutów. Chociaż sprzedawca internetowy realizujący projekt eksploracji danych ograniczył wstępną analizę do około 30 000 klientów zarejestrowanych w serwisie, to nadal w dziennikach sieciowych dostępne są miliony rekordów.

Większość danych z tych źródeł ma charakter symboliczny. Są to na przykład daty i godziny, odwiedzane strony lub odpowiedzi na pytania wielokrotnego wyboru z kwestionariusza rejestrowego. Wybrane zmienne zostaną wykorzystane do utworzenia nowych zmiennych numerycznych, np. liczby odwiedzanych stron i czasu spędzonego w serwisie. Kilka istniejących zmiennych numerycznych w źródłach danych uwzględnia numer każdego zamówionego produktu, ilość czasu poświęconego na zakup, wagę produktu oraz specyfikacje wymiaru z produktowej bazy danych.

Schematy kodowania dla różnych źródeł danych raczej nie zachodzą na siebie, ponieważ źródła danych zawierają różne atrybuty. Jedyne zmienne, które na siebie zachodzą, to tzw. klucze, czyli na przykład identyfikatory klienta lub kody produktów. Te zmienne muszą mieć takie same schematy kodowania w każdym źródle danych. W przeciwnym razie scalenie dwóch źródeł danych będzie niemożliwe. Aby ponownie zakodować te zmienne kluczowe na potrzeby scalania, konieczne będzie dodatkowe przygotowanie danych.

Opracowanie raportu dotyczącego opisu danych

Aby skutecznie realizować projekt eksploracji danych, warto opracować dokładny raport dotyczący opisu danych za pomocą następujących metryk:

Ilość danych

- Jaki jest format danych?
- Należy określić metodę przechwytywania danych, np. ODBC.
- Jaka jest wielkość bazy danych (liczba wierszy i kolumn)?

Jakość danych

- Czy dane zawierają cechy mające znaczenie w ujęciu biznesowym?
- Jakie występują rodzaje danych (symboliczne, numeryczne itd.)?
- Czy dla kluczowych atrybutów zostały wyliczone podstawowe statystyki? Jakie wnioski dla biznesu zostały wyciągnięte na tej podstawie?
- Czy można określić priorytety dla istotnych atrybutów? Jeśli nie, czy analitycy biznesowi mogą opracować głębszą analizę?

Eksplorowanie danych

Ta faza CRISP-DM umożliwia eksplorowanie danych znajdujących się w tabelach, wykresach i innych narzędziach wizualizacji wyników dostępnych w programie IBM SPSS Modeler. Takie analizy ułatwiają realizację celów eksploracji danych opracowanych podczas fazy zapoznania z zadaniem. Pomocne są również przy formułowaniu hipotez i określaniu zadań przekształcenia danych mających miejsce podczas przygotowywania danych.

Przykład z dziedziny handlu internetowego — eksplorowanie danych

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Zgodnie z metodologią CRISP-DM wstępna eksploracja powinna mieć miejsce na tym etapie. W opinii przywołanego sprzedawcy internetowego eksploracja danych jest jednak trudna, jeśli nie niemożliwa, przy użyciu surowych dzienników sieciowych. Dane z dzienników sieciowych muszą być zwykle przetworzone najpierw podczas fazy przygotowania danych. W ten sposób tworzone są dane, których eksploracja może przynieść oczekiwane rezultaty. To odejście od metodologii CRISP-DM podkreśla możliwość, a nawet konieczność dostosowania procesu do określonych potrzeb eksploracji danych. Metodologia CRISP-DM ma charakter cykliczny, a specjaliści ds. eksploracji danych zwykle płynnie przechodzą pomiędzy fazami.

Chociaż dzienniki sieciowe powinny być przetworzone przed eksploracją, do eksploracji lepiej nadają się inne źródła danych dostępne dla sprzedawcy internetowego. Korzystanie z bazy danych dotyczącej zakupów podczas eksploracji umożliwia uzyskanie interesujących podsumowań na temat klientów, np. ile wydają pieniędzy, ile artykułów kupują podczas jednorazowych zakupów oraz skąd pochodzą. Podsumowania oparte na bazie danych klientów dają wgląd w rozkład odpowiedzi na pytania zawarte w kwestionariuszu rejestrowym.

Eksploracja może również być pomocna przy wyszukiwaniu błędów w danych. Chociaż większość źródeł danych generowanych jest automatycznie, informacje w produktowej bazie danych zostały wprowadzone ręcznie. Wybrane podsumowania dotyczące wymienionych wymiarów produktu ułatwiają wykrycie literówek, np. monitor "119-calowy" (zamiast "19-calowy").

Opracowanie raportu z eksploracji danych

Podczas tworzenia wykresów i uruchamiania statystyk dla dostępnych danych warto rozpocząć formułowanie hipotez dotyczących tego, w jaki sposób dane posłużą do realizacji celów biznesowych i technicznych.

Lista zadań

Należy robić notatki, które zostaną później umieszczone w raporcie z eksploracji danych. Warto również odpowiedzieć na następujące pytania:

- Jakie zostały postawione hipotezy dotyczące danych?
- Jakie atrybuty nadają się do dalszej analizy?
- Czy badania pozwoliły na odkrycie nowych cech charakteryzujących dane?
- Czy te badania zmieniły wstępne hipotezy?
- Czy można zidentyfikować podzbiory danych do wykorzystania w przyszłości?
- Dokonaj przeglądu celów eksploracji danych. Czy te badania spowodowały zmianę celów?

Weryfikowanie jakości danych

Dane często nie są idealne. W rzeczywistości większość danych zawiera błędy kodowania, braki lub inne niespójności, które utrudniają wykonywanie analiz. Jednym ze sposobów na uniknięcie potencjalnych pułapek jest przeprowadzenie dokładnej analizy danych jeszcze przed modelowaniem.

Narzędzia raportowania IBM SPSS Modeler (np. Audyt danych, Tabela i inne węzły wyników) ułatwiają wyszukiwanie następujących problemów:

- Przez **brak danych** rozumie się puste wartości lub wartości kodowane jako brak odpowiedzi (np. *$null$*, *?* lub *999*).
- **Błędy danych** to zwykle błędy typograficzne popełnione podczas wprowadzania danych.
- **Błędy pomiarowe** uwzględniają dane wprowadzone poprawnie, które są jednak oparte na błędnym schemacie pomiarów.
- **Niespójności kodowania** uwzględniają zwykle niestandardowe jednostki miary lub niespójne dane, na przykład użycie litery *M* i słowa *męska* do określenia płci.
- **Niewłaściwe metadane** uwzględniają niepoprawne dopasowania pomiędzy oczywistym znaczeniem pola a znaczeniem wynikającym z nazwy lub definicji zmiennej.

Warto notować wszystkie problemy związane z jakością danych. Więcej informacji można znaleźć w temacie “Opracowanie raportu jakości danych” na stronie 17.

Przykład z dziedziny handlu internetowego — weryfikowanie jakości danych

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Weryfikacja jakości danych jest często realizowana podczas procesu opisu lub analizy danych. Podczas tych faz sprzedawca internetowy może napotkać na następujące problemy:

Brak danych. Do poznanych braków danych należą niezawierające odpowiedzi ankiety zarejestrowanych użytkowników. Bez dodatkowych informacji dostępnych w kwestionariuszu tacy klienci nie będą prawdopodobnie uwzględnieni w kolejnych modelach.

Błędy danych. Większość źródeł danych generowanych jest automatycznie, więc ten problem nie ma dużego znaczenia. Błędy typograficzne w produktowej bazie danych można znaleźć w procesie eksploracji.

Błędy pomiarowe. Największym źródłem potencjalnych błędów pomiarów jest kwestionariusz. Jeśli którekolwiek z odpowiedzi zostaną nieprecyzyjnie wyrażone lub niewłaściwie sformułowane, sprzedawca internetowy może uzyskać odpowiedzi różniące się od oczekiwanych. Dlatego podczas procesu eksploracji należy zwracać szczególną uwagę na pozycje z nietypowym rozkładem odpowiedzi.

Opracowanie raportu jakości danych

Po przeprowadzeniu eksploracji i weryfikacji jakości danych można przystąpić do opracowywania raportu przydatnego podczas kolejnej fazy metodologii CRISP-DM. Więcej informacji można znaleźć w temacie “Weryfikowanie jakości danych” na stronie 16.

Lista zadań

Jak wskazano wcześniej, istnieje kilka rodzajów problemów związanych z jakością danych. Przed przejściem do kolejnego kroku warto przeanalizować następujące problemy w zakresie jakości i zaplanować odpowiednie rozwiązanie. Wszystkie odpowiedzi należy zanotować w raporcie jakości danych.

- Czy zostały zidentyfikowane brakujące atrybuty i puste zmienne? Jeśli tak, czy braki danych mają znaczenie?
- Czy występują niespójności w pisowni, które mogą utrudnić późniejsze łączenie lub transformacje?
- Czy zostały przeanalizowane odchylenia, tak aby stwierdzić, czy stanowią one szum, czy zjawisko warte dokładniejszego zbadania?
- Czy została przeprowadzona kontrola wiarygodności danych? Należy opisać oczywiste konflikty (np. nastolatki z wysokimi dochodami).
- Czy wzięto pod uwagę wykluczenie danych niemających wpływu na hipotezy?
- Czy dane przechowywane są w plikach płaskich? Jeśli tak, czy we wszystkich plikach występują te same separatory? Czy każdy rekord zawiera tę samą liczbę zmiennych?

Przed wykonaniem kolejnego kroku

Przed przygotowaniem danych do modelingu w programie IBM SPSS Modeler należy odpowiedzieć na następujące pytania:

W jakim stopniu dane są zrozumiałe?

- Czy wszystkie źródła danych zostały przejrzyście zidentyfikowane i udostępnione? Czy występują jakiekolwiek problemy lub ograniczenia?
- Czy na podstawie dostępnych danych zostały zidentyfikowane kluczowe atrybuty?
- Czy te atrybuty ułatwiają sformułowanie hipotez?
- Czy został odnotowany rozmiar wszystkich źródeł danych?
- Czy podzbiory danych można wykorzystywać tam, gdzie jest to wskazane?
- Czy dla każdego interesującego atrybutu zostały obliczone podstawowe statystyki? Czy zostały wyłonione istotne informacje?
- Czy w celu dokładniejszej analizy kluczowych atrybutów zostały wykorzystane rysunki wyjaśniające? Czy otrzymane wnioski spowodowały przeformułowanie hipotez?
- Czy w tym projekcie występują problemy z jakością danych? Czy został opracowany plan rozwiązania tych problemów?
- Czy kroki w zakresie przygotowywania danych są zrozumiałe? Na przykład, czy wiadomo, które źródła danych należy połączyć i które atrybuty należy przefiltrować lub wybrać?

Po zdobyciu wiedzy biznesowej oraz zrozumieniu podstawowych kwestii dotyczących danych nadszedł czas, aby zastosować IBM SPSS Modeler w celu przygotowania danych do modelowania.

Rozdział 4. Przygotowanie danych

Przygotowanie danych — przegląd

Przygotowanie danych jest jednym z najważniejszych i najbardziej czasochłonnych aspektów eksploracji danych. Szacuje się, że przygotowanie danych pochłania 50–70% czasu i wysiłków w całym projekcie. Poświęcenie odpowiedniej ilości energii podczas wcześniejszej fazy zapoznania z zadaniem i zrozumienia danych może zmniejszyć te nakłady pracy, ale wciąż przygotowanie i pakowanie danych w celu eksploracji będzie fazą wymagającą intensywnych działań.

W zależności od organizacji i jej celów w fazie przygotowania danych można zwykle wyróżnić następujące zadania:

- Łączenie zbiorów danych i/lub rekordów
- Wybór przykładowego podzbioru danych
- Agregowanie rekordów
- Opracowywanie nowych atrybutów
- Sortowanie danych na potrzeby modelingu
- Usuwanie lub zastępowanie pustych wartości lub braków danych
- Dzielenie na zbiory uczące i zbiory danych testowych

Wybór danych

W oparciu o początkowo zebrane dane we wcześniejszej fazie CRISP-DM można przystąpić do wyboru danych istotnych z perspektywy celów eksploracji danych. Ogólnie dostępne są dwa sposoby wyboru danych:

- **Wybór pozycji (wierszy)** polega na podejmowaniu decyzji dotyczących uwzględnianych kont, produktów lub klientów.
- **Wybór atrybutów lub znaków (kolumny)** polega na podejmowaniu decyzji dotyczących użycia cech, takich jak kwota transakcji czy dochód gospodarstwa domowego.

Przykład z dziedziny handlu internetowego — wybór danych

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Wiele decyzji sprzedawców internetowych dotyczących wyboru danych podejmowanych jest we wcześniejszych fazach procesu eksploracji danych.

Wybór pozycji. Początkowe badanie będzie ograniczone do ok. 30 000 klientów zarejestrowanych w serwisie. Filtry należy zatem ustawić tak, aby wykluczać zakupy i dzienniki sieciowe niezarejestrowanych klientów. Inne filtry powinny eliminować wezwania dotyczące plików graficznych i innych nieinformacyjnych wpisów w dziennikach sieciowych.

Wybór atrybutów. Zakupowa baza danych będzie zawierać dane wrażliwe dotyczące klientów sprzedawcy internetowego. Istotne jest więc, aby filtrować atrybuty, takie jak imię i nazwisko/nazwa klienta, adres, numer telefonu i numery karty kredytowej.

Uwzględnianie lub wykluczanie danych

Po podjęciu decyzji dotyczącej uwzględnienia lub wykluczenia podzbiorów danych należy pamiętać o zanotowaniu motywów tej decyzji.

Ważne pytania

- Czy dany atrybut jest istotny z perspektywy celów eksploracji danych?

- Czy jakość określonego zbioru danych lub atrybutu stanowi przeszkodę dla ważności uzyskanych wyników?
- Czy można wykorzystać takie dane?
- Czy istnieją ograniczenia dotyczące wykorzystania określonych zmiennych, takich jak *płeć* lub *rasa*?

Czy decyzje podjęte na tym etapie różnią się od hipotez sformułowanych podczas fazy zrozumienia danych? Jeśli tak, należy pamiętać o zanotowaniu w raporcie projektu przyczyn takich rozbieżności.

Czyszczenie danych

Czyszczenie danych umożliwia dokładniejsze poznanie problemów występujących w danych wybranych do analizy. W rozwiązaniu IBM SPSS Modeler dostępnych jest kilka sposobów czyszczenia danych za pomocą węzłów rekordów i operacji zmiennych.

Tabela 2. Czyszczenie danych

Problem związany z danymi	Możliwe rozwiązanie
Brak danych	Należy wykluczyć wiersze lub cechy. Opcjonalnie można uzupełnić puste wartości szacowaną wartością.
Błędy danych	Posługując się logiką, należy wykryć błędy i je usunąć. Opcjonalnie można wykluczyć cechy.
Niespójności kodowania	Należy wybrać jeden schemat kodowania, a następnie przeliczyć i zastąpić wartości.
Brak danych lub niepoprawne dane	Należy ręcznie przeanalizować podejrzane zmienne i określić poprawne znaczenie.

Raport jakości danych przygotowywany podczas fazy zrozumienia danych zawiera szczegóły dotyczące typów problemów występujących w danych. Raport może stać się punktem wyjściowym do manipulowania danymi w rozwiązaniu IBM SPSS Modeler.

Przykład z dziedziny handlu internetowego — czyszczenie danych

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Sprzedawca internetowy korzysta z procesu czyszczenia danych w celu rozwiązania problemów odnotowanych w raporcie jakości danych.

Brak danych. Istnieje możliwość, że klienci, którzy nie uzupełnili kwestionariusza internetowego, zostaną wyłączeni z wybranych modeli w późniejszych fazach procesu. Klientów tych można ponownie poprosić o uzupełnienie kwestionariusza, jednak będzie to kosztowe i czasochłonne zadanie. Sprzedawca internetowy nie możne sobie na to pozwolić. Sprzedawca może jednak wykonać modelowanie różnic w zakupach pomiędzy klientami, którzy uzupełnili kwestionariusz, a klientami, którzy tego nie zrobili. Jeśli obie grupy klientów mają podobne zwyczaje zakupowe, wówczas brakujące kwestionariusze nie są tak poważnym problemem.

Błędy danych. W tym miejscu można poprawić błędy wykryte w procesie eksploracji. W większości przypadków poprawne wpisy danych wymuszane są w serwisie, zanim klient prześle stronę do bazy danych zalecza.

Błędy pomiarowe. Nieprecyzyjnie sformułowane pytania zawarte w kwestionariuszu mogą znacząco wpłynąć na jakość danych. Jest to dość poważny problem, podobnie jak brakujące kwestionariusze. Firma może nie mieć dodatkowych środków ani czasu, by zebrać odpowiedzi na nowo zadane pytania. Najlepszym rozwiązaniem w odniesieniu do problematycznych pozycji może być powrót do procesu wyboru i wyeliminowanie tych pozycji z dalszej analizy poprzez zastosowanie odpowiedniego filtru.

Opracowanie raportu czyszczenia danych

Notowanie działań w zakresie czyszczenia danych ma kluczowe znaczenie przy rejestrowaniu zmian danych. Podczas realizacji projektów eksploracji danych w przyszłości pomocne będzie korzystanie ze szczegółowych notatek dotyczących wcześniejszych doświadczeń.

Lista zadań

Podczas pisania raportu warto wziąć pod uwagę następujące pytania:

- Jaki rodzaj szumu wystąpił w danych?
- Jakie podejście zastosowano w celu wyeliminowania szumu? Które techniki były skuteczne?
- Czy istnieją przypadki lub atrybuty, których nie można odzyskać/wykorzystać? Należy zanotować dane węzła wykluczone ze względu na szum.

Tworzenie nowych danych

Tworzenie nowych danych jest dość częstą czynnością. Na przykład pomocne może być utworzenie nowej kolumny, w której zakupy z wydłużoną gwarancją są oflagowywane podczas każdej transakcji. Nową zmienną, *purchased_warranty*, można łatwo utworzyć w programie IBM SPSS Modeler za pomocą węzła Flagowanie.

Nowe dane można utworzyć na dwa sposoby:

- Opracowywanie atrybutów (kolumny lub cechy)
- Generowanie rekordów (wiersze)

IBM SPSS Modeler oferuje wiele sposobów tworzenia danych za pomocą węzłów operujących na rekordach i zmiennych.

Przykład z dziedziny handlu internetowego — tworzenie danych

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Podczas przetwarzania dzienników sieciowych można utworzyć wiele nowych atrybutów. Dla zdarzeń zarejestrowanych w dziennikach sprzedawca internetowy będzie chciał utworzyć znaczniki czasu, zidentyfikować odwiedzających i sesje oraz zanotować odwiedzane strony i typ działania reprezentowany przez zdarzenie. Wybrane zmienne zostaną wykorzystane do utworzenia większej liczby atrybutów, np. czas pomiędzy zdarzeniami w sesji.

Kolejne atrybuty można utworzyć po scaleniu lub innej przebudowie danych. Na przykład, po podsumowaniu dzienników sieciowych przedstawiających zdarzenie dla każdego wiersza (każdy wiersz stanowi sesję) zostaną utworzone nowe atrybuty rejestrujące łączną liczbę działań, łączny czas i łączne zakupy podczas jednej sesji. Po scaleniu dzienników sieciowych z bazą danych klientów, tak aby każdy klient miał swój wiersz, zostaną utworzone nowe atrybuty rejestrujące liczbę sesji, łączną liczbę działań, całkowity przeznaczony czas oraz łączne zakupy każdego klienta.

Po utworzeniu nowych danych sprzedawca internetowy realizuje proces eksploracji, aby sprawdzić, czy proces tworzenia danych został przeprowadzony poprawnie.

Opracowywanie atrybutów

W programie IBM SPSS Modeler istnieje możliwość użycia następujących węzłów do opracowania nowych atrybutów:

- Nowe zmienne można utworzyć na podstawie istniejących za pomocą **węzła Wyliczanie**.
- Zmienną flagi można utworzyć za pomocą **węzła Flagowanie**.

Lista zadań

- Podczas opracowywania atrybutów należy wziąć pod uwagę wymagania dotyczące danych. Czy dla algorytmu modelowania oczekiwany jest określony rodzaj danych, na przykład dane numeryczne? Jeśli tak, należy wykonać wymagane działania.
- Czy przed modelowaniem dane należy znormalizować?
- Czy brakujące atrybuty można utworzyć poprzez agregację, uśrednianie lub indukcję?
- Czy korzystając z posiadanej wiedzy, można opracować istotne fakty (np. czas spędzony w serwisie) na podstawie istniejących zmiennych?

Integrowanie danych

Często zdarza się, że dla tego samego zestawu pytań biznesowych dostępnych jest kilka źródeł danych. Na przykład dla tego samego zbioru klientów dostępne mogą być dane dotyczące kredytów hipotecznych oraz dane demograficzne. Jeśli zbiory danych zawierają ten sam unikalny identyfikator (np. numer ubezpieczenia społecznego), w rozwiązaniu IBM SPSS Modeler można je połączyć za pomocą tej zmiennej kluczowej.

Istnieją dwa podstawowe sposoby integracji danych:

- **Scalanie** danych polega na scaleniu dwóch zbiorów danych z podobnymi rekordami, ale różnymi atrybutami. Dane są scalane za pomocą tego samego kluczowego identyfikatora dla poszczególnych rekordów (tj. identyfikator klienta). Dane wynikowe narastają w kolumnach lub znakach.
- **Dołączanie** danych polega na integrowaniu dwóch lub większej liczby zbiorów danych z podobnymi atrybutami, ale różnymi rekordami. Dane są integrowane w oparciu o podobne zmienne (np. nazwa produktu lub długość kontraktu).

Przykład z dziedziny handlu internetowego — integrowanie danych

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Istnieje wiele źródeł danych, a zatem wiele różnych sposobów na integrację danych:

- **Dodawanie atrybutów klientów i produktów do danych zdarzenia.** Aby wykonać modelowanie zdarzeń dzienników sieciowych za pomocą atrybutów z innych baz danych, każdy identyfikator klienta, numer produktu i numer zamówienia związany z poszczególnymi zdarzeniami musi być poprawnie zidentyfikowany, a powiązane atrybuty muszą być scalone z przetwarzanymi dziennikami sieciowymi. Należy pamiętać, że w scalonym pliku informacje dotyczące klienta i produktu replikowane są za każdym razem, kiedy klient lub produkt zostaje powiązany ze zdarzeniem.
- **Dodawanie informacji dotyczących zakupów i dzienników sieciowych do danych o klientach.** Aby wykonać modelowanie wartości klienta, zakupów i sesji, informacje należy wybrać z odpowiednich baz danych, podsumować i scalić z bazą danych klientów. Konieczne jest wówczas utworzenie nowych atrybutów. Szczegóły zostały omówione w procesie tworzenia danych.

Po integracji baz danych sprzedawca internetowy realizuje proces eksploracji, aby sprawdzić, czy proces scalania danych został przeprowadzony poprawnie.

Zadania integracji

Integracja danych może stać się złożonym procesem, jeśli w organizacji nie poświęci się dostatecznej ilości czasu na poznanie i zrozumienie danych. Warto zastanowić się nad pozycjami i atrybutami, które wydają się najistotniejsze z perspektywy celów eksploracji danych. Następnie można przystąpić do integracji danych.

Lista zadań

- Korzystając z węzłów dołączania lub scalania w rozwiązaniu IBM SPSS Modeler, należy przeprowadzić integrację zbiorów danych przydatnych podczas modelowania.
- Przed przystąpieniem do modelowania warto zapisać uzyskane wyniki.
- Po scaleniu dane można uprościć poprzez **agregację** wartości. Agregacja polega na wyliczeniu nowych wartości poprzez podsumowanie informacji z wielu rekordów i/lub tabel.
- Konieczne może być również wygenerowanie nowych rekordów (np. średnich potrąceń z kilku lat składania połączonych deklaracji podatkowych).

Formatowanie danych

Przed przejściem do budowania modelu warto sprawdzić, czy określona metoda nie wymaga specjalnego formatu lub kolejności danych. Na przykład często zdarza się, że w celu uruchomienia modelu algorytm kolejności wymaga wstępnie posortowanych danych. Nawet jeśli sam model umożliwia sortowanie danych, użycie węzła sortowania przed modelowaniem może przyspieszyć proces przetwarzania.

Lista zadań

Podczas formatowania danych należy wziąć pod uwagę następujące pytania:

- Jakie modele zostaną użyte?
- Czy te modele wymagają określonego formatu lub kolejności danych?

W przypadku konieczności wprowadzenia zmian narzędzia do przetwarzania dostępne w rozwiązaniu IBM SPSS Modeler ułatwiają wymagane manipulacje danymi.

Ostatnie kroki przed modelowaniem

Przed przystąpieniem do budowania modeli w rozwiązaniu IBM SPSS Modeler należy odpowiedzieć na następujące pytania.

- Czy wszystkie dane są dostępne w rozwiązaniu IBM SPSS Modeler?
- Czy w oparciu o początkową eksplorację i zrozumienie możliwy był wybór właściwych podzbiorów danych?
- Czy dane zostały skutecznie oczyszczone lub czy zostały usunięte pozycje, których nie można naprawić? Wszystkie podejmowane decyzje należy zanotować w raporcie końcowym.
- Czy zbiory dane zostały właściwie zintegrowane? Czy wystąpiły problemy ze scalaniem, które należy zanotować?
- Czy wykonano analizę wymagań wykorzystywanych narzędzi do modelowania?
- Czy występują problemy z formatowaniem, które można rozwiązać przed modelowaniem? Mowa tu o wymaganych formatach oraz zadaniach, które mogą skrócić czas modelowania.

Po udzieleniu odpowiedzi na powyższe pytania można przystąpić do istoty eksploracji danych — modelowania.

Rozdział 5. Modelowanie

Przegląd modelowania

To etap, podczas którego zaczynają być widoczne efekty wcześniejszych intensywnych działań. W rozwiązaniu IBM SPSS Modeler przygotowywane dane są przenoszone do narzędzi analitycznych. Wyniki rzucają nieco światła na problem biznesowy zgłoszony podczas fazy Zapoznanie z zadaniem.

Modelowanie jest zwykle przeprowadzane w formie wielu iteracji. Eksperci w dziedzinie eksploracji danych uruchamiają zwykle kilka modeli przy użyciu domyślnych parametrów, a następnie dostosowują parametry lub wracają do fazy przygotowania danych w celu wykonania manipulacji wymaganych w wybranym modelu. W organizacjach rzadko zdarza się, że odpowiedź pytanie, dla którego przeprowadzona jest eksploracja danych, znajduje się po jednokrotnym wykonaniu jednego modelu. To dlatego eksploracja danych jest tak interesująca — proponuje wiele sposobów analizy jednego problemu, a IBM SPSS Modeler dostarcza szeregu narzędzi ułatwiających taką analizę.

Wybór technik modelowania

Chociaż osoby pracujące w projekcie mogą już wiedzieć, jakie typy modelowania najlepiej realizują potrzeby firmy, to teraz właśnie nadszedł czas na podjęcie decyzji o ich wykorzystaniu. Wybór odpowiedniego modelu będzie zwykle oparty na następujących zagadnieniach:

- **Typy danych dostępnych do eksploracji.** Na przykład, czy interesujące zmienne to zmienne jakościowe (symboliczne)?
- **Cele eksploracji danych.** Czy w organizacji dąży się do uzyskania wglądu w składnice danych transakcyjnych i odkrycia interesujących wzorców zakupowych? A może konieczne jest uzyskanie wyniku wskazującego np. skłonność do niespłacania kredytu studenckiego?
- **Określone wymagania w zakresie modelowania.** Czy w modelu wymagany jest określony rozmiar lub typ danych? Czy wymagany jest model z wynikami, które można łatwo zaprezentować?

Aby uzyskać więcej informacji o typach modeli w rozwiązaniu IBM SPSS Modeler oraz powiązanych wymogach, należy zapoznać się z dokumentacją IBM SPSS Modeler lub skorzystać z pomocy online.

Przykład z dziedziny handlu internetowego — techniki modelowania

Wybór technik modelowania wykorzystywanych przez sprzedawcę internetowego dyktowany jest celami eksploracji danych:

Udoskonalone rekomendacje. W najprostszym ujęciu polega to na grupowaniu zamówień w celu określenia, które produkty są najczęściej kupowane razem. Aby wyniki były bardziej kompleksowe, można dodać dane klientów oraz rekordy odwiedzin. W tym typie modelowania można skorzystać z techniki dwustopniowej analizy skupień lub sieci Kohonena. Skupienia można następnie profilować za pomocą zestawu reguł C5.0. Można w ten sposób określić, które rekomendacje są najbardziej optymalne w dowolnym momencie podczas wizyty klienta.

Udoskonalona nawigacja w serwisie. W tej chwili sprzedawca internetowy będzie identyfikował najczęściej odwiedzane strony, których wyszukanie wymaga kilku kliknięć. Takie działanie wymaga zastosowania w dziennikach sieciowych algorytmu sekwencji, tak aby wygenerować unikalne ścieżki klientów w serwisie. Następnie konieczne jest wyszukanie sesji, podczas których użytkownicy odwiedzili wiele stron, ale nie podjęli żadnych działań. W następnej kolejności należy przeprowadzić dokładniejszą analizę oraz zastosować techniki analizy skupień. W ten sposób można zidentyfikować typy odwiedzin i odwiedzających, a treść na stronie można uporządkować i zaprezentować w zależności od tego typu.

Wybór odpowiednich technik modelowania

W rozwiązaniu IBM SPSS Modeler dostępnych jest wiele technik modelowania. Eksperci w dziedzinie eksploracji danych często do analizy problemu wykorzystują różne podejścia i perspektywy.

Lista zadań

Podczas podejmowania decyzji dotyczącej wyboru modelu/modeli należy zastanowić się nad następującymi zagadnieniami oraz ich wpływem na dokonane wybory:

- Czy w modelu należy podzielić dane na zbiór uczący i zbiór testowy?
- Czy dostępne są wystarczające dane, aby uzyskać wiarygodne wyniki dla danego modelu?
- Czy w modelu wymagany jest określony poziom jakości danych? Czy taki poziom występuje w bieżących danych?
- Czy rodzaj posiadanych danych pasuje do określonego modelu? Jeśli nie, czy można wykonać niezbędne konwersje za pomocą węzłów manipulowania danymi?

Aby uzyskać więcej informacji o typach modeli w rozwiązaniu IBM SPSS Modeler oraz powiązanych wymogach, należy zapoznać się z dokumentacją IBM SPSS Modeler lub skorzystać z pomocy online.

Założenia dotyczące modelowania

Podczas wybierania narzędzi modelowania warto robić notatki dotyczące procesu decyzyjnego. Należy zanotować wszystkie założenia dotyczące danych oraz czynności manipulowania danymi wykonane w celu spełnienia wymogów modelu.

Na przykład węzły regresji logistycznej i sieci neuronowej wymagają typów danych, które są w pełni **zrealizowane** (typy danych są znane) przed wykonaniem. Oznacza to, że konieczne będzie dodanie węzła typu do strumienia i wykonanie go w celu uruchomienia danych przed zbudowaniem i uruchomieniem modelu. Podobnie modele predykcyjne, np. C5.0, mogą korzystać z ponownego ważenia danych podczas przewidywania reguł dla rzadkich zdarzeń. Podczas tworzenia takich predykcji często można uzyskać lepsze wyniki poprzez dodanie węzła ważenia do strumienia i wprowadzenia do modelu bardziej zrównoważonego podzbioru.

Należy pamiętać o zanotowaniu takich decyzji.

Tworzenie projektu testów

Ostatnim krokiem przed rozpoczęciem budowania modelu powinna być ponowna analiza sposobu testowania wyników wygenerowanych przez model. Opracowywanie kompleksowego projektu testów powinno składać się z dwóch części:

- Opisanie kryteriów "dobroci" dla modelu
- Zdefiniowanie danych, które zostaną wykorzystane do testowania tych kryteriów

Istnieje kilka sposobów zmierzenia **dobroci** modelu. W modelach nadzorowanych, np. C5.0 i C&R Tree, w pomiarach dobroci szacuje się zwykle poziom błędu określonego modelu. W modelach nienadzorowanych, np. sieci skupień Kohonena, w pomiarach uwzględnia się takie kryteria jak łatwość interpretacji, wdrożenie lub wymagany czas przetwarzania.

Należy pamiętać, że budowanie modelu jest procesem wielokrotnym. Oznacza to, że przed podjęciem decyzji dotyczącym użycia i wdrożenia określonego modelu testowane będą wyniki kilku modeli.

Tworzenie projektu testów

Projekt testów to opis kroków, które należy podjąć w celu sprawdzenia utworzonych modeli. Ponieważ modelowanie jest procesem powtarzalnym, należy wiedzieć, kiedy zakończyć dostosowywanie parametrów i wypróbować inną metodę lub model.

Lista zadań

Podczas tworzenia projektu testów należy odpowiedzieć na następujące pytania:

- Jakie dane zostaną wykorzystane do testowania modeli? Czy dane zostały podzielone na zbiory uczące i zbiory testowe? (Jest to powszechnie używane podejście w modelowaniu).
- W jaki sposób można zmierzyć sukces modeli nadzorowanych (np. C5.0)?
- W jaki sposób można zmierzyć sukces modeli nienadzorowanych (np. sieci skupień Kohonena)?
- Ile razy będzie ponownie uruchamiany model ze zmienionymi ustawieniami przed przejściem do kolejnego modelu?

Przykład z dziedziny handlu internetowego — projekt testów

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Kryteria oceny modeli zależą od analizowanych modeli i celów eksploracji danych:

Udoskonalone rekomendacje. Do czasu zaprezentowania udoskonalonych rekomendacji rzeczywistym klientom nie ma możliwości przeprowadzenia obiektywnej oceny tych rekomendacji. Sprzedawca internetowy może jednak zażądać, aby reguły generujące rekomendacje były na tyle proste, aby miały sens z perspektywy biznesowej. Podobnie w celu wygenerowania różnych rekomendacji dla różnych klientów i sesji reguły powinny być bardziej skomplikowane.

Udoskonalona nawigacja w serwisie. Mając informacje o odwiedzanych stronach w serwisie, sprzedawca internetowy może ocenić zaktualizowany projekt serwisu pod kątem łatwego dostępu do ważnych stron. Podobnie jednak przy rekomendacjach, trudno jest z góry ocenić, jak klienci zareagują na zmieniony serwis. Jeśli firma dysponuje dodatkowymi środkami i czasem, warto przeprowadzić testy w zakresie przydatności.

Budowanie modelu

Na tym etapie należy być już dobrze przygotowanym do budowania rozpatrywanych wcześniej modeli. Przed sformułowaniem ostatecznych wniosków warto poeksperymentować z kilkoma różnymi modelami. Większość specjalistów w zakresie eksploracji danych buduje zwykle kilka modeli i przed ich wdrożeniem lub integracją dokładnie porównuje wyniki.

W celu śledzenia postępów podczas pracy z różnymi modelami warto robić notatki dotyczące ustawień oraz danych wykorzystywanych w każdym modelu. Notatki ułatwią omawianie wyników oraz w razie konieczności umożliwią śledzenie podjętych działań. Na koniec procesu budowy modeli dostępne będą trzy rodzaje informacji, które warto wykorzystać do podejmowania decyzji dotyczących eksploracji danych:

- **Ustawienia parametrów** uwzględniają notatki dotyczące parametrów generujących najlepsze wyniki.
- Utworzone rzeczywiste **modele**.
- **Opis wyników modelu**, w tym informacje dotyczące problemów z wydajnością oraz danymi występujących podczas wykonywania modelu i eksploracji uzyskanych wyników.

Przykład z dziedziny handlu internetowego — budowanie modelu

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Udoskonalone rekomendacje. Tworzone są skupienia na różnych poziomach integracji danych, zaczynając od bazy danych o zakupach, a następnie uwzględniając informacje o klientach i sesjach. Na każdym poziomie integracji skupienia tworzy się przy użyciu różnych ustawień parametrów dla algorytmów dwustopniowych i sieci Kohonena. Dla każdego z tych zgrupowań tworzonych jest kilka różnych zestawów reguł C5.0 z różnymi ustawieniami parametrów.

Udoskonalona nawigacja w serwisie. Węzeł modelowania sekwencji stosowany jest do wygenerowania ścieżek klientów. Algorytm umożliwia określenie kryterium minimalnego zaangażowania wsparcia technicznego. Jest to użyteczne w celu skoncentrowania się na najpopularniejszych ścieżkach klientów. Sprawdzane są różne ustawienia parametrów.

Ustawienia parametrów

Większość technik modelowania korzysta z całej gamy parametrów lub ustawień, które można dostosować w celu sterowania procesem modelowania. Na przykład drzewami decyzyjnymi można sterować poprzez dostosowanie głębokości drzewa i podziałów oraz za pomocą wielu innych ustawień. Na początku większość użytkowników zwykle buduje model, korzystając z opcji domyślnych, a następnie dostosowuje parametry w kolejnych sesjach.

Po określeniu parametrów dających najdokładniejsze wyniki należy pamiętać, aby zapisać strumień oraz wygenerowane węzły modelu. Notowanie optymalnych ustawień może być również pomocne przy automatyzowaniu lub przebudowie modeli za pomocą nowych danych.

Uruchamianie modeli

Uruchamianie modeli w programie IBM SPSS Modeler jest bardzo proste. Aby wyświetlić wyniki, należy dodać węzeł modelowania do strumienia i dokonać edycji parametrów, a następnie wykonać model. Wyniki wyświetlane są w nawigatorze Wygenerowane modele po prawej stronie obszaru roboczego. Aby przeglądać wyniki, można kliknąć model prawym przyciskiem myszy. W przypadku większości modeli wygenerowany model można dodać do strumienia w celu dokładniejszej oceny i dalszego użycia wyników. Program IBM SPSS Modeler umożliwia zapisanie modeli w celu łatwiejszego wykorzystania w przyszłości.

Opis modelu

Podczas oceny wyników modelu należy pamiętać o robieniu notatek dotyczących czynności modelowania. Notatki można zapisać razem z modelem, korzystając z okna dialogowego adnotacji węzła lub narzędzi projektowych.

Lista zadań

Dla każdego modelu należy zapisać następujące informacje:

- Czy z tego modelu można wyciągnąć znaczące wnioski?
- Czy ten model daje możliwość uzyskania nowych spostrzeżeń lub nietypowych wzorców?
- Czy występowały problemy z wykonywaniem tego modelu? Czy czas przetwarzania był optymalny?
- Czy w tym modelu wystąpiły problemy z jakością danych, na przykład duża liczba braków danych?
- Czy wystąpiły niespójności w obliczeniach, które warto zanotować?

Ocena modelu

Dysponując zestawem modeli początkowych, można dokonać dokładniejszej analizy i ocenić ich dokładność i skuteczność, a następnie wybrać modele finalne. Modele można uznać za finalne z kilku powodów. Mogą na przykład nadawać się do bezpośredniego wdrożenia lub przedstawiać interesujące wzorce. Z punktu widzenia organizacji w ocenie modeli pomocne może być odniesienie się do utworzonego wcześniej planu testów.

Kompleksowa ocena modelu

Dla każdego ocenianego modelu warto przeprowadzić ocenę metodyczną opartą na kryteriach wygenerowanych w planie testów. Na tym etapie wygenerowany model można dodać do strumienia i użyć wykresów ewaluacyjnych lub węzłów analizy do przeanalizowania skuteczności wyników. Należy również ocenić, czy wyniki mają logiczny sens lub, czy są zbyt daleko idącym uproszczeniem celów biznesowych (na przykład wyniki przedstawiające sekwencję zakupów wino > wino > wino).

Po wykonaniu oceny modele należy ustać w kolejności w oparciu zarówno o kryteria obiektywne (dokładność modelu), jak i subiektywne (łatwość użycia lub interpretacji wyników).

Lista zadań

- Należy ocenić wyniki uzyskanie z modelu, korzystając z narzędzi do eksploracji danych dostępnych w rozwiązaniu IBM SPSS Modeler, np. wykresów ewaluacyjnych, węzłów analizy lub wykresów walidacji krzyżowej.

- Należy dokonać przeglądu wyników w oparciu o własne zrozumienie problemu biznesowego. Należy skonsultować się z analitykami danych lub innymi specjalistami, którzy mogą posiadać informacje dotyczące znaczenie poszczególnych wyników.
- Należy zastanowić się, czy wyniki modelu można łatwo wdrożyć. Czy organizacja wymaga wdrożenia wyników w sieci, czy też odesłania ich do hurtowni danych?
- Należy dokonać analizy wpływu wyników na kryteria sukcesu. Czy wyniki realizują cele ustalone w fazie zapoznania z zadaniem?

Po uzyskaniu zadowalających odpowiedzi na powyższe pytania i uzyskaniu pewności, że bieżące modele realizują cele organizacji nadszedł czas, aby przystąpić do dokładniejszej oceny modeli i końcowego wdrożenia. Opcjonalnie można ponownie uruchomić modele z dostosowanymi parametrami, wykorzystując zdobytą dotychczas wiedzę.

Przykład z dziedziny handlu internetowego — ocena modelu

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Udoskonalone rekomendacje. Jedna z sieci Kohonena i dwustopniowe grupowanie dały osobno zadawalające wyniki. Sprzedawca internetowy ma jednak trudności z wyborem jednej techniki. Z biegiem czasu firma będzie chciała korzystać z obu rozwiązań, akceptując rekomendacje prezentowane w obu technikach i bardziej szczegółowo analizując sytuacje, w których się różnią. Przy niewielkim nakładzie pracy i wiedzy biznesowej sprzedawca internetowy może dalej opracowywać reguły eliminujące różnice pomiędzy dwoma technikami.

Sprzedawca internetowy odkrył również, że wyniki zawierające informacje o sesji są zaskakująco dobre. Pojawił się dowód sugerujący, że rekomendacje można powiązać z nawigacją po serwisie. Zestaw reguł definiujący, gdzie klient może udać się w następnej kolejności, może być użyty w czasie rzeczywistym. Będzie to miało bezpośredni wpływ na treści wyświetlone na stronach podczas wizyty klienta.

Udoskonalona nawigacja w serwisie. Model sekwencji daje sprzedawcy internetowemu dużo pewności co do możliwości przewidywania ścieżek klienta. Dzięki temu można wygenerować wyniki sugerujące konkretną liczbę zmian w projekcie serwisu.

Rejestrowanie skorygowanych parametrów

Wykorzystując wiedzę zdobytą podczas fazy oceny, jeszcze raz przyjrzymy się modelom. Na tym etapie dostępne są dwa działania:

- Dostosowanie parametrów do istniejących modeli.
- Wybór innego modelu w celu rozwiązania problemu związanego z eksploracją danych.

W obu przypadkach konieczne będzie powtórzenie zadań z zakresu budowania modeli do czasu osiągnięcia oczekiwanych rezultatów. Nie należy martwić się powtarzaniem tego kroku. Bardzo często zdarza się, że specjaliści z zakresu eksploracji danych kilkakrotnie oceniają i uruchamiają modele, zanim otrzymają taki, który odpowiada ich potrzebom. To dobry powód, aby zbudować jednocześnie kilka modeli i porównać wyniki przed skorygowaniem parametrów dla każdego z nich.

Przed wykonaniem kolejnego kroku

Przed przejściem do końcowej oceny modeli należy zastanowić się, czy początkowa ocena była dostatecznie dokładna.

Lista zadań

- Czy wyniki modelu są zrozumiałe?
- Czy wyniki modelu są logiczne? Czy występują oczywiste rozbieżności, które należy dokładniej przeanalizować?
- Czy spoglądając pobieżnie na wyniki, można stwierdzić, że zawierają odpowiedź na pytanie biznesowe postawione przez organizację?
- Czy zostały wykorzystane węzły analizy i wykresy przyrostu lub korzyści w celu porównania i oceny dokładności modelu?

- Czy został przeanalizowany więcej niż jeden typ modelu i czy porównano wyniki?
- Czy wyniki modelu nadają się do wdrożenia?

Jeśli wyniki modelowania danych wydają się dokładne lub istotne, nadszedł czas, aby przed ostatecznym wdrożeniem przeprowadzić gruntowną ocenę.

Rozdział 6. Ocena

Ocena — przegląd

Na tym etapie w projekcie eksploracji danych większość zadań została już zakończona. W fazie modelowania oceniono również, że zbudowane modele są technicznie poprawne i skuteczne, zgodnie ze zdefiniowanymi wcześniej **kryteriami sukcesu eksploracji danych**.

Przed przejściem do kolejnych kroków należy jednak ocenić wyniki przeprowadzonych działań w oparciu o **kryteria sukcesu biznesowego** określone na początku projektu. W ten sposób można dowieść, że organizacja będzie mogła wykorzystać uzyskane wyniki. W procesie eksploracji danych można wygenerować dwa rodzaje wyników:

- Ostateczne **modele** wybrane w poprzedniej fazie metodologii CRISP-DM.
- Wszystkie wnioski lub sugestie uzyskane na podstawie modeli oraz w procesie eksploracji danych. Są to tak zwane **spostrzeżenia**.

Ocena wyników

Na tym etapie należy dokonać formalnej oceny, czy wyniki projektu spełniły kryteria sukcesu biznesowego. Ten krok wymaga dobrego zrozumienia sformułowanych celów biznesowych. Tak więc w zespole oceniającym projekt należy uwzględnić kluczowe osoby odpowiedzialne za podejmowanie decyzji.

Lista zadań

W pierwszej kolejności należy odnotować, czy wyniki procesu eksploracji danych spełniają kryteria sukcesu biznesowego. W raporcie należy odpowiedzieć na następujące pytania:

- Czy wyniki zostały jasno sformułowane? Czy format wyników pozwala na ich łatwą prezentację?
- Czy występują zupełnie nowe lub unikalne spostrzeżenia, które warto podkreślić?
- Czy można uszeregować modele i spostrzeżenia w kolejności uwzględniającej to, w jakim stopniu odnoszą się do celów biznesowych?
- W jakim stopniu uzyskane wyniki nawiązują do celów biznesowych organizacji?
- Jakie dodatkowe kwestie zostały podniesione w otrzymanych wynikach? Jak takie kwestie można sformułować w kontekście biznesowym?

Po ocenie wyników należy opracować listę zatwierdzonych modeli, które zostaną uwzględnione w raporcie końcowym. Na liście powinny znaleźć się modele realizujące cele procesu eksploracji danych oraz cele biznesowe organizacji.

Przykład z dziedziny handlu internetowego — ocena wyników

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Ogólne wyniki z pierwszych czynności eksploracji danych u sprzedawcy internetowego łatwo jest przełożyć na język biznesu: w badaniu opracowano zalecenia dotyczące skuteczniejszych produktów oraz udoskonalony projekt serwisu. Udoskonalony projekt serwisu opiera się na kolejności przeglądania stron przez klientów. Informacja ta wskazuje funkcje w serwisie, którymi zainteresowany jest klient, a które są dostępne dopiero po wykonaniu kilku kroków. Trudniej jest natomiast udowodnić skuteczność zaleceń produktowych, ponieważ reguły w zakresie podejmowania decyzji mogą być dość skomplikowane. W celu opracowania raportu końcowego analitycy będą próbowali zidentyfikować wybrane ogólne trendy, posługując się zestawami reguł, które łatwiej jest wyjaśnić.

Rangowanie modeli. Użycie kilku początkowych modeli wydawało się uzasadnione z biznesowego punktu widzenia, dlatego też rangowanie w tej grupie zostało oparte na kryteriach statystycznych, łatwości interpretacji oraz różnorodności. W ten sposób na podstawie modelu opracowano różne zalecenia dla różnych sytuacji.

Nowe pytania. Jednym z najważniejszych pytań stawianych w tym badaniu jest to, w jaki sposób sprzedawca internetowy może uzyskać więcej informacji o swoich klientach. Informacje znajdujące się w bazie danych klientów w dużym stopniu wpływają na tworzenie skupień dla zaleceń. Chociaż istnieją specjalne reguły tworzenia zaleceń dla klientów z brakującymi informacjami, to mają one jednak bardziej ogólny charakter niż te, które można sformułować dla zarejestrowanych klientów.

Proces przeglądu

W skutecznych metodologiach uwzględnia się zwykle czas na analizę sukcesu oraz słabości zakończonego właśnie procesu. Eksploracja danych nie różni się w tym zakresie. Częścią metodologii CRISP-DM jest nauka na podstawie wcześniejszego doświadczenia. Dzięki temu przyszłe projekty eksploracji danych będą skuteczniejsze.

Lista zadań

W pierwszej kolejności należy podsumować działania i decyzje podejmowane w każdej fazie, w tym czynności przygotowania danych, budowanie modelu itd. Następnie należy dla każdej fazy odpowiedzieć na następujące pytania i przedstawić sugestie dotyczące poprawy:

- Czy ten etap wpłynął na wartość wyników końcowych?
- Czy istnieją sposoby na usprawnienie lub poprawę tej konkretnej fazy lub tego konkretnego działania?
- Jakie niepowodzenia lub błędy pojawiły się w tej fazie? W jaki sposób można ich uniknąć w przyszłości?
- Czy występowały "ślepe zaułki", tzn. modele, które były nieskuteczne? Czy istnieją sposoby na uniknięcie ślepych zaułków, tak aby ukierunkować wysiłki na bardziej produktywne działania?
- Czy wystąpiły jakieś niespodzianki (pozytywne i negatywne) podczas tej fazy? Czy z perspektywy czasu istnieje oczywisty sposób, aby przewidzieć takie zdarzenia?
- Czy istnieją alternatywne decyzje lub strategie, które mogły być wykorzystane w tej fazie? Należy je zanotować dla przyszłych projektów eksploracji danych.

Przykład z dziedziny handlu internetowego — przegląd raportu

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

W wyniku przeglądu projektu eksploracji danych sprzedawca internetowy lepiej zrozumiał wzajemne powiązania pomiędzy krokami procesu. Chociaż sprzedawca internetowy początkowo był niechętny do odtwarzania procesu CRISP-DM, teraz zdał sobie sprawę, że cykliczny charakter procesu zwiększa jego skuteczność. Przegląd procesu pozwolił również sprzedawcy na sformułowanie następujących wniosków:

- Powrót do procesu eksploracji jest zawsze uzasadniony, jeśli w kolejnej fazie procesu CRISP-DM występują nieoczekiwane zdarzenia.
- Przygotowanie danych, zwłaszcza dzienników sieciowych, wymaga cierpliwości, gdyż jest to bardzo czasochłonne zadanie.
- Duże znaczenie ma skoncentrowanie się na istniejącym problemie biznesowym. Kiedy dane będą już gotowe do analizy, bardzo łatwo jest rozpocząć budowę modelu bez odniesienia się do ogólnego kontekstu.
- Po zakończeniu fazy modelowania faza zapoznania z zadaniem będzie miała jeszcze większe znaczenie przy decydowaniu o sposobie wdrożenia wyników oraz o przeprowadzeniu dalszych badań.

Określenie kolejnych kroków

Na tym etapie wygenerowane zostały wyniki i ocenione zostało doświadczenie związane z eksploracją danych. Jakie są więc **kolejne kroki?** Faza ta pozwoli na uzyskanie odpowiedzi na powyższe pytanie w kontekście celów biznesowych eksploracji danych. Na tym etapie można wybrać jedną z poniższych ścieżek działania:

- **Kontynuacja do fazy wdrożenia.** Kolejna faza ułatwi wdrożenie wyników modelu do procesu biznesowego i utworzenie raportu końcowego. Nawet jeśli eksploracja danych zakończyła się niepowodzeniem, należy zrealizować fazę wdrażania metodologii CRISP-DM w celu utworzenia raportu końcowego, który zostanie przekazany sponsorowi projektu.

- **Powrót i udoskonalenie lub zastąpienie używanych modeli.** Jeśli okaże się, że wyniki są prawie, ale nie całkowicie, optymalne, należy zastanowić się nad kolejną rundą modelowania. Doświadczenie zdobyte w tej fazie można wykorzystać do poprawy modeli i uzyskania lepszych wyników.

Na tym etapie podejmowana decyzja będzie dotyczyła dokładności i znaczenia wyników modelowania. Jeśli wyniki realizują cele eksploracji danych i cele biznesowe, można śmiało przystąpić do fazy wdrożenia. Bez względu na podejmowaną decyzję proces oceny należy dokładnie udokumentować.

Przykład z dziedziny handlu internetowego — kolejne kroki

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Sprzedawca internetowy jest przekonany o dokładności i znaczeniu wyników projektu, więc przechodzi do fazy wdrożenia.

Zespół projektowy jest również gotowy do powrotu na wcześniejszy etap procesu i dodanie modeli uwzględniających techniki predykcji. Na tym etapie zespół projektowy czeka na otrzymanie raportu końcowego i zielone światło od osób podejmujących decyzje.

Rozdział 7. Wdrożenie

Wdrożenie — przegląd

Wdrożenie polega na wykorzystaniu nowych spostrzeżeń do wprowadzenia udoskonaleń w organizacji. Może to polegać na formalnej integracji, np. wdrożeniu modelu IBM SPSS Modeler generującego oceny poziomu odejścia, które są następnie wczytywane do hurtowni danych. Wdrożenie może również oznaczać wykorzystanie spostrzeżeń uzyskanych w procesie eksploracji danych do wywołania zmiany w organizacji. Na przykład zostały odkryte niepokojące wzorce danych wskazujące na zmianę zachowania klientów powyżej 30. roku życia. Wyniki te nie mogą być formalnie zintegrowane z systemami informatycznymi w organizacji, jednak bez wątpienia będą użyteczne podczas planowania i podejmowania decyzji marketingowych.

Faza wdrażania w metodologii CRISP-DM składa się zasadniczo z dwóch rodzajów działań:

- Planowanie i monitorowanie wrażania wyników
- Realizacja zadań końcowych, np. opracowanie raportu finalnego i przeprowadzenie przeglądu projektu

W zależności od wymogów organizacji konieczne może być wykonanie jednego z powyższych kroków lub z obu.

Planowanie wdrożenia

Chociaż pracownicy będą chcieli szybko opublikować rezultaty eksploracji danych, należy dokładnie zaplanować płynne i kompleksowe wdrożenie wyników.

Lista zadań

- Pierwszym krokiem jest podsumowanie wyników — zarówno modeli, jak i spostrzeżeń. Ułatwi to określenie modeli, które można łatwo zintegrować w systemach baz danych, oraz spostrzeżeń, które należy przekazać współpracownikom.
- Dla każdego modelu nadającego się do wdrożenia należy utworzyć szczegółowy plan wdrożenia i integracji z systemami w organizacji. Należy zanotować wszystkie szczegóły techniczne, np. wymogi bazy danych dla danych wyjściowych modelu. Na przykład w systemie może istnieć wymóg wdrożenia danych wyjściowych modelowania w formacie z oddzielającymi znakami tabulacji.
- Dla każdego ostatecznego spostrzeżenia należy utworzyć plan, zgodnie z którym osoby odpowiedzialne za strategię będą o tym informowane.
- Czy istnieją alternatywne plany wdrożenia obu typów wyników, które warto podkreślić?
- Należy zastanowić się nad sposobem monitorowania wdrożenia. Na przykład w jaki sposób aktualizować model wdrożony za pomocą rozwiązania IBM SPSS Modeler Solution Publisher? W jaki sposób podejmowana będzie decyzja o tym, że model nie jest już adekwatny?
- Należy zidentyfikować problemy związane z wdrożeniem i utworzyć plan awaryjny. Na przykład osoby podejmujące decyzje mogą wymagać więcej informacji dotyczących wyników modelowania oraz szczegółów technicznych.

Przykład z dziedziny handlu internetowego — planowanie wdrażania

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Pomyślne wdrożenie wyników eksploracji danych u sprzedawcy internetowego wymaga przekazania właściwych informacji właściwym osobom.

Osoby podejmujące decyzje. Osoby podejmujące decyzje należy informować o zaleceniach i proponowanych zmianach w serwisie. Należy również przekazywać im krótkie wyjaśnienia dotyczące wpływu takich zmian. Po zaakceptowaniu wyników badania przez te osoby stosowane informacje należy przekazać zespołowi wdrażającemu zmiany.

Programiści. Osoby prowadzące serwis będą zobowiązane do wdrożenia nowych zaleceń i układu treści w serwisie. Należy poinformować ich o *możliwych* zmianach wynikających z badania, tak aby już teraz poczynili stosowne przygotowania. Przygotowanie zespołu do budowania serwisu na bieżąco w oparciu analizę kolejności w czasie rzeczywistym może być przydatne na późniejszym etapie.

Specjaliści w zakresie bazy danych. Osoby prowadzące bazy danych dotyczące klientów, zakupów i produktów należy powiadomić o sposobie wykorzystywania informacji z bazy danych oraz o atrybutach, jakie mogą zostać dodane do bazy danych w przyszłych projektach.

Przede wszystkim zespół projektowy powinien na bieżąco kontaktować się z każdą z tych grup w celu koordynowania procesu wdrażania wyników i planowania przyszłych projektów.

Planowanie monitorowania i utrzymania

W przypadku pełnego wdrożenia i integracji wyników modelowania działania w zakresie eksploracji danych mogą mieć charakter ciągły. Na przykład, jeśli model wdrażany jest w celu przewidzenia kolejności zakupów internetowych, konieczna będzie okresowa ocena takiego modelu. Ma to na celu zapewnienie skuteczności modelu oraz bieżącej optymalizacji. Podobnie model wdrożony w celu zdobycia większej lojalności wśród najbardziej wartościowych klientów zostanie prawdopodobnie zaktualizowany po osiągnięciu założonego poziomu lojalności. Model ten może zostać wówczas zmodyfikowany i ponownie użyty w celu zdobycia lojalności klientów na niższym, ale wciąż dochodowym poziomie w piramidzie wartości.

Lista zadań

Należy zrobić notatki dotyczące następujących zagadnień i uwzględnić je w raporcie końcowym.

- Które czynniki lub wpływy (np. wartość rynkowa lub zmienność sezonowa) należy śledzić dla poszczególnych modeli lub spostrzeżeń?
- W jaki sposób można zmierzyć i monitorować ważność i dokładność poszczególnych modeli?
- W jaki sposób zostanie określone, kiedy określony model utracił ważność? Należy podać szczegóły dotyczące na przykład progów ważności czy oczekiwanych zmian danych.
- Co stanie się po wygaśnięciu modelu? Czy możliwe będzie ponowne zbudowanie modelu w oparciu o nowsze dane lub wprowadzenie niewielkich korekt? A jeśli zmiany będą poważne, czy konieczne będzie utworzenie nowego projektu eksploracji danych?
- Czy po wygaśnięciu modelu możliwe będzie jego wykorzystanie dla rozwiązania podobnych problemów biznesowych? Na tym etapie kluczowe znaczenie dla oceny celu biznesowego poszczególnych projektów eksploracji danych będzie miała przejrzysta i wyczerpująca dokumentacja.

Przykład z dziedziny handlu internetowego — monitoring i utrzymanie

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Bezpośrednim celem monitorowania jest określenie, czy nowy układ serwisu i udoskonalone rekomendacje dają wymierne rezultaty. Innymi słowy, czy użytkownicy mogą w sposób bardziej bezpośredni dotrzeć do interesujących ich stron. Czy wzrosła sprzedaż wiązana rekomendowanych produktów. Po kilku tygodniach monitorowania sprzedawca internetowy będzie mógł ocenić, czy badanie zakończyło się powodzeniem.

Dodawanie nowych zarejestrowanych użytkowników przebiega teraz automatycznie. Po zarejestrowaniu się na stronie do informacji dotyczących klientów można zastosować bieżące zestawy reguł. Dzięki temu możliwe będzie wskazanie odpowiednich rekomendacji.

Decyzja dotycząca czasu aktualizowania zestawu reguł w celu określenia rekomendacji jest trudniejszym procesem. Aktualizacja zestawu reguł nie jest procesem automatycznym. Tworzenie skupień wymaga ręcznego wprowadzenia informacji dotyczącej trafności poszczególnych rozwiązań dla skupień.

Ponieważ przyszłe projekty generują bardziej złożone modele, potrzeba oraz intensywność monitorowania będą z całą pewnością rosły. W miarę możliwości należy zautomatyzować większość czynności monitorowania, a regularnie opracowywane raporty powinny być udostępniane do przeglądu. Firma może być również zainteresowana tworzeniem modeli generujących predykcje na bieżąco. Taki kierunek wymaga podjęcia bardziej skomplikowanych działań niż te realizowane przez zespół w pierwszym projekcie eksploracji danych.

Tworzenie raportu końcowego

Raport końcowy nie tylko stanowi wspólny mianownik dla wcześniejszych dokumentów, ale również może posłużyć do przekazania uzyskanych wyników. Chociaż może wydawać się to oczywiste, należy pamiętać o zaprezentowaniu wyników różnym grupom osób zainteresowanym wynikami. Może to być zespół administratorów technicznych odpowiedzialnych za wdrożenie reguł modelowania bądź sponsor (np. zarząd lub dział marketingu) podejmujący decyzje na podstawie uzyskanych wyników.

Lista zadań

W pierwszej kolejności należy zastanowić się nad grupą docelową raportu. Czy są to eksperci techniczni, czy kierownicy ds. rynku? W przypadku bardzo rozbieżnej grupy docelowej konieczne może być opracowanie odrębnych raportów. W każdym przypadku raport powinien uwzględniać możliwie jak najwięcej poniższych zagadnień:

- Dokładny opis początkowego problemu biznesowego
- Proces wykorzystywany do eksploracji danych
- Koszy projektu
- Notatki dotyczące wszystkich odchyleń od pierwotnego planu projektu
- Podsumowanie wyników eksploracji danych, zarówno modeli, jak i spostrzeżeń
- Przegląd proponowanego planu dotyczącego wdrożenia
- Zalecenia dotyczące dalszej eksploracji danych, w tym interesujące wskazówki uzyskane podczas eksploracji i modelowania

Przygotowanie prezentacji końcowej

Oprócz raportu z projektu konieczne może być również zaprezentowanie spostrzeżeń uzyskanych w ramach projektu grupie sponsorów lub innym zainteresowanym działom. W takim przypadku można wykorzystać większość informacji z raportu. Należy je jednak przedstawić w szerszej perspektywie. W IBM SPSS Modeler diagramy i wykresy można łatwo wyeksportować do tego rodzaju prezentacji.

Przykład z dziedziny handlu internetowego — raport końcowy

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Największe odchylenie od pierwotnego planu projektu stanowi również interesującą wskazówkę do dalszej eksploracji danych. Celem pierwotnego planu było określenie sposobu, który spowoduje, że klienci będą spędzali więcej czasu w serwisie i przeglądali więcej stron podczas jednej wizyty.

Jak się okazuje, zadowolenie klienta to coś więcej niż tylko jego obecność w serwisie. Z rozkładu częstości dla czasu spędzonego podczas jednej sesji, z podziałem na to, czy sesja została zakończona zakupem czy też nie, wynika, że długość większości sesji zakończonych zakupem jest równa długości sesji dla dwóch skupień sesji bez zakupów.

Dysponując takimi informacjami, warto dowiedzieć się teraz, czy klienci spędzający dużo czasu w serwisie bez robienia zakupów wyłącznie przeglądają strony, czy też nie mogą znaleźć interesujących ich produktów. Kolejnym krokiem jest określenie sposobu zaprezentowania takim klientom interesujących ich produktów, tak aby zachęcić ich do zakupu.

Wykonywanie końcowego przeglądu projektu

Ostatni krok w metodologii CRISP-DM umożliwia sformułowanie końcowych wniosków i uporządkowanie doświadczenia zdobytego w proces eksploracji danych.

Lista zadań

Należy przeprowadzić krótką rozmowę z osobami w największym stopniu zaangażowanymi w proces eksploracji danych. Ważne pytania, które należy zadać podczas takich rozmów, to m.in.:

- Jakie są ogólne wrażenia po zakończeniu projektu?
- Czego można było nauczyć się podczas tego procesu, zarówno w zakresie ogólnej eksploracji danych, jak i dostępnych danych?
- Które części projektu udały się? Gdzie pojawiły się trudności? Czy były informacje, które mogły rozwiązać powstałe trudności?

Po wdrożeniu wyników eksploracji danych warto również porozmawiać z osobami, na które wyniki miały negatywny wpływ, np. z klientami lub partnerami biznesowymi. Celem tych działań jest określenie, czy projekt był celowy i czy wygenerował zakładane korzyści.

Wyniki tych rozmów można podsumować razem z wewnętrznymi wnioskami dotyczącymi projektu w raporcie końcowym. Raport ten powinien koncentrować się na doświadczeniu zdobytym w procesie eksploracji dostępnych danych.

Przykład z dziedziny handlu internetowego — przegląd końcowy

Scenariusz eksploracji sieci za pomocą metodologii CRISP-DM

Rozmowy z członkami projektu. Sprzedawca internetowy doszedł do wniosku, że członkowie projektu najbliżej związani ze wszystkimi etapami badania mają entuzjastyczne podejście do wyników i z niecierpliwością czekają na dalsze projekty. Osoby pracujące z bazami danych wyraziły umiarkowany optymizm — doceniły użyteczność badania, ale wskazywały na nałożone na nich dodatkowe obowiązki. Podczas badania dostępny był konsultant, ale w przyszłych projektach konieczne będzie zaangażowanie dodatkowego pracownika zajmującego się utrzymaniem baz danych. Wynika to ze zwiększającego się zakresu projektu.

Rozmowy z klientami. Opinie klientów uzyskane do tej pory są w dużej mierze pozytywne. Jedną z kwestii, które nie zostały dokładnie przemyślane, był wpływ zmian układu serwisu na istniejących klientów. Przez kilka lat korzystania z serwisu zarejestrowani klienci nabrali pewnych przyzwyczajeń dotyczących układu stron. Opinia zarejestrowanych użytkowników nie jest tak pozytywna jak niezarejestrowanych użytkowników. Co więcej, kilku z nich wyraziło duże niezadowolenie ze zmian. Sprzedawca internetowy musi pamiętać o tej kwestii. Powinien również dokładnie przeanalizować, czy zmiana spowoduje zdobycie takiej liczby nowych klientów, która zrekompensuje utratę istniejących.

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 od 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.

Indeks

R

S

T

U

W

Z

Drukowane w USA