IBM SPSS Modeler 18.2.1 モデル作成ノード

注記

本書および本書で紹介する製品をご使用になる前に、 409 ページの『特記事項』に記載されている情報をお読みください。

本書は、IBM SPSS Modeler バージョン 18 リリース 2 モディフィケーション 0 および新しい版で明記されない限り、以降のすべてのリリースおよびモディフィケーションに適用されます。

お客様の環境によっては、資料中の円記号がバックスラッシュと表示されたり、バックスラッシュが円記号と表示されたりする場合があります。

原典： IBM SPSS Modeler 18.2.1 Modeling Nodes

発行： 日本アイ・ビー・エム株式会社

担当： トランスレーション・サービス・センター

目次

前書き

IBM® SPSS® Modeler は、IBM Corp. が開発した企業強化用のデータ・マイニング・ワークベンチです。SPSS Modeler を使用すると、企業はデータを詳しく調べることにより、顧客および一般市民との関係を強化することができます。SPSS Modeler から得られた情報を利用することで、大口の契約を見込める顧客を獲得し、クロスセルを提案できる案件を見つけだし、新たな顧客層を獲得し、不正を発見し、リスクを減少させ、政府機関へのサービスの提供を改善することができます。

SPSS Modeler の視覚的インターフェースを使用すると、特定ビジネスの専門知識を適用し、より強力な予測モデルを実現し、解決までの時間を短縮します。SPSS Modeler では、予測、分類、セグメンテーション、および関連性検出アルゴリズムなど、さまざまなモデル作成手法を提供しています。モデルを作成した後は、IBM SPSS Modeler Solution Publisher により、企業全体の意思決定者やデータベースにモデルを配布することが可能になります。

IBM Business Analytics について

IBM Business Analytics ソフトウェアは、意思決定者がビジネス・パフォーマンスの改善のために使用可能な、完全で整合性があり正確な情報を提供します。ビジネス・インテリジェンス、予測分析、財務実績および戦略管理、分析アプリケーション の包括的なポートフォリオを利用することによって、現在の実績を明確、迅速に理解し、将来の結果を予測することができます。豊富な業界ソリューション、効果が実証済みのプラクティス、熟練の専門サービスを組み合せることにより、あらゆる規模の会社組織が、最高の生産性を達成し、信頼できる意思決定を自動化し、よりよい結果を実現させることができます。

このポートフォリオの一部として、IBM SPSS Predictive Analytics ソフトウェアを使用する組織は、将来のイベントを予測し、その洞察に基づいて積極的に行動し、より優れた業績を実現することができます。全世界の企業、政府、学術分野のお客様が IBM SPSS の技術を活用し、不正行為を減少させ、リスクを軽減させながら、顧客の獲得、保持、成長において、競争優位を高めることができます。IBM SPSS ソフトウェアを日々の業務に取り入れることによって、組織は業務目標を達成し、大きな競争的優位を獲得することができるよう、意思決定を方向付け、自動化することができるようになります。詳細な情報、または営業担当者へのお問い合わせ方法については、http://www.ibm.com/spss を参照してください。

技術サポート

保守担当のお客様向けに技術サポートが提供されています。IBM Corp. 製品の使用方法、または対応するハードウェア環境へのインストールについてサポートが必要な場合は、テクニカル・サポートにご連絡ください。テクニカル・サポートの詳細は、IBM Corp. Web ページ http://www.ibm.com/support を参照してください。支援を要請される場合は、事前にユーザー、会社組織、およびサポート契約を明確にしておいていただくよう、お願いします。

第 1 章 IBM SPSS Modeler について

IBM SPSS Modeler は、ビジネスの専門知識を活用して予測モデルを迅速に作成したり、また作成したモデルをビジネス・オペレーションに展開して意思決定を改善できるようにする、一連のデータ・マイニング・ツールです。IBM SPSS Modeler は業界標準の CRISP-DM モデルをベースに設計されたものであり、データ・マイニング・プロセス全体をサポートして、データに基づいてより良いビジネスの成果を達成できるようにします。

IBM SPSS Modeler には、マシン学習、人工知能、および統計に基づいたさまざまなモデル作成方法が用意されています。「モデル作成」パレットを利用して、データから新しい情報を引き出したり、予測モデルを作成することができます。各手法によって、利点や適した問題の種類が異なります。

SPSS Modeler は、 スタンドアロン製品として購入または SPSS Modeler Server と組み合わせてクライアントとして使用することができます。後のセクションで説明されているとおり、多くの追加オプションも使用することができます。 詳しくは、https://www.ibm.com/analytics/us/en/technology/spss/を参照してください。

IBM SPSS Modeler 製品

製品と関連するソフトウェアの IBM SPSS Modeler ファミリーの構成は次のとおりです。

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console (IBM SPSS Deployment Manager に付属)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Collaboration and Deployment Services 用 IBM SPSS Modeler Server アダプター

IBM SPSS Modeler

SPSS Modeler はこの製品のすべての機能を搭載したバージョンであり、ユーザーのパーソナル・コンピューターにインストールし、そのコンピューターで実行します。スタンドアロン製品としてローカル・モードで SPSS Modeler を実行できるだけでなく、大規模なデータ・セットを使用する場合にパフォーマンスを向上させるために IBM SPSS Modeler Server と組み合わせて実行することもできます。

SPSS Modeler には難しいプログラミングは必要ありません、正確な予測モデルを迅速かつ直感的に構築することができます。独自のビジュアル・インターフェースを使用すると、データ・マイニング・プロセスを簡単に視覚化することができます。製品に組み込まれている高度な分析機能から得られるデータを活用して、データ内に隠れたパターンやトレンドを発見することができます。結果をモデル化し、その結果に影響を与える要因を理解することにより、ビジネスチャンスをさらに活用するとともに、リスクを軽減できるようにもなります。

SPSS Modeler は SPSS Modeler Professional および SPSS Modeler Premium の 2 つのエディションで使用できます。 詳しくは、トピック 3 ページの『IBM SPSS Modeler のエディション』を参照してください。

IBM SPSS Modeler Server

SPSS Modeler は、クライアント/サーバー・アーキテクチャーを使用して、リソース集中型の操作が必要な要求を、強力なサーバー・ソフトウェアへ分散します。これにより、大規模データ・セットにおけるパフォーマンスが向上します。

SPSS Modeler Server は、1 つまたは複数の IBM SPSS Modeler のインストールと組み合わせてサーバー・ホストで分散分析モードで継続的に実行する、別途ライセンスが必要な製品です。このように、SPSS Modeler Server では、メモリー集中型の操作を、クライアント コンピューターにデータをダウンロードせずにサーバー上で実行できるため、大きなデータ・セットで優れたパフォーマンスを発揮します。また IBM SPSS Modeler Server は、SQLの最適化とデータベース内のモデリング機能をご利用になれますので、さらなるパフォーマンスの向上と各種デー3タ処理の自動化を図ることができます。

IBM SPSS Modeler Administration Console

Modeler Administration Console は、SPSS Modeler Server 構成オプションの多くを管理するグラフィカル・ユーザー・インターフェースです。それらの構成オプションは、オプション・ファイルで設定することも可能です。コンソールは、IBM SPSS Deployment Manager に含まれています。コンソールを使用すると、SPSS Modeler Server インストール済み環境をモニターしたり、構成したりできます。SPSS Modeler Server の現在の顧客は、コンソールを無料で利用できます。アプリケーションは Windows コンピューターにのみインストールできますが、サポートされる任意のプラットフォームにインストールされたサーバーを管理できます。

IBM SPSS Modeler Batch

データマイニングは、通常、対話型のプロセスですが、グラフィカル・ユーザー・インターフェースを使用せずに、コマンドラインから SPSS Modeler を実行することも可能です。例えば、ユーザーの介入なしで実行する長期実行または反復的なタスクなどがあげられます。SPSS Modeler Batch は、通常のユーザー・インターフェースにアクセスせずに SPSS Modeler の完全な分析機能のサポートを提供する製品の特別バージョンです。SPSS Modeler Batch を使用するには、SPSS Modeler Server が必要です。

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher は、外部ランタイム・エンジンで実行するか、外部アプリケーションに埋め込むことができる SPSS Modeler ストリームのパッケージ版を作成することができるツールです。このように、SPSS Modeler がインストールされていない環境で使用するための完全な SPSS Modeler ストリームを公開して展開することができます。SPSS Modeler Solution Publisher は、個別のライセンスが必要とされている IBM SPSS Collaboration and Deployment Services - Scoring サービスの一部として配布されています。このライセンスを使用すると、SPSS Modeler Solution Publisher Runtime が得られ、これによって公開されたストリームを実行することができます。

SPSS Modeler Solution Publisher について詳しくは、IBM SPSS Collaboration and Deployment Services の資料を参照してください。IBM SPSS Collaboration and Deployment Services Knowledge Center に『IBM SPSS Modeler Solution Publisher』と『IBM SPSS Analytics Toolkit』というセクションがあります。

IBM SPSS Collaboration and Deployment Services のための IBM SPSS Modeler Server アダプター

さまざまな IBM SPSS Collaboration and Deployment Services 用のアダプターを使用すると、SPSS Modeler および SPSS Modeler Server を IBM SPSS Collaboration and Deployment Services リポジト

リーとインタラクティブに機能させることができます。このように、リポジトリーに展開された SPSS Modeler ストリームは、複数のユーザーで共有したり、シンクライアント アプリケーションである IBM SPSS Modeler Advantage からアクセスしたりできます。リポジトリーをホストするシステムに、アダプターをインストールします。

IBM SPSS Modeler のエディション

SPSS Modeler は次のエディションで使用できます。

SPSS Modeler Professional

SPSS Modeler Professional は、CRM システムで追跡する行動や対話、人口統計データ、購入行動や販売データなど、多くの構造化データを処理するために必要なすべてのツールを提供しています。

SPSS Modeler Premium

SPSS Modeler Premium は、特化したデータ、または構造化されていないテキスト・データを処理するために SPSS Modeler Professional を拡張する、別途ライセンスが必要な製品です。SPSS Modeler Premium には、以下の IBM SPSS Modeler Text Analytics が含まれます。

IBM SPSS Modeler Text Analytics は、高度な言語技術と Natural Language Processing (NLP) を使用して、構造化されていない多様なテキスト・データをすばやく処理し、重要なコンセプトを抽出および組織化し、そしてそのコンセプトをカテゴリー別に分類します。抽出されたコンセプトとカテゴリーを、人口統計のような既存の構造化データと組み合わせ、IBM SPSS Modeler の豊富なデータ・マイニング・ツールを使用したモデル作成に適用し、焦点を絞ったより良い決定を下すことができます。

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription は、従来の IBM SPSS Modeler クライアントとすべて同じ予測分析機能を提供します。Subscription エディションの場合、定期的に製品アップデートをダウンロードできます。

資料

資料は、SPSS Modeler の「ヘルプ」メニューから参照できます。ここからオンラインの Knowledge Center が開きます。オンライン版 Knowledge Center は、製品の外部で常に利用できます。

各製品の完全な資料 (インストール手順を含む) は、PDF 形式でも提供されており、製品ダウンロードの一部として、個別の圧縮フォルダーに格納されています。また、最新の PDF 文書を Web サイト http://www.ibm.com/support/docview.wss?uid=ibm10874788 からダウンロードすることもできます。

SPSS Modeler Professional ドキュメント

SPSS Modeler Professional のドキュメント スイート (インストール手順を除く) は次のとおりです。

- **IBM SPSS Modeler** ユーザーズ・ガイド**:** SPSS Modeler の使用への全体的な入門で、データ ストリームの作成方法、欠損値の処理方法、CLEM 式の作成方法、プロジェクトおよびレポートの処理方法、および IBM SPSS Collaboration and Deployment Services または IBM SPSS Modeler Advantage に展開するためのストリームのパッケージ方法が含まれています。
- **IBM SPSS Modeler** 入力ノード、プロセス・ノード、出力ノード**:** 各種形式のデータの読み取り、処理、および出力に使用するすべてのノードの説明です。これは、モデル作成ノード以外のすべてのノードについての説明です。

- **IBM SPSS Modeler** モデル作成ノード**:** データ・マイニング・モデルの作成に使用するすべてのノードについての説明です。IBM SPSS Modeler には、マシン学習、人工知能、および統計に基づいたさまざまなモデル作成方法が用意されています。
- **IBM SPSS Modeler** アプリケーション・ガイド**:** このガイドの例では、特定のモデル作成手法および技法について、簡単に対象を絞って紹介します。本ガイドのオンライン バージョンは、「ヘルプ」メニューからも利用できます。詳しくは、トピック『アプリケーションの例』を参照してください。
- **IBM SPSS Modeler Python** スクリプトとオートメーション**:** Python スクリプトによるシステムの自動化に関する情報です。ノードおよびストリームの操作に使用できるプロパティーを含めて説明します。
- **IBM SPSS Modeler** 展開ガイド**:** IBM SPSS Deployment Manager のもとで処理されるジョブ内のステップとして IBM SPSS Modeler ストリームを実行することに関する情報。
- **IBM SPSS Modeler CLEF** 開発者ガイド**:** CLEF では、IBM SPSS Modeler のノードとしてデータ処理ルーチンやモデル作成アルゴリズムなどのサード・パーティー製のプログラムを統合できます。
- **IBM SPSS Modeler** データベース内 マイニング・ガイド**:** サード・パーティー製アルゴリズムを使用してご使用のデータベースの能力を利用してパフォーマンスを向上させ、分析機能の範囲を拡張する方法に関する情報を示します。
- **IBM SPSS Modeler Server** 管理およびパフォーマンス・ガイド**:** IBM SPSS Modeler Server の構成方法と管理方法に関する情報。
- 「**IBM SPSS Deployment Manager** ユーザー・ガイド」。IBM SPSS Modeler Server の監視や構成を行うための Deployment Manager アプリケーションに組み込まれている管理コンソール・ユーザー・インターフェースの使用法に関する情報。
- **IBM SPSS Modeler CRISP-DM** ガイド**:** SPSS Modeler でのデータ・マイニングに対する CRISP-DM 方法の使用に関するステップバイステップのガイドです。
- **IBM SPSS Modeler Batch** ユーザーズ・ガイド**:** IBM SPSS Modeler をバッチ・モードで使用するための完全ガイドで、バッチ・モードでの実行およびコマンド・ライン引数の詳細について説明します。このガイドは、PDF 形式のみです。

SPSS Modeler Premium ドキュメント

SPSS Modeler Premium のドキュメント スイート (インストール手順を除く) は次のとおりです。

- 「**SPSS Modeler Text Analytics** ユーザーズ・ガイド」。SPSS Modeler でテキスト分析を使用する場合の情報。テキスト・マイニング・ノード、インタラクティブ・ワークベンチ、テンプレートなどについて説明します。

アプリケーションの例

SPSS Modeler のデータ・マイニング・ツールは、多様なビジネスおよび組織の問題解決を支援しますが、アプリケーションの例では、特定のモデル作成手法および技術に関する簡単で、目的に沿った説明を行います。ここで使用するデータ・セットは、データ・マイニング担当者が管理するような大規模データ・ストアと比較すると非常に小規模ですが、関係する概念および手法は実際のアプリケーションにも拡張できます。

その例を参照するには、SPSS Modeler の「ヘルプ」メニューから「アプリケーションの例」をクリックしてください。

データ・ファイルとサンプル・ストリームは、製品のインストール・ディレクトリーの `Demos` フォルダーにインストールされています。詳しくは、 5 ページの『Demos フォルダー』を参照してください。

データベース・モデル作成の例 : 例は、『*IBM SPSS Modeler* データベース内マイニング・ガイド』を参照してください。

スクリプトの例 : 例は、『*IBM SPSS Modeler* スクリプトとオートメーション ガイド』を参照してください。

Demos フォルダー

アプリケーションの例で使用されるデータ・ファイルとサンプル・ストリームは、製品のインストール ディレクトリの Demos フォルダー (例: C:¥Program Files¥IBM¥SPSS¥Modeler¥<version>¥Demos) にインストールされています。このフォルダーには、Windowsの「スタート」メニューの IBM SPSS Modeler プログラム・グループから、または「ファイル」 > 「ストリームを開く」ダイアログ・ボックスの最近使ったディレクトリーのリストで「Demos」をクリックしてアクセスすることもできます。

ライセンスの追跡

SPSS Modeler を使用すると、ライセンスの使用状況が一定の間隔で追跡され、ログに記録されます。ログに記録されるライセンスメトリックは *AUTHORIZED_USER* と *CONCURRENT_USER* であり、ログに記録されるメトリックのタイプは、SPSS Modeler に使用するライセンスのタイプによって決まります。

作成されたログファイルは IBM License Metric Tool によって処理可能であり、そのファイルからライセンス使用状況レポートを生成できます。

ライセンスログファイルは、SPSS Modeler クライアントログファイルが記録されるディレクトリと同じディレクトリに作成されます (デフォルトでは %ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log)。

第 2 章 モデル作成の紹介

モデルは、一連の入力フィールドまたは変数に基づいて結果を予測するために使用できるルール、式、または方程式のセットです。例えば、金融機関はモデルを使用して、 過去の申請者に関して既に認識されている情報に基づき、融資申請者のリスクが低いか高いかを予測します。

結果を予測する能力は予測分析の主な目標であり、このモデル作成のプロセスを理解することが、IBM SPSS Modeler を使用するうえで鍵となります。

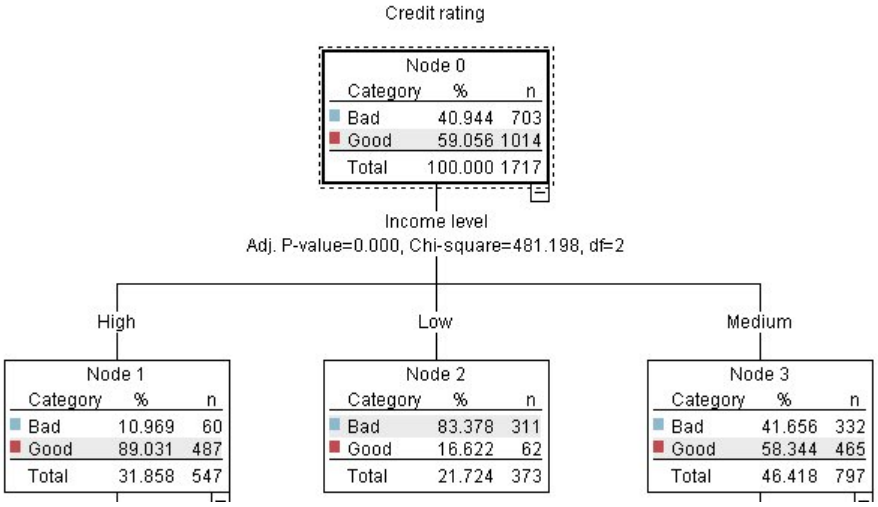


図 *1.* 簡単なディシジョン・ツリー・モデル

この例では、次のような一連のディシジョン・ルールを使用して、レコードの分類 (および回答の予測) を行うディシジョン ツリー モデルを使用します。

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

この例では、一般的な概要を説明する意図で CHAID (カイ 2 乗自動反復検出) モデルを使用しますが、ほとんどの概念は IBM SPSS Modeler のほかのモデル タイプにも広く適用します。

モデルを理解するには、まずそれにあてはめるデータを理解する必要があります。この例のデータには、銀行の顧客に関する情報が含まれます。次のフィールドが使用されています。

フィールド名	説明
Credit_rating	信用度:0=悪い、1=良い、9=欠損値
Age	年齢
Income	収入レベル:1=低、2=中、3=高
Credit_cards	所有するクレジット カード数:1=5 枚未満、2=5 枚以上
Education	学歴:1=高校、2=大学
Car_loans	利用中のカーローン数:1=1 件以下、2=2 件以上

銀行は、ローンを返済したか (信用度 = 良い) 否か (信用度 = 悪い) ということを含めて、銀行から融資を受けている顧客に関する履歴情報のデータベースを管理します。この既存データを使用して、銀行は今後の融資申請者が債務不履行となる可能性がどれほど高いかを予測できるモデルを構築します。

ディシジョン・ツリー・モデルを使用して、顧客の 2 つのグループの特性を分析し、債務不履行の尤度を予測できます。

この例では、*Demos* フォルダー内の *streams* サブフォルダーにある *modelingintro.str* という名前のストリームを使用します。データ・ファイルは、*tree_credit.sav* です。詳しくは、トピック 5 ページの『Demos フォルダー』を参照してください。

ここで、ストリームを詳しく見ていくことにしましょう。

1. メイン・メニューから次の各項目を選択します。

   「ファイル」 > 「ストリームを開く」
2. 「開く」ダイアログ・ボックスのツールバーの金のナゲット・アイコンをクリックし、Demos フォルダーを選択します。
3. *streams* フォルダーをダブルクリックします。
4. *modelingintro.str* という名前のファイルをダブルクリックします。

ストリームの構築

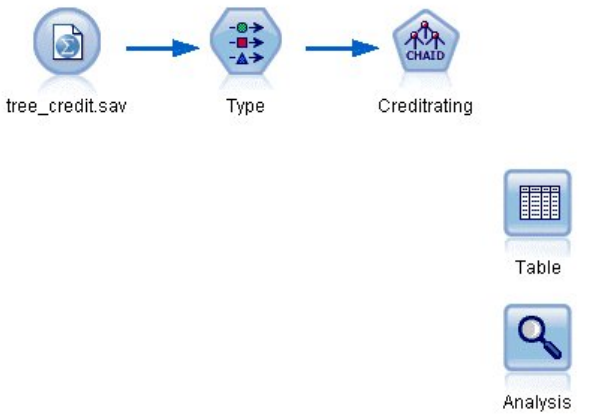


図 2. モデル作成ストリーム

モデルを作成するストリームを構築するには、少なくとも次の 3 つの要素が必要です。

- 外部のソースからデータを読み込む入力ノード。ここでは、IBM SPSS Statistics データ・ファイル。
- 測定の尺度 (フィールドが含んでいるデータの種類) など、フィールド・プロパティーを指定する入力ノードまたはデータ型ノードと、モデル作成の対象または入力値としての各フィールドの役割。
- ストリームが実行されたときにモデル・ナゲットを生成するモデル作成ノード。

この例では、CHAID モデル作成ノードを使用しています。CHAID (Chi-squared Automatic Interaction Detection) は、最適な分割を識別するために、カイ 2 乗統計という種類の統計を使用してディシジョン・ツリーを構築し、ディシジョン・ツリーを分割するのに最適な位置を検出する分類方法です。

測定の尺度が入力ノード内で指定された場合、別個のデータ型ノードは除外できます。機能的に、結果は同じとなります。

ストリームには、モデル・ナゲットが作成されてストリームに追加されたあとスコアリングされた結果を表示するのに使用されるテーブル・ノードおよび精度分析ノードもあります。

Statistics ファイル入力ノードは *tree_credit.sav* データ・ファイルから IBM SPSS Statistics 形式のデータを読み込みます。このデータ・ファイルは *Demos* フォルダーにあります (現在の IBM SPSS Modeler インストールの下のこのフォルダーを参照するには、*$CLEO_DEMOS* という名前の特別な変数が使用されます。これによって、現在のインストール・フォルダーやバージョンにかかわらず、パスが有効になります)。

図 *3. Statistics* ファイル入力ノードを使用してデータを読み込む

データ型ノードが各フィールドの測定の尺度を指定します。測定の尺度は、フィールドのデータの種類を示すカテゴリーです。入力データ・ファイルは、3 つの異なる測定の尺度を使用します。

連続型フィールド (年齢 フィールドなど) には連続した数値が含まれるのに対し、名義型フィールド (信用度 フィールド) には 悪い、良い、またはクレジット履歴なし などの複数の値があります。順序型フィールド (収入レベル・フィールドなど) は、特有の順序を持つ (この場合は低、中 および高) 複数の値を含むデータについて説明します。

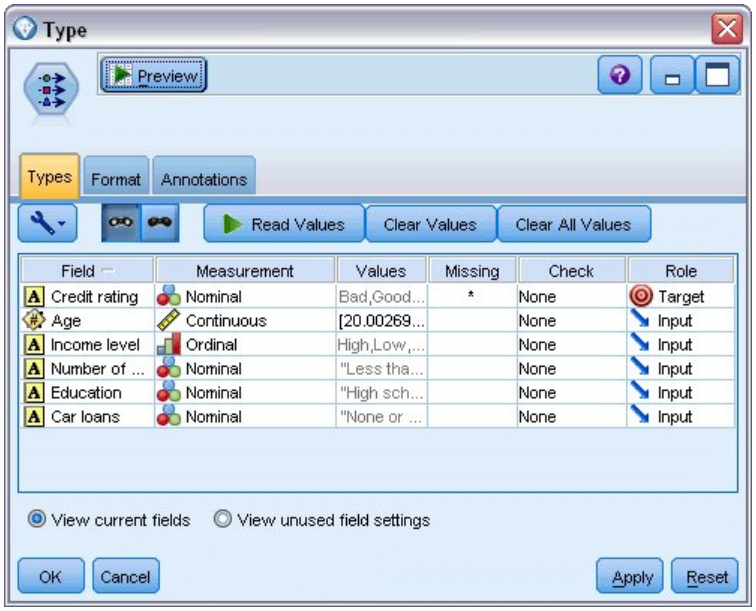


図 4. データ型ノードによる対象フィールドおよび入力フィールドの設定

また、各フィールドについて、データ型ノードは役割を指定して、モデル作成で各フィールドが果たす役割を示します。信用度フィールドの役割は対象と設定されています。これにより、指定された顧客が債務不履行したかどうかが示されます。これが対象、つまり値を予測したいフィールドです。

その他のフィールドについては役割を「入力」に設定します。入力フィールドは、予測フィールドと呼ばれる場合があります。値はモデル作成アルゴリズムによって使用され、対象フィールドの値を予測します。

CHAID モデル作成ノードはモデルを生成します。

モデル作成ノードの「フィールド」タブで、「定義済みの役割を使用」 オプションが選択されています。つまり、データ型ノードで指定された対象と入力値が使用されます。この時点ではフィールドの役割を変更できますが、この例ではそのまま使用します。

1. 「作成オプション」タブをクリックします。

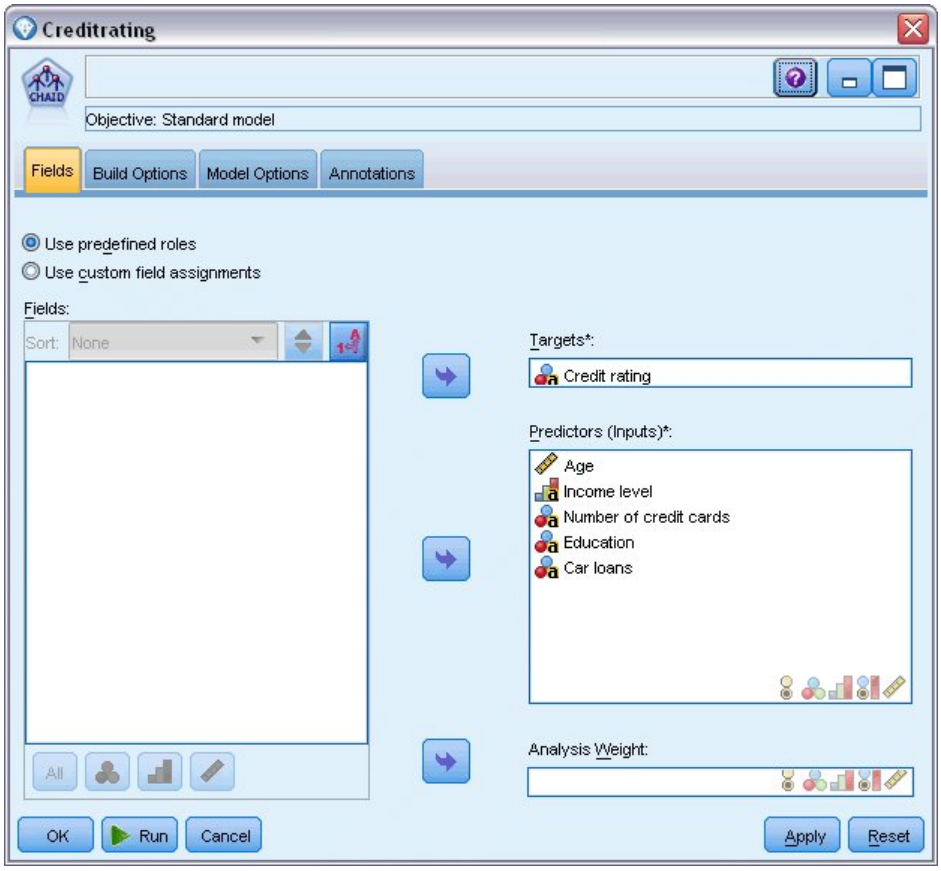


図 *5. CHAID* モデル作成ノードの「フィールド」タブ

作成するモデルの種類を指定できるオプションがいくつかあります。

新規モデルが必要であるため、デフォルト・オプション「新規モデルの作成」 を使用します。

また、拡張機能のない単一の標準ディシジョン・ツリー・モデルが必要であるため、デフォルトの目的オプション 「単一ツリーの構築」 のままにします。

オプションで、インタラクティブなモデル作成セッションを起動して、モデルの微調整を行うことも可能ですが、この例では、デフォルトのモード設定 「モデルの生成」 を使用して単純にモデルを生成します。

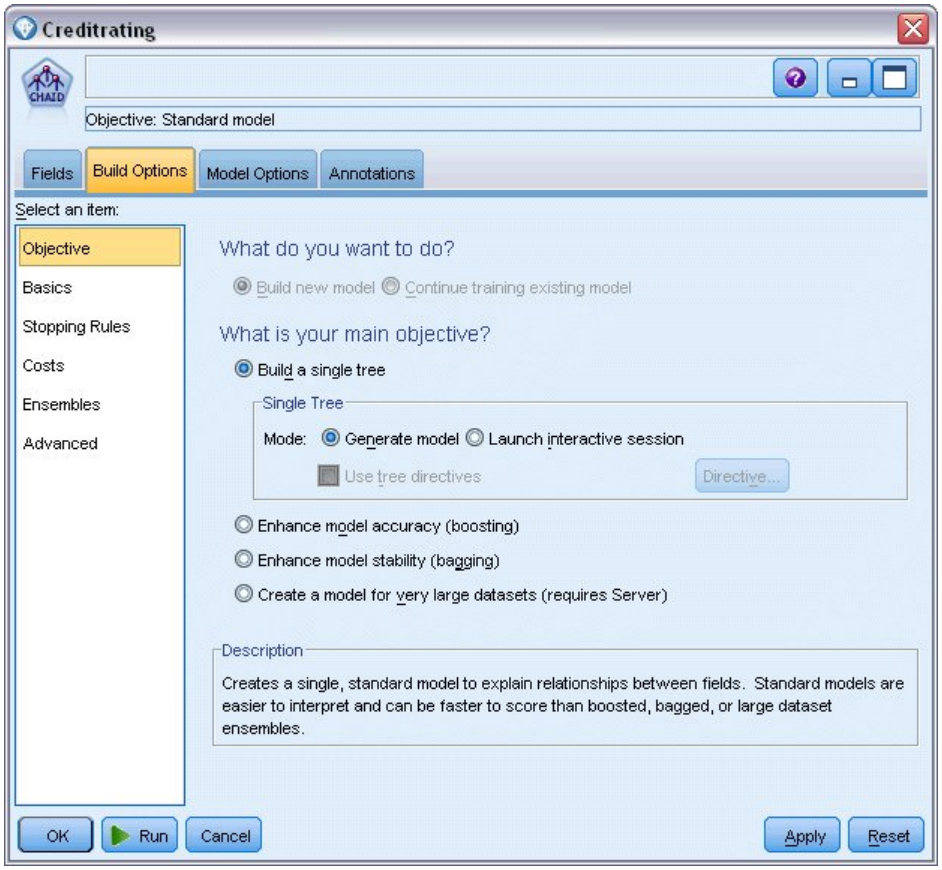


図 *6. CHAID* モデル作成ノードの「作成オプション」タブ

この例では、ツリーを単純にして、親ノードおよび子ノードのケースの最小数を大きくすることにより、ツリーの成長を制限します。

2. 「作成オプション」タブで、左側のナビゲータ・ペインから 「停止規則」 を選択します。
3. 「絶対値を使用」 オプションを選択します。
4. 「親枝の最小レコード」 を 400 に設定します。
5. 「子枝の最小レコード」 を 200 に設定します。

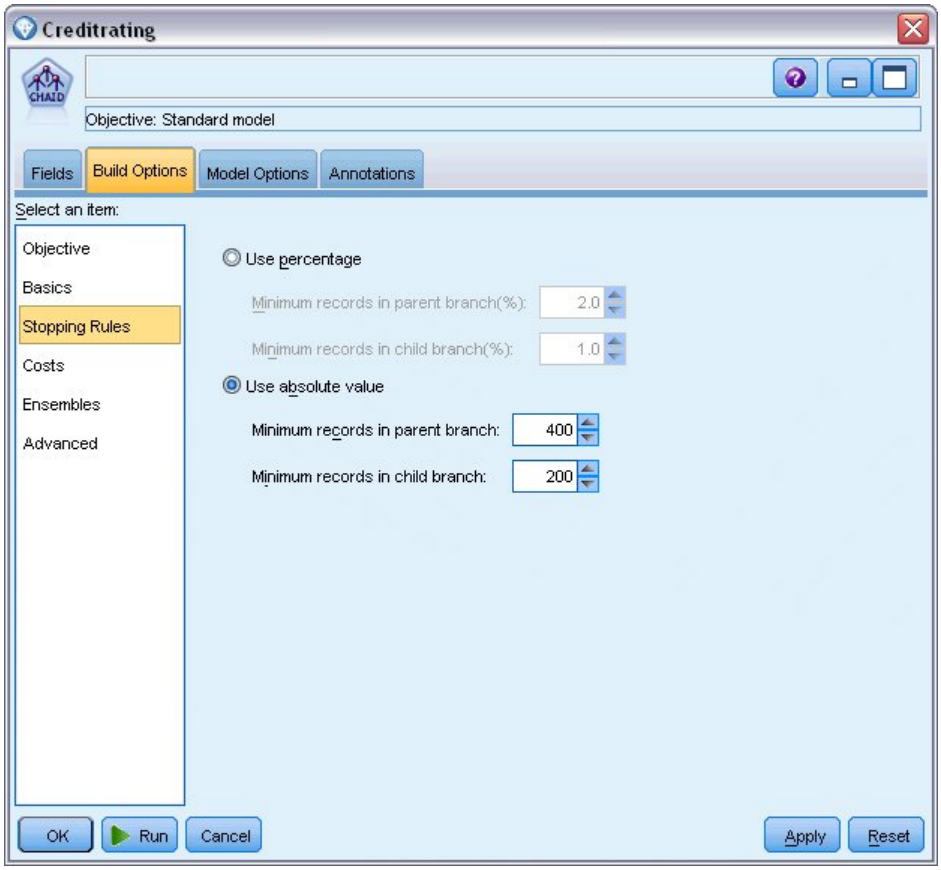


図 7. ディシジョン・ツリー構築の停止基準の設定

この例では、他のすべてのデフォルト・オプションを使用できるため、「実行」をクリックしてモデルを作成します。(または、ノードを右クリックし、コンテキスト・メニューから「実行」を選択するか、あるいはノードを選択し、「ツール」メニューから「実行」を選択します)。

モデルの参照

実行が完了すると、モデル・ナゲットがアプリケーション・ウィンドウの右上角のモデル・パレットに追加されます。また、モデルが作成されたモデル作成ノードへリンクした状態でストリーム・キャンバス内に配置されます。モデルの詳細を表示するには、モデル・ナゲットを右クリックして、モデル・パレットの「参照」 またはキャンバスの 「編集」 を選択します。

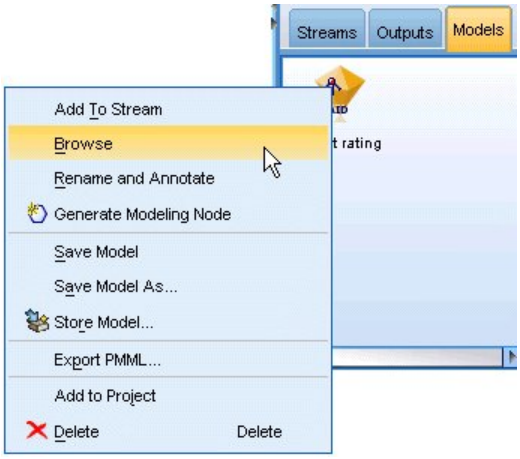


図 *8.* モデル・パレット

CHAID ナゲットの場合、「モデル」タブには、ルール・セットのかたちで詳細が表示されます。これは、基本的に、異なる入力フィールドの値に基づいて、子ノードに個別のレコードを割り当てるために使用できる一連のルールです。

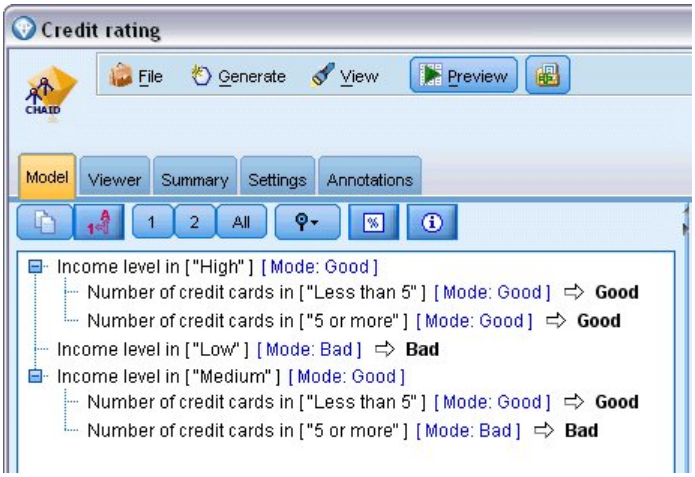


図 *9. CHAID* モデル・ナゲット、ルール・セット

各ディシジョン・ツリー・ターミナル・ノード (それ以上分割されないツリー・ノード) では、「良い」または 「悪い」 の予測が返されます。どちらの場合も、予測はモード、つまり、そのノード内に収まるレコードの最も一般的な回答によって決定されます。

ルール・セット右側の、「モデル」タブには予測変数の重要度のグラフが表示されます。そのグラフには、モデル推定時の各予測値の相対的な重要度が表示されます。これから、「収入レベル」 がこの場合最も有意であり、その他の唯一の有意な因子は 「クレジット カード数」 であることが分かります。

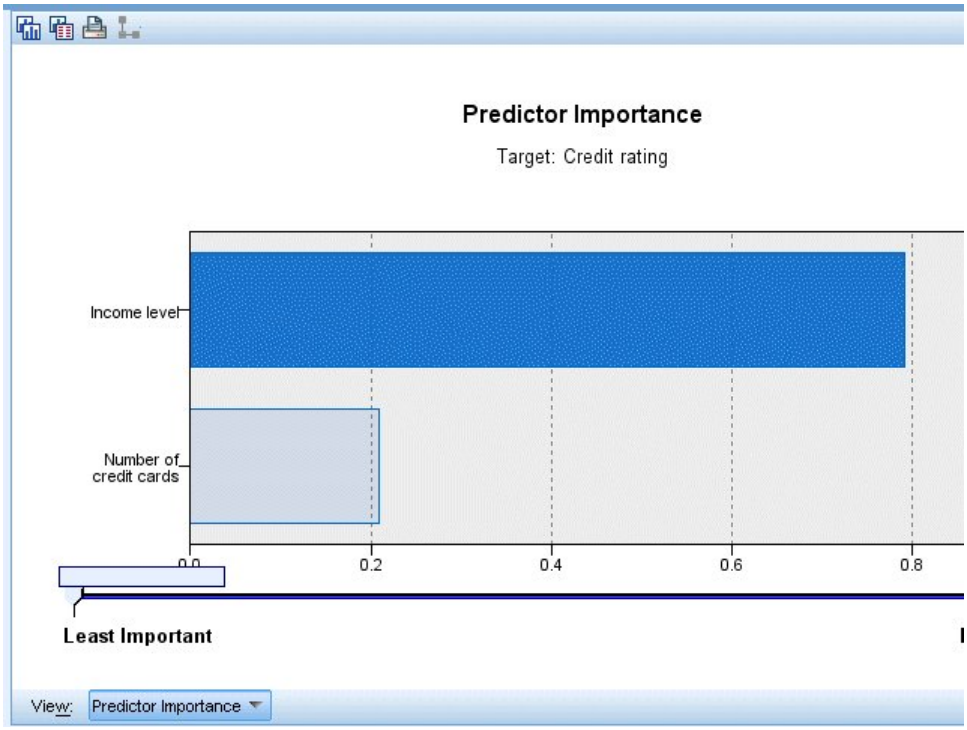


図 *10.* 予測変数の重要度グラフ

モデル・ナゲットの「ビューアー」タブでは、同じモデルを、各ディシジョン・ポイントにノードを配したツリーのかたちで表示します。■ツールバーの「ズーム」コントロールを使用すると、特定のノードをズーム・インして表示したり、ズーム・アウトしてツリー内を広く見たりできます。

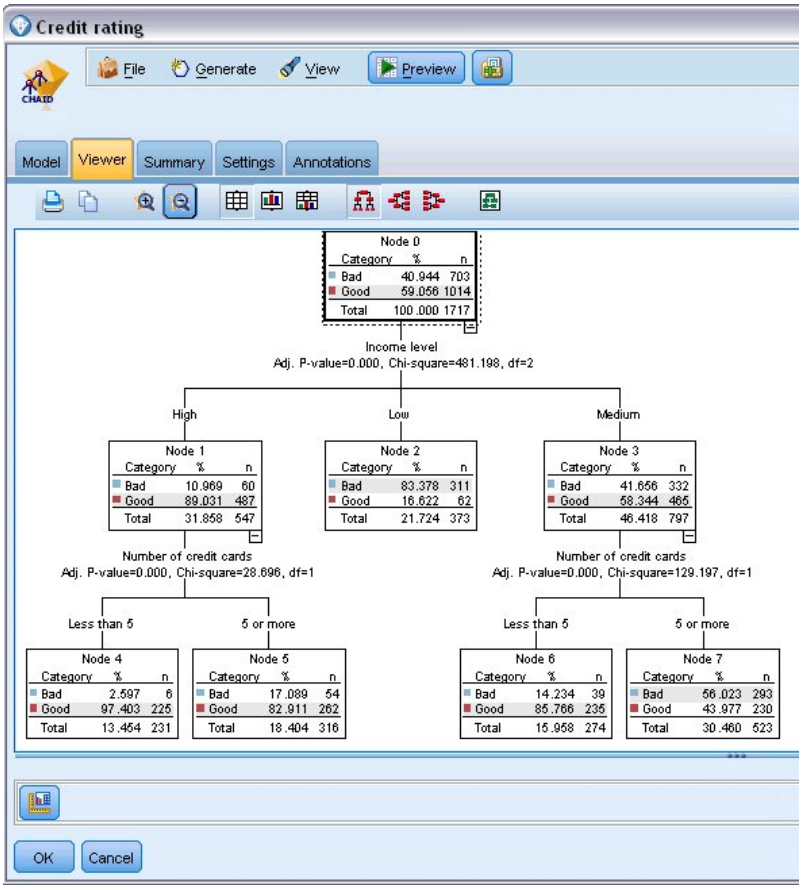


図 *11.* モデル・ナゲットの「ビューアー」タブ、ズーム・アウトを選択

ツリーの上部を見ると、最初のノード (ノード 0) はデータ・セット内のすべてのレコードの要約を示します。 データ・セット内の 40% を超えるケースが、高リスクと分類されています。これはきわめて高い割合です。重要な因子についてツリーがヒントを示すことができるかどうかを見てみましょう。

最初の分割は収入レベルを示すことが分かります。収入レベルが低 カテゴリーのレコードはノード 2 に割り当てられます。このカテゴリーの債務不履行者の割合は最も高くなっています。明らかに、このカテゴリーの顧客に融資することは、高いリスクを有します。

ただし、このカテゴリーの 16% の顧客は債務不履行となっておらず、予測が常に正しいとは限りません。すべての回答をうまく予測できるモデルはありません。しかし、良いモデルは、使用可能なデータに基づいて、各レコードに最も見込みの高い 回答を予測することを可能にします。

同じように、収入の多い顧客 (ノード 1) を見ると、大部分 (89%) の顧客のリスクが低いことが分かります。しかし、これらの顧客の 10 人に 1 人が 債務不履行に陥っています。こうしたリスクを最小限に抑えるために、融資基準を調整できるのでしょうか?

ここで、このモデルがこれらの顧客をどのように 2 つのサブカテゴリー (ノード4 および 5) に分類しているかについて注目してみましょう。分類は顧客の保有しているクレジット カードの数に基づいて行われています。高収入の顧客について、所有クレジット カード数が 5 枚未満の顧客にのみ融資した場合、成功比率が 89% から 97% まで伸びます。これは、さらに満足のゆく結果です。

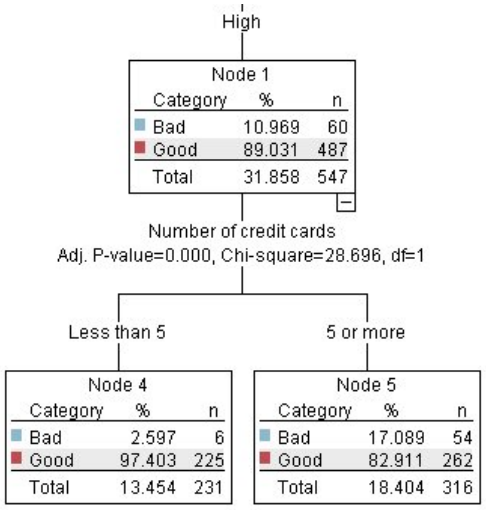


図 *12.* 高収入の顧客のツリー・ビュー

しかし、中程度の収入カテゴリー (ノード 3) の顧客についてはどうでしょうか。 「良い」評価と「悪い」評価に、均等に分かれています。

ここでも、サブカテゴリー (この場合ノード 6 および 7) が役立ちます。今回、所有カード数が 5 枚未満の中程度の収入の顧客にのみ融資すると、「良い」の評価が 58% から 85% に伸び、大幅な改善を示します。

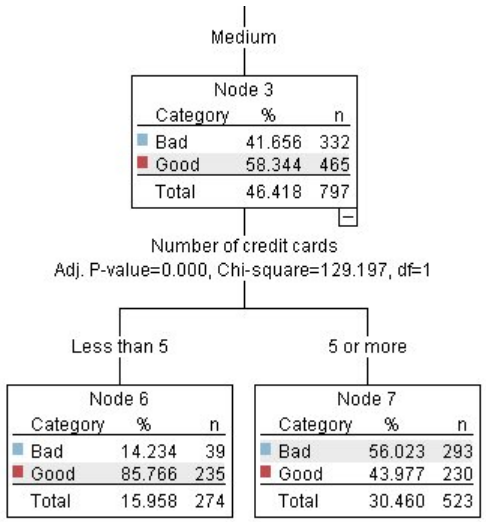


図 *13.* 中程度の収入の顧客のツリー・ビュー

このモデルに入力されたすべてのレコードが特定のノードに割り当てられ、ノードの最も一般的な回答に基づいて、「良い」または「悪い」の予測を割り当てます。

個々のレコードに予測値を割り当てるこのプロセスは、スコアリングとして知られています。モデルを推定するために使用したのと同じレコードをスコアリングすることにより、モデルが学習データ (結果が分かっているデータ) に対してどれだけ正確に実行できるかを評価できます。その方法について説明します。

モデルの評価

モデルを参照すると、スコアリングが機能する方法を理解できます。ただし、それがどれほど正確に 機能するかを評価するには、いくつかのレコードのスコアリングを行って、モデルによって予測された回答と実際の結果とを比較する必要があります。これで、モデルを推定するのに使用されたのと同じレコードをスコアリングし、観測回答と予測回答とを比較することができます。

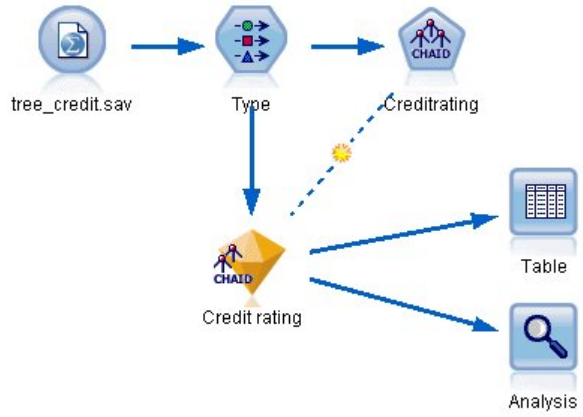
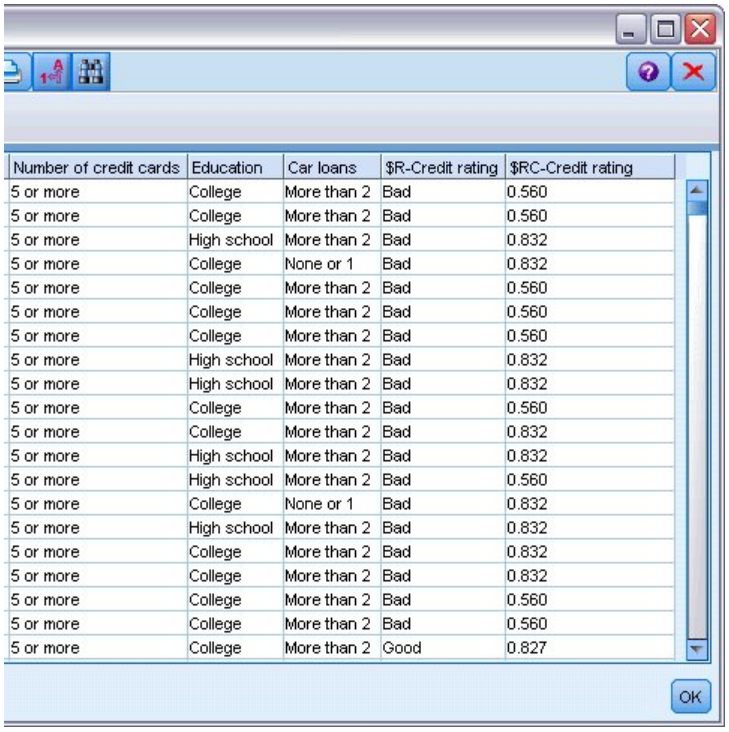


図 *14.* モデル評価を行うためのモデル・ナゲットを出力ノードに接続

1. スコアまたは予測値を確認するには、テーブル・ノードをモデル・ナゲットに接続し、テーブル・ノードをダブルクリックして「実行」をクリックします。

   モデルによって作成された *$R-Credit rating* という名前のフィールドに予測されたスコアが表示されます。これらの値を、実際の回答が含まれている元の 信用度 フィールドの値と比較できます。

   規則により、スコアリング時に生成されるフィールドの名前は、対象フィールドに基づきますが、標準の接頭辞が付加されます。接頭辞 *$G* および *$GE* は、一般化線形モデルにより生成されます。*$R* は、この場合 CHAID モデルにより生成される予測で使用される接頭辞です。*$RC* は、確信度値用に使用されます。*$X* は、通常アンサンブルを使用して生成されます。*$XR*、*$XS*、*$XF* は、対象フィールドがそれぞれ連続型フィールド、カテゴリー型フィールド、セット型フィールド、フラグ型フィールドの場合に接頭辞として使用されます。それぞれのモデル タイプでそれぞれの接頭辞を使用します。確信度値は予測値がどれだけ正確であるかに関するモデル独自の推定で、スケールは 0.0 ～ 1.0 です。



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

OK

図 *15.* 生成されたスコアおよび確信度値を示すテーブル

予測されたとおり、多くのレコードについては予測値と実際値が一致していますが、すべてがそうではありません。その理由は、各 CHAID ターミナル・ノードに回答が混在しているためです。予測は、最も一般的 なものと一致していますが、そのノードのほかのすべてのものに関しては間違っています。(低収入の顧客の 16% は債務不履行に陥っていません。)

これを回避するには、すべてのノードが純粋に 100%、つまり、すべて良い または混在回答のない 悪い になるまで、ツリーを小さいブランチに分割し続けます。ただし、そのようなモデルは非常に複雑で、ほかのデータセットに対してうまく一般化できないことが考えられます。

正しい予測の数を確認するには、テーブルを読み込み、予測フィールド 「*$R-Credit rating*」 の値が「信用度」 の値に一致するレコード数を集計します。精度分析ノードを使用すると自動的に行われるため、より簡単に予測数が分かります。

2. モデル・ナゲットを精度分析ノードに接続します。
3. 精度分析ノードをダブルクリックし、「実行」 をクリックします。

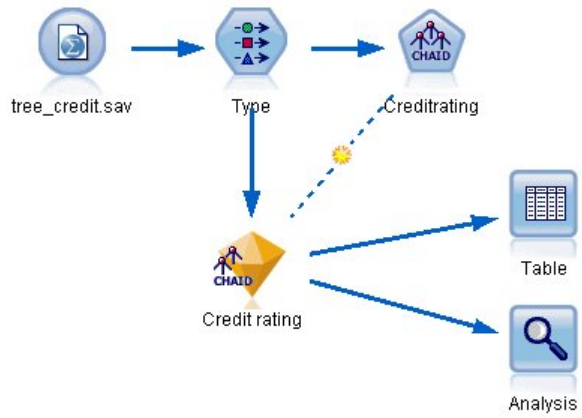


図 *16.* 精度分析ノードの接続

分析の結果、2464 個のレコード中 1899 個 (77% 強) で、モデルによって予測された値と実際の回答が一致したことがわかります。

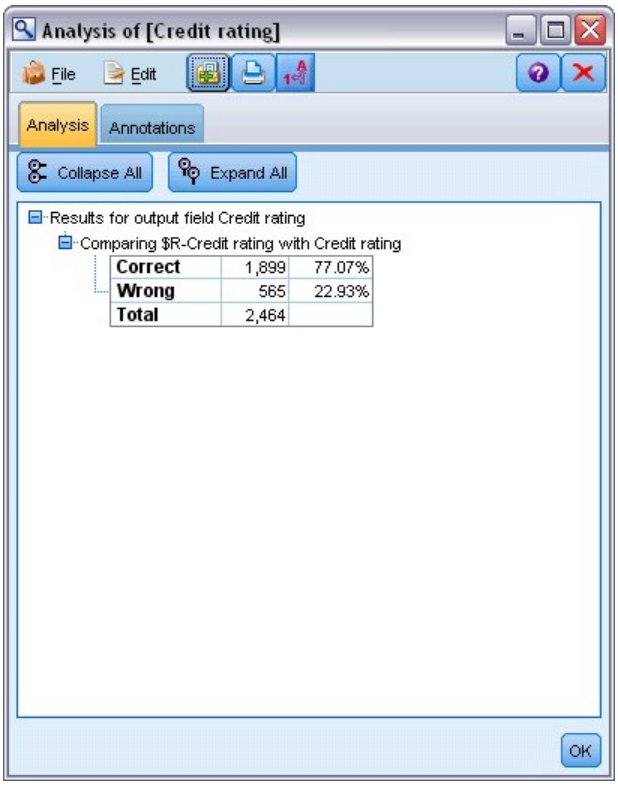


図 *17.* 観測回答と予測回答の比較の分析結果

この結果は、スコアリングされるレコードがモデルの推定に使用されるものと同じであるという事実に制限されます。実際の状況では、データ区分ノードを使用して、データを別個のサンプルに分割し、学習および評価を行います。

1 つのデータ区分サンプルをモデルの生成に使用し、別のデータ区分サンプルをテストに使用することにより、それが、いかにうまくほかのデータセットに一般化できるかについての良い目安を得ることができます。

精度分析ノードを使用すると、すでに実際の結果がわかっているレコードに対してモデルをテストすることができます。次の段階では、結果のわからないレコードをスコアリングするためにモデルをどのように使用するかについて説明します。例えば、このレコードには現在銀行の顧客ではありませんが、販促メールで見込み客となりうる人々が含まれています。

レコードのスコアリング

前の段階で、モデルの精度を評価するためにモデルの推定に使用するものと同じレコードをスコアリングしました。モデルの作成に使用したものとは異なるレコードのセットをスコアリングする方法について説明します。対象フィールドを使用したモデル作成の目的は、結果が分かっているレコードを調べ、まだ分からない結果について予測できるパターンを特定することです。

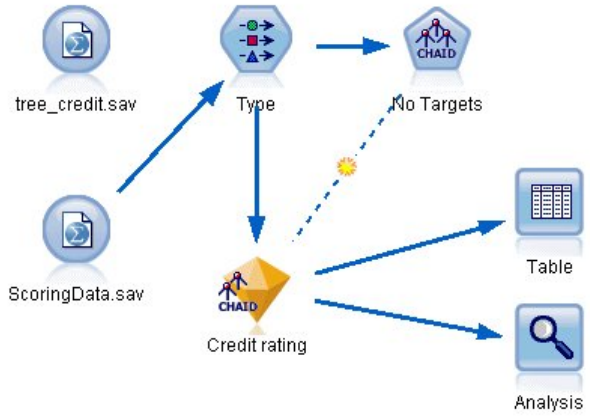


図 *18.* スコアリングのための新しいデータの追加

Statistics ファイル入力ノードを更新して別のデータ・ファイルを指すか、またはスコアリングするデータを読み込む新しい入力ノードを追加します。どちらの場合も、新しいデータセットには、対象フィールド信用度は含まれず、モデルによって使用されたのと同じ入力フィールド (年齢、収入レベル、学歴など) が含まれている必要があります。

別の方法として、必要な入力フィールドを含む任意のストリームにモデル・ナゲットを追加する方法もあります。フィールド名とタイプがモデルによって使用されたものと同じであるかぎり、ファイルからの読み込みであろうとデータベースからの読み込みであろうと、ソース・タイプは関係ありません。

モデル・ナゲットを別のファイルとして保存したり、PMML フォーマットをサポートするその他のアプリケーションで使用するために PMML フォーマットでモデルをエクスポートしたり、IBM SPSS Collaboration and Deployment Services リポジトリーにモデルを格納したりできます。これにより、モデルを全社的に展開して、スコアリングや管理を行うことができます。

モデル自体は、使用されるインフラストラクチャに影響を受けず、同様に機能します。

要約

この例では、モデルの作成、評価、およびスコアリングの基本的なステップを紹介しています。

- モデル作成ノードは、結果がわかっているレコードを調べてモデルを推定し、モデル・ナゲットを作成します。これはモデルの学習と呼ばれることもあります。

- モデル・ナゲットは、レコードのスコアリングを行う予定のフィールドを含む任意のストリームに追加できます。すでに結果がわかっているレコード (既存の顧客など) をスコアリングすることによって、モデルがどれほどうまく実行されているかを評価できます。
- モデルが適度にうまく実行されていると満足したら、新しいデータ (見込み客など) のスコアリングを行って、その回答を予測することができます。
- モデルの学習または推定に使用されるデータは、解析データまたは履歴データと呼ばれる場合があります。また、スコアリング・データはオペレーショナル・データと呼ばれることもあります。

第 3 章 モデル作成の概要

モデル作成ノードの概要

IBM SPSS Modeler には、マシン学習、人工知能、および統計に基づいたさまざまなモデル作成方法が用意されています。「モデル作成」パレットを利用して、データから新しい情報を引き出したり、予測モデルを作成することができます。各手法によって、利点や適した問題の種類が異なります。

IBM SPSS Modeler アプリケーション ガイド では、これらの手法の多くの例が、モデル作成プロセスの概要とともに提供されています。このガイドは、オンライン・チュートリアル、および PDF 形式で使用できます。詳しくは、トピック 4 ページの『アプリケーションの例』を参照してください。

モデル作成方法は、次のカテゴリに分けられます。

- 教師あり学習
- アソシエーション
- セグメンテーション

教師あり学習モデル

教師あり学習モデルでは、1 つまたは複数の入力フィールドの値を使用し、1 つまたは複数の出力、または対象フィールドの値を予測します。これらの手法の例として、ディシジョン ツリー (C&R Tree、QUEST、CHAID および C5.0 アルゴリズム)、回帰 (線型、ロジスティック、一般化線型、Cox 回帰アルゴリズム)、ニューラル・ネットワーク、サポート・ベクター・マシン、Bayesian Network (ベイズ) があります。

組織は教師あり学習モデルを活用して、既知の結果に基づく予測に役立てることができます。例えば、顧客が購入するか立ち去るか、または特定の取引が既知の詐欺パターンに当てはまるかどうか、です。モデル作成手法には、マシン学習、ルール算出、サブグループ識別、統計的手法、および多重モデル生成が含まれます。

教師あり学習ノード

自動分類ノードは、2種類の結果 (yes/no、 churn/don't churn など) を生じる多くの異なるモデルを作成および比較し、与えられた分析への最善のアプローチを選ぶことができるようになります。多くのモデル作成アルゴリズムに対応し、希望する方法、各特定のオプション、そして結果を比較するための基準を選択することができます。このノードで、指定されたオプションに基づいてモデルのセットが生成され、指定された基準に基づいて最善の候補がランク付けされます。

自動数値ノードでは、多くのさまざまな方法を使用し、連続する数値範囲の結果を求めてモデルを推定し比較します。このノードは、自動分類ノードと同じ方法で動作し、1 回のモデル作成のパスで、複数の組み合わせのオプションを使用し試すアルゴリズムを選択することができます。使用できるアルゴリズムには、ニューラル・ネットワーク、C&R Tree、CHAID、線型、一般化線型、サポート・ベクトル・マシン (SVM) が含まれています。モデルは、相関、相対エラー、または使用された変数の数に基づいて比較できます。

C&R Tree (分類と回帰ツリー) ノードは、ディシジョン・ツリーを生成し、将来の観測値を予測または分類できるようにします。この方法は再帰的なデータ区分を使用して学習レコードを複数のセグメントに分割し、各ステップで不純性を最小限に抑えます。ツリーのノードが「純粋」であると考えられるのは、ノード中にあるケースの 100% が、対象フィールドのある特定のカテゴリーに分類される場合です。対象フィールドおよび入力フィールドは、数値範囲またはカテゴリー (名義型、順序型、フラグ) が使用できます。すべての分岐は 2 分割です (2 つのサブグループのみ)。

QUEST ノードには、ディシジョン・ツリーの構築用に2 分岐の方法が用意されています。これは、大規模な C&R Tree 分析が必要とする処理時間を短縮すると同時に、より多くの分割を可能にする入力値が優先される分類ツリー内の傾向を低減するように設計されています。入力フィールドは、数値範囲 (連続型) にできますが、目標変数はカテゴリーでなければなりません。すべての分割は 2 分岐です。

CHAID ノードはディシジョン・ツリーを生成し、カイ二乗統計値を使用して最適な分割を識別します。C&R Tree および QUEST ノードと違って、CHAID は、非 2 分岐ツリーを生成できます。これは、ある分岐が 3 個以上のブランチを持つことを意味します。対象フィールドおよび入力フィールドは、数値範囲 (連続型) またはカテゴリーとなります。Exhaustive CHAID は CHAID の修正版で、可能性のある分割すべてを調べることで、よりよい結果を得られますが、計算時間も長くなります。

C5.0 ノードは、ディシジョン・ツリーとルール・セットのどちらかを構築します。このモデルは、各レベルで最大の情報の対応をもたらすフィールドに基づいてサンプルを分割します。対象フィールドは、カテゴリーでなければなりません。複数の分割を 2 つ以上のサブグループに分割できます。

ディシジョン・リスト・ノードは、母集団に関連する与えられた 2 値の結果の高いもしくは低い尤度を示すサブグループまたはセグメントを識別します。例えば、離れる可能性の少ないもしくはキャンペーンに好意的に答える可能性のある顧客を探すことができます。顧客区分を追加し、結果を比較するために他のモデルを並べて表示することによって、ビジネスに関する知識をモデルに導入することができます。ディシジョン・リスト・モデルは、ルールのリストから構成され、各ルールには条件と結果が含まれます。ルールは順番に適用され、一致する最初のルールで、結果が決まります。

線型モデルは、対象と 1 つまたは複数の予測値との線型の関係に基づいて連続型対象を予測します。

因子分析モデル・ナゲットには、データの複雑性を整理する強力なデータ分解手法が 2 種類あります。主成分分析 (PCA) : 入力フィールドの線型結合が検出されます。成分が互いに直交する (直角に交わる) 場合に、フィールドのセット全体の分散を把握するのに役立ちます。因子分析 : 一連の観測フィールド内の相関パターンを説明する基本因子が識別されます。どちらの手法でも、元のフィールド・セットの情報を効果的に要約する少数の派生フィールドの検出が目標です。

特徴量選択ノードで、（欠損値の割合などの）諸基準に基づいて入力フィールドをスクリーニングして削除にかけ、指定した目標に相対的な残りの入力フィールドの重要度をランク付けします。例えば、数百の潜在的入力フィールドを含むデータセットがあるとして、患者予後のモデリングにはどれが役に立つのでしょう?

判別分析によって、ロジスティック回帰より厳密な仮説を立てることができますが、これらの仮説が一致した場合、ロジスティック回帰分析に対する様々な代替あるいは補足になります。

ロジスティック回帰は、入力フィールドの値に基づいてレコードを分類する統計手法です。線型と似ていますが、数値範囲ではなくカテゴリー対象フィールドを使用します。

一般化線型モデルは、指定したリンク関数によって従属変数が因子および共変量と線型関係になるよう、一般線型モデルを拡張したものです。さらにこのモデルでは、非正規分布の従属変数を使用することができます。 線型、ロジスティック回帰、カウント・データに関するログ線型モデル、そして区間打切り生存モデルなど、統計モデルの機能性が数多く含まれています。

一般化線型混合モデル (GLMN) は線型モデルを拡張したため、対象が非正規分布となる場合があり、指定されたリンク関数を介して因子および共変量に線形に関連し、観測が相関できるようになりました。一般化線型混合モデルには、単純な線型から、非正規分布の縦断的データを取り扱う複雑なマルチレベル・モデルまで、さまざまなモデルがあります。

Cox 線型回帰ノードを使用すると、打ち切りレコードの存在下でイベントまでの時間のデータの生存モデルを構築します。モデルは、対象のイベントが入力変数の指定の値で指定の時間 (*t*) に発生する確率を予測する生存関数を作成します。

サポート・ベクター・マシン (SVM) ノードを使用すると、オーバーフィットすることなく、データを 2 つのグループのいずれかに分類することができます。SVM は、非常に多数の入力フィールドを含むデータセットなど、広範なデータセットを処理することができます。

Bayesian network (ベイズ) ノードを使用すると、観測された情報および記録された情報を実際の知識を組み合わせることによって確率モデルを作成し、発生の尤度を確立できます。このノードは、主に分類に使用される Tree Augmented Naïve Bayes (TAN) および Markov Blanket ネットワークに焦点を当てています。

SLRM (自己学習応答モデル) ノードを使用するとモデルを構築でき、単一または少数の新しいケースを使用して全データを使用するモデルの保持をすることなく、モデルの再見積もりを行うことができます。

時系列ノードは、時系列データから指数平滑法、1 変量の自己回帰型統合移動平均法 (ARIMA)、および多変量 ARIMA (または伝達関数) モデルを推測し、将来のパフォーマンスの予測を作成します。この時系列ノードは、SPSS Modeler バージョン 18 で廃止された以前の時系列ノードと類似しています。ただし、この新しい時系列ノードは、IBM SPSS Analytic Server の機能を活用してビッグ データを処理するよう設計されており、結果モデルは SPSS Modeler バージョン 17 で追加された出力ビューアーに表示されます。

k が整数である場合、*k* 最近傍 (KNN) ノードは、新しいケースを、予測領域の新しいケースに最も近い *k* 個のオブジェクトのカテゴリーまたは値と関連付けます。類似したケースはお互いに近く、類似していないケースはお互いに離れています。

時空間予測 (STP) ノードは、ロケーション・データ、予測用の入力フィールド (予測値)、時間フィールド、および対象フィールドを使用します。各ロケーションには、それぞれの測定時の各予測値を表すデータの行が多数あります。データを分析すると、そのデータを使用して、分析で使用される形状データ内の任意のロケーションの対象値を予測できます。

アソシエーション・モデル

アソシエーション・モデルでは、イベント、購入、属性など、1 つまたは複数のエンティティーが 1 つまたは複数のその他のエンティティーと関連するデータ内のパターンを検出します。モデルは、これらの関係性を定義するルール・セットを構築します。データ内のフィールドは、入力および対象のいずれのフィールドとしても機能します。これらのアソシエーションは手動で検出できますが、アソシエーション・ルール・アルゴリズムはより迅速に検出が可能で、より複雑なパターンも検証できます。Apriori および Carma モデルが、アソシエーション・アルゴリズムの使用例です。他にアソシエーション・モデルの 1 つとしてシーケンス検出モデルがあり、時間構造データのシーケンス・パターンを検索します。

アソシエーション モデルは、例えば、商品 X を購入した顧客は Y と Z も購入するなど、複数の結果を予測する場合に最も役立ちます。アソシエーション・ルールのアルゴリズムは、どのような属性の間にも連関を成立させることができるという点で、より一般的なディシジョン・ツリーのアルゴリズム (C5.0 や C&R Trees など) より勝っています。ディシジョン・ツリーのアルゴリズムは、一つの結果にいたるルールを構築するのに対し、アソシエーション・ルールのアルゴリズムは、それぞれが異なる結果にいたる多数のルールを見つけようとします。

アソシエーション・ノード

Apriori ノードで、データからルール・セットを抽出し、情報内容が最も充実したルールを引き出します。Apriori には、5 種類のルール選択方法があり、高度なインデックス作成方法を使用して、大きなデータ・セットが効率的に処理されます。大きな問題の場合は、一般に、Apriori の方が高速に学習できます。保持できるルール数に特に制限はありません。また、最大 32 の前提条件を持つルールを処理できます。Apriori では、入力フィールドと出力フィールドのすべてがカテゴリーであることが必要ですが、この種類のデータに合わせて最適化されているので、よりよいパフォーマンスを実現します。

CARMA モデルは、入力または対象フィールドを指定しなくても、データからルールのセットを抽出します。Apriori とは対照的に、CARMA ノードでは、前提条件サポートだけではなく、ルール・サポート (前提条件と結果の両方のサポート) を対象とした構築の設定が可能です。これは、生成されたルールをさまざまなアプリケーションで活用できることを意味します。例えば、この休暇シーズンに販売促進する項目を結果とする、商品またはサービス (前提条件) のリストを調べることができます。

シーケンス・ノードで、シーケンシャルな、または時間経過が伴うデータ内のアソシエーション・ルールを検出します。予測可能な順序で起こる傾向にあるアイテム・セットのリストを、シーケンスと呼びます。例えば、顧客がひげそりとアフター・シェーブ ローションを購入した場合、その顧客は次の購入時にシェービング クリームを購入する可能性があります。シーケンス・ノードは CARMA アソシエーション・ルール・アルゴリズムに基づいているため、効率的な 2 段階通過法でシーケンスが検出されます。

アソシエーション・ルール・ノードは Apriori ノードに似ていますが、Apriori とは異なり、アソシエーション・ルール・ノードはリスト・データを処理できます。さらに、アソシエーション・ルール・ノードを IBM SPSS Analytic Server と共に使用すると、ビッグデータの処理や高速な並列処理の利用が可能になります。

セグメンテーション・モデル

セグメンテーション・モデルでは、データを入力フィールドの類似したパターンを持つレコードのセグメント、またはクラスターに分割します。入力フィールドにのみ関心があるため、セグメンテーション・モデルには出力フィールドまたは対象フィールドの概念はありません。セグメンテーション・モデルの例として、Kohonen ネットワーク、K-Means クラスタリング、TwoStep クラスタリングおよび異常値検査があります。

「クラスタリング・モデル」とも呼ばれるセグメンテーション・モデルは、特定の結果が不明である場合に適しています（例えば、詐欺の新しいパターンを識別する場合や、既存の顧客ベースから関心の対象となるグループを識別する場合です）。クラスタリング・モデルは、類似したレコードのグループを識別し、そのグループに従ってレコードにラベルを付けます。この作業は、各グループとそれぞれの特性に関する事前の知識を活用せずに実行されます。これは、クラスタリング・モデルと他のモデル作成技法との違いであり、クラスタリング・モデルには、モデルが予測する定義済みの出力フィールドや対象フィールドはありません。これらのモデルには、正、誤という回答はありません。モデルの価値は、データのグループ構成を把握し、それらのグループについて役に立つ説明を提供できるかどうかで決まります。クラスタリング・モデルは、クラスターやセグメントを作成するためによく利用されます。このクラスターやセグメントは、後の分析で入力として使用されます（例えば、潜在的な顧客を、等質のサブグループに分類する方法です）。

セグメンテーション・ノード

自動クラスタリング・ノードは、同様の特性を持つレコードのグループを識別するクラスタリング・モデルを推定し、比較します。ノードは他の自動化モデル作成ノードと同じように動作し、複数の組み合わせのオプションを単一のモデル作成の実行で検証できます。モデルは、クラスター・モデルの有用性をフィルタリングおよびランク付けする基本的な指標を使用して比較し、特定のフィールドの重要度に基づいて指標を提供します。

K-Means ノードで、データ・セットが異なるグループ (つまりクラスター) へ、クラスタリングされます。この方法で、固定数のクラスターを定義し、クラスターにレコードを繰り返し割り当てて、これ以上調整してもモデルが改善されなくなるまで、クラスターの中心を調整します。*K*-means では、結果を予測するのではなく、入力フィールドのセット内のパターンを明らかにするために、「教師なし学習」として知られるプロセスが使用されます。

Kohonen ノードは、ニューラル・ネットワークの一種であり、データ・セットをクラスター化して異なるグループを形成する目的で使用できます。ネットワークの学習が完了すると、類似のレコードは出力マップで互い近くに表示され、違いの大きいレコードほど離れたところに表示されます。強度の高いユニットを識別するために生成されたモデル内で、各ユニットが獲得した観察の数値を調べることができます。これは、適切なクラスター数についてのヒントになる場合があります。

TwoStep ノードで、2 段階のクラスター化手法が使用されます。最初のステップでは、データを 1 度通過させて、未処理の入力データを管理可能な一連のサブクラスターに圧縮します。2 番目のステップでは、階層クラスター化手法を使用して、サブクラスターをより大きなクラスターに結合させていきます。TwoStep には、学習データに最適なクラスター数を自動的に推定するという利点があります。また、フィールド・タイプの混在や大規模データ・セットも効率よく処理できます。

異常値検査ノードで、「正常な」データのパターンに合致しない異常ケースや外れ値を識別します。このノードで、外れ値が既知のパターンに当てはまらなかったり、何を探しているのかはっきりしなかったりする場合でも、外れ値を識別できます。

データベース内マイニング・モデル

IBM SPSS Modeler は、データベース・ベンダーから入手できる、Oracle Data Miner や Microsoft Analysis Services などのデータ・マイニングおよびモデル作成のツールとの統合をサポートしています。データベース内および IBM SPSS Modeler アプリケーション内のモデル、スコアおよびストア モデルすべての作成が可能です。詳細は、『*IBM SPSS Modeler* データベース内マイニング・ガイド』を参照してください。

IBM SPSS Statistics モデル

コンピューターに IBM SPSS Statistics をインストールしライセンスが付与されている場合、IBM SPSS Modeler 内から特定の IBM SPSS Statistics ルーチンにアクセスおよび実行して、モデルを作成およびスコアリングできます。

分割モデルの作成

分割モデル作成の場合、単一ストリームを使用して、フラグ型、名義型、または連続型の入力フィールドの可能性のある値ごとに別個のモデルを作成できます。作成されたすべてのモデルは、単一のモデル・ナゲットからアクセスできます。入力フィールドの値は、モデルにさまざまな影響を与えます。分割モデル作成によって、ストリームの一度の実行で可能な各フィールド値に最も適合するモデルを取得できます。

インタラクティブモデル作成セッションでは、分割は使用できません。インタラクティブ・モデル作成によってモデルを個別に指定すると、複数のモデルを自動的に作成する分割を使用する利点がありません。

分割モデル作成は、特定の入力フィールドを分割フィールドとして指定することによって動作します。データ型指定でフィールドの役割を「分割」に指定します。

測定の尺度が「フラグ型」、「名義型」、「順序型」、または「連続型」のフィールドのみ、分割フィールドとして指定できます。

複数の入力フィールドを分割フィールドとして割り当てることができます。ただし、この場合、作成されるモデル数が大幅に増加します。選択された分割フィールドに値の組み合わせについて、モデルが作成されます。例えば、それぞれ 3 つの値を持つ 3 つの入力フィールドが分割フィールドとして指定されている場合、27 種類のモデルが作成されます。

1 つまたは複数のフィールドを分割フィールドとして割り当てている場合でも、モデル作成ノードのダイアログに設定されているチェック・ボックスを使用して、分割モデルを作成するか、単一モデルを作成するかを選択できます。

分割フィールドが定義されているにもかかわらずチェック・ボックスが選択されていない場合、モデルは 1 つだけ生成されます。チェック・ボックスが選択されているにもかかわらず分割フィールドが定義されていない場合、分割は無視され、モデルは 1 つだけ生成されます。

ストリームを実行すると、各モデルは分割フィールドの可能な各値のバックグラウンドで作成されますが、モデル・パレットおよびストリーム・キャンバスにはモデル・ナゲットは 1 つだけ配置されます。分割モデル・ナゲットは、分割の記号で示されます。この記号は、ナゲット・イメージに重なった 2 つの灰色の四角形です。

分割モデル・ナゲットを参照して、作成された各モデルをリストから調べることができます。

ビューアーのナゲットのアイコンをダブルクリックすると、リスト内の各モデルを検証することができます。アイコンをダブルクリックすると、各モデルの標準ブラウザー・ウィンドウが開きます。ナゲットがキャンバス上にある場合、グラフのサムネイルをダブルクリックすると、フルサイズのグラフが開きます。詳しくは、トピック 48 ページの『分割モデル・ビューアー』を参照してください。

モデルが分割モデルとして作成されると、モデルから分割プロセスを削除することも、分割モデル作成ノードまたはナゲットから下流で分割を取り消すこともできません。

例: 小売業者は国内の店舗ごとの製品カテゴリーによって販売額を推定したいと考えています。分割モデル作成を使用して、入力データの「店舗」フィールドを分割フィールドとして指定し、一度の操作で店舗ごとの各カテゴリーにモデルを個別に作成できるようにします。その結果生成される情報を使用し、単一モデルの使用時より正確に在庫レベルを制御することができます。

分割および区分

分割にはデータ区分と共通する機能がいくつかありますが、2 つの使用方法は異なります。

データ区分では、データセットをランダムに学習、検定、および (任意で) 検証の 2 つまたは 3 つの部分に分割し、単一モデルのパフォーマンスの検定に使用します。

分割では、データセットを分割フィールドと同じ数のグループに分け、複数のモデルを作成するために使用されます。

区分および分割は、お互い完全に独立して操作されます。モデル作成ノードで一方または両方を選択したり、あるいはどちらを選択しないということもできます。

モデル作成ノードの分割モデルのサポート

多くのモデル作成ノードでは分割モデルを作成できます。例外として、自動クラスタリング、因子分析、特徴量選択、SLRM、Random Trees、Tree-AS、Linear-AS, LSVM、アソシエーション・モデル (Apriori、Carma およびシーケンス)、クラスタリング・モデル (K-Means、Kohonen、Two Step および異常値検査)、Statistics モデル、データベース内モデリングで使用されるノードがあります。

分割モデル作成をサポートするモデル作成ノードは次のとおりです。


C&R Tree


ベイズ・ネット


線型


QUEST


一般化線型


GLMM


CHAID


KNN


STP


C5.0


Cox


One-Class SVM


ニューラル・ネットワーク


自動分類


XGBoost ツリー


ディシジョン・リスト


自動数値


XGBoost Linear


線型回帰


ロジスティック回帰


HDBSCAN


判別分析


SVM


時系列

分割の影響を受ける機能

分割モデルを使用すると、多くの IBM SPSS Modeler の機能にさまざまな点で影響を与えます。ここでは、ストリーム内のその他のノードと共にスプリット・モデルを使用する方法について説明します。

レコード設定ノード

サンプリング・ノードを含むストリームで分割モデルを使用すると、レコードを均等にサンプルするために、分割フィールドによってレコードが層化されます。このオプションを使用できるのは、サンプル方法として「コンプレックス」が選択されている場合です。

バランス調整は、ストリーム内にバランス・ノードがある場合、分割内のレコードのサブセットではなく、入力レコードのセット全体に適用されます。

レコード集計ノードでレコードを集計している場合、分割ごとに集計を計算するには、分割フィールドをキー フィールドに設定します。

フィールド設定ノード

データ型ノードは、分割フィールドとして使用するフィールドを指定するノードです。

注: アンサンブル・ノードは、複数のモデル・ナゲットを組み合わせる際に使用しますが、単一のモデル・ナゲット内には複数の分割モデルが含まれているため、分割アクションの逆を行う際には使用できません。

モデル作成ノード

分割モデルは、予測変数の重要度 (モデル推定時の予測入力フィールドの相対的な重要度) の計算をサポートしていません。分割モデルの作成時、予測変数の重要度の設定は無視されます。

注: 分割モデルの使用時は、調整済み傾向スコア設定が無視されます。

KNN (最近傍) ノードは、対象フィールドを予測するよう設定されている場合にのみ、分割モデルをサポートします。代替設定 (最近傍を識別するのみ) ではモデルを作成しません。オプション「自動的に **k** を選択」が選択されている場合、各分割モデルで最近傍の数が異なる場合があります。そのため、モデル全体では、すべての分割モデルにまたがって検出された最近傍の数と等しい数の列が生成されます。この最大数より最近傍の数が少ない分割モデルの場合、`$null` 値が入力された列数に相当する数の列が存在します。詳しくは、トピック 365 ページの『KNN ノード』を参照してください。

データベース・モデル作成ノード

データベース内モデル作成ノードでは分割モデルをサポートしていません。

モデル・ナゲット

ナゲットには複数のモデルがあり、PMML はこのようなパッケージをサポートしていないため、分割モデルからの **PMML** へのエクスポートはできません。テキストまたは HTML へのエクスポートが可能です。

モデル作成ノードのフィールド・オプション

すべてのモデル作成ノードには、「フィールド」タブがあり、そこからモデルの作成に使用するフィールドを指定できます。

モデルを作成する前に、対象フィールドや入力フィールドを指定する必要があります。いくつかの例外を除いて、すべてのモデル作成ノードは、上流のデータ型ノードからのフィールド情報を使用します。データ型ノードを使用して入力フィールドおよび対象フィールドを選択する場合、このタブで何も変更する必要はありません(ただしシーケンス・ノードとテキスト抽出ノードは例外です。これらのノードは、モデル作成ノードでフィールド設定を指定する必要があります)。

データ型ノードの設定を使用**:** このオプションを選択すると、上流のデータ型ノードからのフィールド情報がこのノードで使用されます。これがデフォルトです。

ユーザー設定を使用**:** このオプションを選択すると、上流のデータ型ノードからのフィールド情報ではなく、ここで指定したフィールド情報がこのノードで使用されます。このオプションを選択した後に、必要に応じて以下のフィールドを指定します。

注 : すべてのノードのすべてのフィールドが表示されるわけではありません。

- トランザクション形式 **(Apriori**、**CARMA**、**MS** アソシエーション・ルール、および **Oracle Apriori** ノードのみ**)** を使用 **:** 入力データがトランザクション形式の場合に選択します。この形式のレコードには、ID 用と内容用の 2 つのフィールドがあります。各レコードは単一のトランザクションまたは項目を示し、同じ ID を指定することによって関連する項目をリンクさせます。データがテーブル形式である場合は子のボックスをオフにします。項目はそれぞれのフラグで示され、各フラグ・フィールドは特定の項目の有無を示し、各レコードは関連する項目の完全セットを示します。 詳しくは、トピック 272 ページの『テーブル形式データとトランザクション形式・データ』を参照してください。
  - **ID:** トランザクション形式データの場合は、リストから ID フィールドを選択します。ID フィールドとして使用できるのは、数値またはシンボル値のフィールドです。選択したフィールドでは、一意の値がそれぞれ、ある分析ユニットを示している必要があります。例えば、マーケット・バスケット分析なら、各 ID が 1 人の顧客を表します。Web ログ分析なら、各 ID が 1 台のコンピューター (IP アドレス) あるいは 1 人のユーザー (ログイン・データ) を表します。
  - 連続する **ID :** (Apriori ノードおよび CARMA ノードのみ) データ・ストリーム中で同じ ID を持つすべてのレコードが一緒に表示されるようにデータをソートしている場合、このオプションを選択すると処理を高速化することができます。データがあらかじめソートされていない場合 (またはわからない場合) は、このオプションは選択しないでください。この場合、ノードが自動的にデータをソートします。

    注**:** データがソートされていない場合にこのオプションを選択すると、モデルで意味のない結果しか得られない可能性があります。
  - 内容。モデルの内容フィールドを指定します。これらのフィールドには、アソシエーション・モデリングで関心の対象となる項目が含まれています。複数のフラグ・フィールド (データがテーブル形式の場合) または単一の名義型フィールド (データがトランザクション形式の場合) を指定できます。
- 目標: 1 つ以上の対象フィールドが必要なモデルの場合に、対象フィールドを選択します。これは、データ型ノードのフィールドの役割を 「対象」 に設定するのと似ています。
- 評価 **:** (自動クラスタリング・モデルのみ。)クラスター・モデルに対象は指定されません。ただし、評価フィールドを指定して、重要度のレベルを識別することができます。また、クラスターがこのフィールドの値をどれほど正確に区別しているかを評価し、クラスターを使用してこのフィールドを予測できるかどうかを識別します。注: 評価フィールドは、複数の値を含む文字列で指定する必要があります。
  - 入力**:** 1 つ以上の入力フィールドを選択します。これは、データ型ノードのフィールドの役割を 「入力」 に設定するのと似ています。
  - データ区分**:** このフィールドでは、モデル構築の学習、テスト、および検証の各ステージ用に、データを独立したサブセット (サンプル) に分割するフィールドを指定できます。1 組のサンプルをモデ

ルの生成に使用し、別のサンプルで生成したモデルをテストすることにより、そのモデルが、このデータに似た性質を持つより大きなデータセットにどの程度適用できるかについての良い目安を得ることができます。データ型ノードまたはデータ区分ノードを使用して複数のデータ区分フィールドが定義されている場合、データ区分を使用する各モデル作成ノードの「フィールド」タブで単一のデータ区分フィールドを選択する必要があります (1 つのデータ区分だけが存在している場合、データ分割を有効にすると、そのデータ区分が必ず自動的に使用されます)。また、選択したデータ区分を分析に適用するには、そのノードの「モデル・オプション」タブでデータ区分が有効になっている必要があります (このオプションの選択を解除すると、フィールド設定を変更することなくデータ区分を無効にすることができます)。

- 分割。分割モデルについて、分割フィールドを選択します。これは、データ型ノードのフィールドの役割を 「分割」 に設定するのと似ています。測定の尺度が「フラグ型」、「名義型」、「順序型」または「連続型」のフィールドのみ、分割フィールドとして指定できます。分割フィールドとして選択されたフィールドは、対象フィールド、入力フィールド、データ区分フィールド、度数フィールドまたは重みフィールドとして使用できません。 詳しくは、トピック 28 ページの『分割モデルの作成』を参照してください。
- 度数フィールドを使用: このオプションにより、フィールドを出現頻度の重みとして選択することができます。例えば、集計データの使用時など、学習データのレコードに複数のユニットが存在する場合に選択します。フィールド値は、レコードごとに示した単位数です。 詳しくは、トピック『度数フィールドと重みフィールドの使用』を参照してください。

注 : エラー・メッセージ「メタデータ **(**入力**/**出力フィールド**)** が無効です」が表示される場合、度数フィールドなどの必須フィールドがすべて指定されていることを確認してください。

- 重みフィールドを使用: このオプションにより、フィールドをケースの重みとして選択することができます。ケースの重みを使用して、出力フィールドのレベル間の分散における相違を処理します。 詳しくは、トピック『度数フィールドと重みフィールドの使用』を参照してください。
- 結果**:** ルール算出ノード (Apriori) の場合は、得られたルール・セットの結果として使用するフィールドを選択します (このフィールドは、データ型ノードで役割が「対象」または「両方」になっているフィールドに対応します)。
- 前提条件**:** ルール算出ノード (Apriori) の場合は、得られたルール・セットの前提条件として使用するフィールドを選択します (このフィールドは、データ型ノードで役割が「入力」または「両方」になっているフィールドに対応します)。

この項の説明とは異なる「フィールド」タブのあるモデルもあります。

- 詳しくは、トピック 289 ページの『シーケンス・ノードの「フィールド」オプション』を参照してください。
- 詳しくは、トピック 276 ページの『CARMA ノードのフィールド・オプション』を参照してください。

度数フィールドと重みフィールドの使用

度数および重みを使用して、いくつかのレコードに他のレコード以上の重要度を与えます。例えば、母集団の 1 つのセクションが学習データ内で低く示されていることがわかっている場合 (重み) や、1 つのレコードが多くの同一ケースを示している場合 (度数) です。

- 度数フィールドの値には正の整数を指定する必要があります。度数が負または 0 のレコードは、分析から除外されます。度数の重みが整数でない場合は、四捨五入された整数になります。
- ケースの重み付けに使用する値には、正の数値を指定する必要がありますが、整数でなくてもかまいません。ケースの重みが負または 0 のレコードは、分析から除外されます。

度数フィールドと重みフィールドのスコアリング

度数フィールドと重みフィールドは学習モデルに使用されますが、スコアリングには使用されません。各レコードのスコアは、それがどれほど多くのケースを表しているかにかかわらず、その特徴に基づくからです。例えば、次の表のデータがあるとします。

表 *1.* データの例

既婚	応答
はい	はい
はい	はい
はい	はい
はい	いいえ
いいえ	はい
いいえ	いいえ
いいえ	いいえ

これに基づくと、4 人の既婚者のうちの 3 人が販売促進活動に応答し、3 人の未婚者のうち 2 人が応答しなかったと結論付けることができます。したがって、次の表に示すように、新しいレコードはすべて、これに基づいてスコアリングします。

表 *2.* スコアリングされたレコードの例

既婚	**$-**応答	**$RP-**応答
はい	はい	0.75 (3/4)
いいえ	いいえ	0.67 (2/3)

代わりに、次の表に示すように、度数フィールドを使用して、学習データをよりコンパクトに格納することもできます。

表 *3.* スコアリングされたレコードの代替例

既婚	応答	度数
はい	はい	3
はい	いいえ	1
いいえ	はい	1
いいえ	いいえ	2

これは全く同じデータセットを表しているため、配偶者の有無だけを基にモデルを構築し、回答を予測します。スコアリングするデータに 10 人の既婚者がいる場合、それらが 10 個の個別のレコードであろうと、度数が 10 の 1 個のレコードであろうと、全員にはいの予測をたてるでしょう。重みは通常は整数ではありませんが、同様にレコードの重要性を示すと考えられます。したがって、度数フィールドと重みフィールドはレコードをスコアリングする場合に使用されません。

モデルの評価および比較

モデルの種類によって、度数フィールドをサポートするものや、重みフィールドをサポートするもの、また、その両方をサポートするものがあります。どの場合も、それが適用される場合は、モデル構築にのみ使

用され、評価ノードまたは精度分析ノードを使用してモデルを評価したり、自動分類ノードおよび自動数値ノードでサポートされる多くの手法を使用してモデルをランク付けする際には考慮されません。

- 例えば、評価グラフを使用してモデルを比較する場合は、度数と重みは無視されます。これにより、これらのフィールドを使用するモデルと使用しないモデルとのレベルを比較することができます。ただし、正確な評価を行うには、度数や重みのフィールドに依存しない母集団を正確に表すデータ・セットを使用する必要があります。実際には、度数または重みのフィールドの値が常にヌルまたは 1 であるテスト・サンプルを使用してモデルを評価するようにすることによって、これができます (この制限は、モデルを評価する際にのみ適用します。度数または重みの値が学習サンプルとテスト・サンプルの両方で常に 1 なら、これらのフィールドを使用する理由がそもそもありません)。
- 自動分類を使用する場合、プロフィットを基にモデルをランク付けする場合に度数を考慮する場合があります。その場合は、この手法をお勧めします。
- 必要な場合は、データ区分ノードを使用して、データを学習サンプルとテスト・サンプルに分割します。

モデル作成ノードの分析オプション

多くのモデル作成ノードには「分析」タブがあり、そこで、生スコアおよび調整済み傾向スコアとともに予測変数の重要度の情報を取得できます。

モデル評価

予測変数の重要度を計算**:** 重要度の適切な測定基準を作成するモデルの場合、モデル推定時に各予測値の相対重要度を示すグラフを表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。モデルによっては、特に大きなデータセットを使用する場合、予測変数の重要度の計算に時間がかかることがあります。そのため、一部のモデルではデフォルトでオフになっています。予測変数の重要度は、ディシジョン・リスト・モデルには使用できません。 詳しくは、 44 ページの『予測変数の重要度』を参照してください。

傾向スコア

傾向スコアは、モデル作成ノードで、またはモデル・ナゲットの「設定」タブで有効にできます。この機能は、選択された対象がフラグ型フィールドである場合にのみ使用できます。 詳しくは、トピック 36 ページの『傾向スコア』を参照してください。

未調整傾向スコアを計算**:** 生の傾向スコアは学習データだけに基づいたモデルから得られます。モデルが *true* 値 (応答する) を予測する場合、傾向は P と同じになります。ここで P は、予測値の確率です。モデルが false 値を予測する場合、傾向は (1 - P) と算出されます。

- モデルを構築する際にこのオプションを選択すると、傾向スコアはそのモデル・ナゲット内でデフォルトで有効になります。ただし、モデル作成ノードで選択したかどうかにかかわらず、モデル・ナゲット内でいつでも生の傾向スコアを有効にできます。
- モデルをスコアリングする際、生の傾向スコアは、標準の接頭辞に *RP* が追加されてフィールドに追加されます。例えば、予測値が *$R-churn* という名前のフィールドにある場合は、傾向スコア フィールドの名前は *$RRP-churn* となります。

調整済み傾向スコアを計算 **:** 生の傾向スコアはモデルによって与えられた推定値のみに基づいて算出されますが、これはオーバフィットしている可能性があり、極端に楽観的な傾向が推定されることがあります。調整済み傾向スコアは、テスト・データ区分や検証データ区分に対するモデルの成果を調べて、傾向を調整することによって、より的確な推定を行うものです。

- この設定では、ストリームに有効なデータ区分フィールドが存在している必要があります。

- 生の傾向スコアと違い、調整済み傾向スコアは、モデルを構築するときに計算されなければなりません。そうでなければ、モデル・ナゲットをスコアリングするときにそれらを使用することはできません。
- モデルをスコアリングする際、調整済み傾向スコアは、標準の接頭辞に *AP* が追加されてフィールドに追加されます。例えば、予測値が *$R-churn* という名前のフィールドにある場合は、傾向スコア フィールドの名前は *$RAP-churn* となります。調整済み傾向スコアは、ロジスティック回帰モデルには使用できません。
- 調整済み傾向スコアを計算する場合、計算に使用するテスト・データ区分または検証データ区分はバランス化されていてはいけません。そのため、上流のバランス・ノードで 「学習データのみをバランス」オプションを必ず選択します。さらに、複雑なサンプルが上流にとられた場合は、それによって調整済み傾向スコアが無効になります。
- 調整済み傾向スコアは、「ブーストされた」ツリーまたはルールセット・モデルには使用できません。詳しくは、トピック 130 ページの『ブーストされた C5.0 モデル』を参照してください。

準拠 **:** 調整済み傾向スコアが計算されるには、ストリームにデータ区分フィールドが存在していなければなりません。この計算にテスト・データ区分または検証データ区分を使用するかどうかを指定できます。最適な結果を得るには、テスト・データ区分または検証データ区分に、少なくとも、その区分が元のモデルを学習するのに使用したのと同じ数のレコードを含める必要があります。

傾向スコア

はいまたはいいえの予測値を返すモデルの場合、標準の予測値と確信度値に加えて、傾向スコアを要求できます。傾向スコアは、特定の結果または回答の尤度を示します。次の表に例があります。

表 *4.* 傾向スコア

顧客	応答の傾向
Joe Smith	35%
Jane Smith	15%

傾向スコアはフラグ型対象を持つモデルにのみ使用できます。フィールドに定義された *True* 値の尤度を、入力またはデータ型ノードで指定されたとおりに示します。

傾向スコアと確信度スコア

傾向スコアと確信度スコアは異なります。確信度スコアは現在の予測値、はい またはいいえ に適用されます。例えば、予測値がいいえ の場合、高い確信度は実際には応答しない 高い尤度を意味します。傾向スコアはこの制限を回避し、すべてのレコード間の比較を簡単にします。例えば、確信度 *0.85* のいいえ予測値は、*0.15* (または *1 − 0.85*) の生の傾向と解釈されます。

表 *5.* 確信度スコア

顧客	予測	確信度
Joe Smith	応答する	.35
Jane Smith	応答しない	.85

傾向スコアの取得

- 傾向スコアは、モデル作成ノードの「分析」タブまたはモデル・ナゲットの「設定」タブで有効にできます。この機能は、選択された対象がフラグ型フィールドである場合にのみ使用できます。 詳しくは、トピック 35 ページの『モデル作成ノードの分析オプション』を参照してください。

- 傾向スコアは、使用するアンサンブル手法によっては、アンサンブル・ノードによって計算されます。

調整済み傾向スコアの計算

調整済み傾向スコアは、モデル構築のプロセスの一部として計算され、そのほかでは使用できません。モデルが構築されると、テスト・データ区分または検証データ区分からのデータを使用してスコアリングされます。調整済み傾向スコアを算出する新しいモデルは、そのデータ区分に対する元のモデルの精度を分析することによって構築されます。モデルの種類によって、2 つの手法のうちいずれかが調整済み傾向スコアの計算に使用されます。

- ルール・セットおよびツリー・モデルの場合、調整済み傾向スコアは、各ツリー・ノードで各カテゴリーの度数 (ツリー・モデルの場合) または各ルールのサポートおよび確信度 (ルール・セット・モデルの場合) を再計算することによって生成されます。新しいルール・セットまたはツリー・モデル内の結果は元のモデルとともに格納され、調整済み傾向スコアが要求されるたびに使用されます。元のモデルが新しいデータに適用されるたびに、新しいモデルが生の傾向スコアに適用されて調整済みスコアが生成されます。
- そのほかのモデルの場合、元のモデルをテスト・データ区分または検証データ区分でスコアリングすることによって生成されたレコードは、それぞれの生の傾向スコアごとに分割されます。次に、各ビン内の平均生傾向から同じビン内の観測傾向へマップする非線型関数を定義するニューラル・ネットワーク・モデルが学習されます。ツリー・モデルに関して前述したように、結果のニューラル・ネットワーク・モデルは元のモデルとともに格納され、調整済み傾向スコアが要求されるたびに生の傾向スコアに適用されます。

テスト用データ区分の欠損値に関する注意 **:**テスト/検証用データ区分の欠損入力値の処理は、モデルによって異なります (詳細は、各モデルのスコアリング・アルゴリズムを参照してください)。C5 モデルは、欠損入力値がある場合、調整済み傾向スコアを計算することはできません。

誤分類コスト

状況によっては、特定の誤りコストが他の誤りコストに比べて高いことがあります。例えば、信用リスクの高い申請者を低リスクに分類した場合 (ある種の誤分類) のコストは、リスクの低い申請者を高リスクに分類した場合 (別種の誤分類) よりも高くなります。誤分類コストでは、さまざまな予測の誤りに対し、相対的な重要度を指定できます。

誤分類コストは、基本的には、特定の結果に対して適用される重みです。これらの重みは、モデルに組み込まれ、(コストの高い誤りを防ぐための手段として) 実際に予測値に影響する場合があります。

C5.0 モデルを例外として、誤分類コストは、モデルのスコアリング時には適用されず、自動分類ノード、評価グラフ、または精度分析ノードを使用してモデルをランク付けまたは比較する場合には考慮されません。コストを含むモデルは、コストを含まないモデルに比べてエラーが少なく、全体の精度の項目で高くランク付けされません。ただし、コストが少ない エラーにより組み込まれたバイアスがあるため、実際の問題でパフォーマンスが優れる場合があります。

コスト行列には、可能な各予測カテゴリーや実際のカテゴリーの組み合わせのコストが表示されます。デフォルトでは、すべての誤分類コストが 1.0 に設定されています。コストの値を自分で入力するには、「誤分類コストを使用」 を選択して、コスト行列に独自の値を入力します。

誤分類コストを変更するには、目的の予測値と実際の値の組み合わせに対応するセルを選択して、セルの内容を削除してから、適切なコストを入力してください。コストは自動的に対称的にはなりません。例えば *A* を *B* として誤分類した場合のコストを 2.0 に設定しても、*B* を *A* として誤分類した場合のコストは、変更しない限りデフォルト値 (1.0) のまま変わりません。

注: ディシジョン ツリー・モデルのみで構築時のコストを指定できます。

モデル・ナゲット

図 *19.* モデル・ナゲット

モデル・ナゲットは、モデルのコンテナーです。つまり、SPSS Modeler のモデル作成操作の結果を示すルール、式または方程式のセットです。ナゲットの主な目的は、データをスコアリングし、予測を生成、またはモデルのプロパティーの詳細な分析を可能にすることです。画面でモデル・ナゲットを開くと、モデル作成時の入力フィールドの相対重要度など、モデルに関する様々な詳細情報を表示できます。予測変数を表示するには、高度なプロセス・ノードまたは出力ノードを接続および実行する必要があります。 詳しくは、トピック 49 ページの『ストリーム内でのモデル・ナゲットの使用』を参照してください。

図 *20.* モデル作成ノードからモデル・ナゲットへのモデル・リンク

モデル作成ノードを正常に実行すると、対応するモデル・ナゲットがストリーム・キャンバスに置かれ、金色のダイヤの形のアイコンで表示されます (名前は「ナゲット」)。ストリーム・キャンバスに、モデル作成ノードの前に最も近い適切なノードへの接続およびモデル作成ノード自体へのリンク (点線) と共にナゲットが表示されます。

またナゲットは、IBM SPSS Modeler ウィンドウの右上隅にある「モデル」パレット内に表示されます。いずれかの場所から、ナゲットを選択して参照し、モデルの詳細を表示できます。

モデル作成ノードが正常に実行されると、ナゲットは常に「モデル」パレットに表示されます。ナゲットをさらにストリーム・キャンバスに投入するかどうかを制御するユーザー・オプションを設定できます。

IBM SPSS Modeler のモデル・ナゲットの使用に関する情報は、次の各トピックを参照してください。使用されているアルゴリズムの詳細は、製品ダウンロードの一部として PDF ファイルで入手可能な「*IBM SPSS Modeler* アルゴリズム・ガイド」を参照してください。

モデル・リンク

デフォルトでは、ナゲットを作成したモデル作成ノードへのリンクと共に、ナゲットがキャンバス内に表示されます。これは、複数のナゲットを含む複雑なストリームで役に立ち、モデル作成ノードによって更新されるナゲットを特定できます。各リンクには、モデル作成ノードを実行するときにモデルを置き換えるかどうかを示す記号が表示されます。 詳しくは、トピック 40 ページの『モデルの置換』を参照してください。

モデル・リンクの定義および削除

キャンバス上のリンクを手動で定義および削除できます。新しいリンクを定義すると、カーソルがリンク カーソルに変わります。

図 *21.* リンク カーソル

新しいリンクの定義 (コンテキスト・メニュー)

1. リンクを開始するモデル作成ノードを右クリックします。
2. コンテキスト・メニューから 「モデル リンクを定義」 を選択します。
3. リンクを終了するナゲットをクリックします。

新しいリンクの定義 (メイン・メニュー)

1. リンクを開始するモデル作成ノードをクリックします。
2. メイン・メニューから次の各項目を選択します。

   「編集」 > 「ノード」 > 「モデル リンクを定義」
3. リンクを終了するナゲットをクリックします。

既存のリンクの削除 (コンテキスト・メニュー)

1. リンクの終点となるナゲットを右クリックします。
2. コンテキスト・メニューから 「モデル リンクを削除」 を選択します。

または、次のように指定します。

1. リンクの中間にある記号を右クリックします。
2. コンテキスト・メニューから 「リンクを削除」 を選択します。

既存のリンクの削除 (メイン・メニュー)

1. リンクを削除するモデル作成ノードまたはナゲットをクリックします。
2. メイン・メニューから次の各項目を選択します。

   「編集」 > 「ノード」 > 「モデル リンクを削除」

モデル・リンクのコピーと貼り付け

モデル作成ノードを除いて、リンクしたナゲットをコピーして同じストリームに貼り付けると、ナゲットがモデル作成ノードへのリンクと共に貼り付けられます。新しいリンクは、元のリンクと同じモデル置換状態になります (40 ページの『モデルの置換』を参照)。

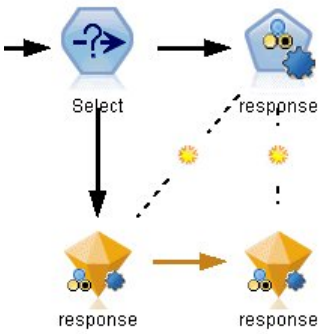


図 22. リンクしたナゲットのコピーと貼り付け

ナゲットをリンクしたモデル作成ノードと一緒にコピーして貼り付けると、オブジェクトが貼り付けられるのが同じストリームであっても新しいストリームであっても、リンクは保持されます。

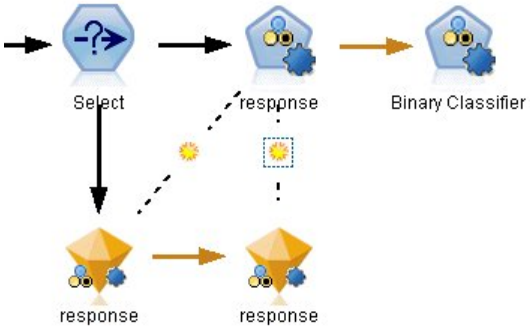


図 23. リンクしたナゲットのコピーと貼り付け

注 : モデル作成ノードを除いて、リンクしたナゲットをコピーして新しいストリーム (またはモデル作成ノードを含まないスーパーノード) に貼り付けると、リンクが解除され、ナゲットのみが貼り付けられます。

モデル・リンクおよびスーパーノード

リンクしたモデルのモデル作成ノードまたはモデル・ナゲット (両方ではない) を含むスーパーノードを定義すると、リンクは解除されます。スーパーノードを拡張してもリンクは保持されません。リンクを保持できる方法は、スーパーノードの作成を取り消すことだけです。

モデルの置換

ナゲットを作成したモデル作成ノードの再実行時に既存のナゲットを置き換える (更新する) かどうかを選択できます。置換オプションを無効にすると、モデル作成ノードの再実行時に新しいナゲットが作成されます。

モデル作成ノードからナゲットへの各リンクには、モデル作成ノードを再実行するときにモデルを置き換えるかどうかを示す記号が表示されます。

図 *24.* モデル置換オプションが有効なモデル・リンク

このリンクは最初、モデルの置換が有効になった状態で表示されます。リンクに日光の記号が示されます。この状態で、リンクの一方の端でモデル作成ノードを再実行すると、もう一方の端のナゲットが更新されます。

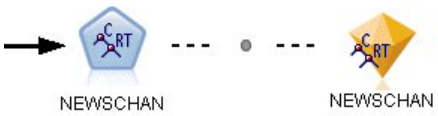


図 *25.* モデル置換オプションが無効なモデル・リンク

モデル置換が無効な場合、リンクの記号が灰色の点に変わります。この状態で、リンクの一方の端でモデル作成ノードを再実行すると、新しく、更新されたバージョンのナゲットがキャンバスに追加されます。

いずれかの場合、「前のモデルを置換」システム・オプションの設定に応じて、「モデル」パレットの既存のナゲットが更新、または新しいナゲットが追加されます。

実行の順序

複数のブランチにモデル・ナゲットが含まれるストリームを実行する場合、ストリームをまず評価して、モデルの置換が有効なブランチがモデル・ナゲットを使用するブランチの前に実行されることを確認します。

要件がより複雑な場合、スクリプトを使用して手動で実行を順序を設定できます。

モデル置換設定の変更

1. リンク上のシンボルを右クリックします。
2. 必要に応じて、「モデル置換を有効 **(**無効**)** にする」 を選択します。

注**:** モデル・リンクのモデル置換設定は、「ユーザー オプション」ダイアログの通知タブの設定 (「ツール」>「オプション」>「ユーザー オプション」) より優先されます。

モデル・パレット

モデル・パレット (マネージャー・ウィンドウの「モデル」タブ) から、モデル・ナゲットの使用、調査、および変更をさまざまな方法で行うことができます。

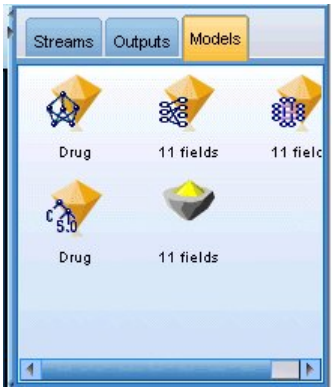


図 *26.* モデル・パレット

モデル・パレット内のモデル・ナゲットを右クリックすると、コンテキスト・メニューが表示されます。次のオプションがあります。

- ストリームに追加**:** モデル・ナゲットを現在アクティブなストリームに追加します。ストリーム内に選択されているノードがある場合、モデル・ナゲットはそのノードと接続されます (接続できる場合)。そうでない場合、使用できる最も近いノードに接続されます。ノードがストリームにある場合、ナゲットがモデルを作成したモデル作成ノードにリンクした状態で表示されます。
- 参照**:** ナゲットのモデル・ブラウザーを開きます。
- 名前の変更と注釈 **:** モデル・ナゲットの名前を変更したり、ナゲットの注釈を変更したりできます。
- モデル作成ノードの生成**:** モデル・ナゲットの変更または更新が必要だが、モデル作成に使用されたストリームが利用できない場合は、このオプションを使用して、最初のモデル作成に使用されたのと同じオプションで、モデル作成ノードを再生成できます。
- モデルを保存、モデルに名前を付けて保存**:** モデル・ナゲットを生成されたモデル (.gm) のバイナリファイルに保存します。
- モデルを格納**:** モデル・ナゲットを IBM SPSS Collaboration and Deployment Services Repository に保存します。
- **PMML** をエクスポート**:** PMML (Predictive Model Markup Language) としてモデル・ナゲットをエクスポートします。PMML は、IBM SPSS Modeler 外部の新規データのスコアリングに使用できます。「**PMML** をエクスポート」 は、生成されたすべてのモデル・ノードで使用できます。
- プロジェクトに追加**:** モデル・ナゲットを保存して、それを現在のプロジェクトに追加します。「クラス」タブで、ナゲットは「生成されたモデル」フォルダーに追加されます。「CRISP-DM」タブでは、ナゲットはデフォルトのプロジェクト・フェーズに追加されます。
- 削除**:** モデル・ナゲットをパレットから削除します。

モデル・パレット内の空の領域を右クリックすると、コンテキスト・メニューが表示されます。次のオプションがあります。

- モデルを開く **:** 前に IBM SPSS Modeler で作成されたモデル・ナゲットをロードします。
- モデルを取得**:** IBM SPSS Collaboration and Deployment Services リポジトリーから、保存されたモデルを取得します。
- パレットのロード **:** 保存されているパレットを外部ファイルからロードします。
- パレットの取得**:** IBM SPSS Collaboration and Deployment Services リポジトリーから、保存されたモデル・パレットを取得します。

- パレットの保存**:** 生成されたモデル・パレットの外部ファイル (.gen) にモデル・パレットの内容全体を保存します。
- パレットを格納**:** IBM SPSS Collaboration and Deployment Services リポジトリーにモデル・パレットの内容全体を保存します。
- パレットを消去**:** すべてのナゲットをパレットから削除します。
- パレットをプロジェクトに追加**:** モデル・パレットを保存して、それを現在のプロジェクトに追加します。「クラス」タブで、ナゲットは「生成されたモデル」フォルダーに追加されます。「CRISP-DM」タブでは、ナゲットはデフォルトのプロジェクト・フェーズに追加されます。
- **PMML** をインポート **:** 外部ファイルからモデルをロードします。IBM SPSS Statistics またはこの形式をサポートする他のアプリケーションで作成された PMML モデルを開いたり、参照、スコアリングを行うことができます。 詳しくは、トピック 50 ページの『PMML としてのモデルのインポートおよびエクスポート』を参照してください。

モデル・ナゲットの参照

モデル・ナゲット・ブラウザーを使用して、モデルの結果を検証したり使用したりできます。ブラウザーから、生成されたモデルの保存、印刷、またはエクスポート、およびモデル要約の検討、モデルの注釈の表示または編集などの作業を行うことができます。モデル・ナゲットの種類によっては、フィルター・ノードやルールセット・ノードのような新規ノードを作成することもできます。さらに一部のモデルでは、ルールやクラスター中心などの、モデル・パラメーターを表示することもできます。モデルの種類によっては (ツリー・ベースのモデルとクラスター・モデル)、モデル構造をグラフィカルに表示することもできます。モデル・ナゲット・ブラウザーの使用方法を、次に説明していきます。

メニュー

「ファイル」メニュー **:** すべてのモデル・ナゲットには「ファイル」メニューがあります。このメニューには、次のオプションのサブセットがあります。

- ノードの保存**:** 生成されたモデル・ナゲット (.nod) をファイルに保存します。
- ノードの保管**:** モデル・ナゲットを IBM SPSS Collaboration and Deployment Services リポジトリーに保存します。
- ヘッダーとフッター**:** ナゲットから印刷するページのヘッダーやフッターを編集できます。
- ページ設定**:** ナゲットから印刷するページの設定を変更できます。
- 印刷プレビュー**:** ナゲットがどのように印刷されるかをプレビュー表示します。サブメニューから、プレビューする情報を選択してください。
- 印刷**:** ナゲットの内容を印刷します。サブメニューから、印刷する情報を選択してください。
- ビューを印刷**:** 現在のビューまたはすべてのビューを印刷します。
- テキストのエクスポート**:** ナゲットの内容をテキスト・ファイルにエクスポートします。サブメニューから、エクスポートする情報を選択してください。
- **HTML** 生成**:** ナゲットの内容を HTML ファイルにエクスポートします。サブメニューから、エクスポートする情報を選択してください。
- **PMML** をエクスポート**:** このファイルは、他の PMML 互換ソフトウェアで利用することができます。詳しくは、トピック 50 ページの『PMML としてのモデルのインポートおよびエクスポート』を参照してください。
- **SQL** のエクスポート**:** モデルを SQL としてエクスポートし、編集し、その他のデータベースとともに使用することができます。

注: SQL のエクスポートは、次のモデルでのみ有効です。C5、C&R Tree、CHAID、QUEST、線型、ロジスティック回帰、ニューラル・ネットワーク、因子/主成分分析、およびディシジョン・リスト・モデル。

- サーバー・スコアリング・アダプタ向けに公開: モデルをスコアリング・アダプターがインストールされたデータベースに公開し、データベース内でモデルのスコアリングを実行できるようにします。 詳しくは、トピック 53 ページの『スコアリング・アダプター向けにモデルを公開』を参照してください。

「ノードの生成」メニュー : 大部分のモデル・ナゲットには、「ノードの生成」メニューもあります。大部分のモデル・ナゲットには「生成」メニューもあり、モデル・ナゲットに基づいて新しいノードを生成することができます。このメニューで利用できるオプションは、参照しているモデルの種類によって異なります。特定のモデルから生成できるノードの詳細は、各モデル・ナゲット・タイプを参照してください。

「表示」メニュー: ナゲットの「モデル」タブで、このメニューを使用すると、現在のモードで使用できるさまざまな視覚化ツールバーを表示または非表示にできます。完全なセットのツールバーを使用できるようにするには、「一般」ツールバーから「編集モード」(刷毛のアイコン) を選択します。

「プレビュー」ボタン : 一部のモデル・ナゲットには「プレビュー」ボタンがあります。これを使用すると、モデル作成プロセスで作成された追加フィールドなど、モデル・データのサンプルを表示できます。表示される行のデフォルト数は 10 行です。ただし、ストリームのプロパティーで変更できます。

「現在のプロジェクトに追加」ボタン: モデル・ナゲットを保存して、それを現在のプロジェクトに追加します。「クラス」タブで、ナゲットは「生成されたモデル」フォルダーに追加されます。「CRISP-DM」タブでは、ナゲットはデフォルトのプロジェクト・フェーズに追加されます。

モデル・ナゲットの要約/情報

モデル・ナゲットの「要約」タブまたは「情報」ビューには、フィールド、構築の設定、およびモデル推定プロセスについての情報が表示されます。結果は、特定の項目をクリックすると開いたり閉じたりできるツリーで表示されます。

精度分析: モデルについての情報を表示します。特定の詳細はモデル タイプによって異なり、各モデル・ナゲットについてのセクションで説明されています。また、このモデル作成ノードに接続されている精度分析ノードを実行した場合、その精度分析情報もこのセクションに表示されます。

フィールド: 対象フィールドおよびモデル構築時の入力として使われるフィールドが表示されます。分割モデルの場合、分割を指定されたフィールドも一覧表示されます。

注: ニューラル ネットワーク モデル、線形モデル、およびブースティングまたはバギングいずれかのモードを使用するその他のモデルの「情報」ビューで表示されるアイコンは、型がフラグ、名義、または順序のいずれであるかに関係なく、同じアイコンです (名義型アイコン)。

構築の設定 / オプション: モデル構築時に使われる設定に関する情報が含まれます。

学習の要約 : モデルの種類、モデルの作成に使われたストリーム、モデルの作成者、モデルの作成日時、およびモデル構築の経過時間などの情報が表示されます。モデル構築の経過時間は、「要約」タブにのみ表示され、「情報」ビューには表示されないことに注意してください。

予測変数の重要度

一般にモデリングの作業では、最も重要な予測フィールドの編集に集中して取り組み、最も重要でない予測フィールドは削除するか無視してしまいたくなるものです。予測値の重要度グラフを使用すると、モデル推

定時に各予測値の相対重要度が示されるので便利です。値が相対的であるため、表示されるすべての予測変数の値の合計は 1.0 となります。予測値の重要度は、モデルの精度に関連しません。予測が正確かどうかに関係なく、予測時における各予測の重要度にのみ関連します。

予測変数の重要度は、重要度の適切な統計的尺度を生成するモデルで使用することができます。そのようなモデルには、ニューラル・ネットワーク、ディシジョン・ツリー (C&R Tree、C5.0、CHAID、および QUEST)、Bayesian network (ベイズ)、判別分析、SVM、SLRM モデル、線型回帰、ロジスティック回帰、一般化線型モデル、および最近傍 (KNN) モデルがあります。これらの多くのモデルについて、予測変数の重要度モデル作成ノードの「分析」タブで有効にできます。詳しくは、トピック 35 ページの『モデル作成ノードの分析オプション』を参照してください。KNN モデルの詳細は、 367 ページの『近傍』を参照してください。

注 : 予測変数の重要度は、分割モデルにサポートされていません。分割モデルの作成時、予測変数の重要度の設定は無視されます。詳しくは、トピック 28 ページの『分割モデルの作成』を参照してください。

予測変数の重要度の計算には、特に大きなデータセットを使用する場合、モデル構築よりもずっと長い時間がかかることがあります。SVM およびロジスティック回帰の場合、他のモデルに比べて時間がかかるため、これらのモデルではデフォルトで無効になっています。多数の予測値を使用したデータセットを使用する場合、特徴量選択ノードを使用した最初のスクリーニングでより速くなる可能性があります (下記を参照)。

- 予測変数の重要度は、テスト・データ区分から計算されます (可能な場合)。そうでなければ、学習データが使用されます。
- SLRM モデルでも予測変数の重要度は使用できますが、SLRM アルゴリズムによって計算されます。詳しくは、トピック 352 ページの『SLRM モデル・ナゲット』を参照してください。
- IBM SPSS Modeler のグラフ ツールを使用して、グラフを対話的に処理、編集、保存できます。
- オプションで、予測変数の重要度グラフ内の情報を基にフィルター・ノードを生成することもできます。 詳しくは、トピック 46 ページの『重要度を基にした変数のフィルタリング』を参照してください。

予測変数の重要度と特徴量選択

モデル・ナゲットに表示される予測変数の重要度グラフは、特徴量選択ノードと同様の結果を生成するように思われるかもしれません。特徴量選択は各入力フィールドを選択された対象との関係の強さに基づいて、他の入力値とは独立してランク付けする一方、予測変数の重要度グラフはこの 特定のモデルに関する各入力値の相対的な重要度を示します。そのため、特徴量選択は入力値のスクリーニングにおいて、より安全策を取ることになります。例えば、役職と仕事のカテゴリーがどちらも給料と密接に関係している場合、特徴量選択では両方ともが重要であると示されます。しかし、モデル作成においては、相互作用と相関も考慮されます。そのため、2 つの入力値の情報が重複する場合は、そのうちの 1 つだけが使用されることに気づかれるでしょう。実際には、特徴量選択は、数多くの変数がある大きなデータセットの処理での予備的スクリーニングに最も便利で、予測変数の重要度はモデルの微調整により便利です。

単一モデルと自動化モデル作成ノードとの間の予測変数の重要度の違い

個別のノードから単一モデルを作成するか、自動化モデル作成ノードを作成して結果を生成するかによって、予測変数の重要度に微細な違いがある場合があります。このような実装での違いは、いくつかの技術的な制約事項によるものです。

例えば、CHAID などの単一の分類子を使用すると、重要度の値を計算するときに、計算に停止規則が適用され、確率値が使用されます。一方、自動分類子は停止規則を使用せず、計算で予測ラベルを直接使用しま

す。これらの違いにより、自動分類を使用して単一モデルを生成すると、重要度の値は、単一の分類子で計算されたものに比べて、おおまかな推定と考えられる可能性があります。最も正確な予測変数の重要度を取得するためには、自動化モデル作成ノードの代わりに単一ノードを使用することをお勧めします。

重要度を基にした変数のフィルタリング

オプションで、予測変数の重要度グラフ内の情報を基にフィルター・ノードを生成することもできます。

該当する場合、グラフ内に含めたい予測値にマークをつけます。そして、メニューから次の各項目を選択します。

「生成」 > 「フィルター ノード **(**予測変数の重要度**)**」

または

> 「フィールド選択 **(**予測変数の重要度**)**」

「変数のトップ数」。指定した上位数までの重要度を有する予測値を含めるか、または除外します。

「次より大きな重要度」。指定した値よりも相対的に重要度が大きい予測値をすべて含めるか、または除外します。

アンサンブル・ビューアー

アンサンブルのモデル

アンサンブルのモデルは、アンサンブル内のコンポーネントモデルの情報、およびアンサンブル全体のパフォーマンスの情報を提供します。

メインの (ビューに依存しない） ツールバーにより、スコア付けにアンサンブルを使用するか、または参照モデルを使用するかを選択できます。スコア付けにアンサンブルを使用する場合、結合ルールも選択できます。この変更にはモデルの再実行は不要です。ただし、選択内容はモデル (ナゲット) に保存され、スコア付けまたは下流のモデル評価、またはその両方に使用されます。また、アンサンブル・ビューアーからエクスポートされた PMML にも影響を与えます。

「結合規則」。アンサンブルのスコアリング時に、この規則を使用して基本モデルの予測値を結合し、アンサンブル・スコア値を計算します。

- カテゴリー目標に対するアンサンブル予測値は、投票、確率が最も高いもの、または平均値の確率が最も高いものを組み合わせることができます。「票決」は、基本モデルで最も頻繁であり、最も確率が高いカテゴリーを選択します。「高確率」は、すべての基本モデルで最も高い単独の確率に達したカテゴリーを選択します。「最高平均確率」は、基本モデルでカテゴリーの確率が平均化された場合の、最も値の高いカテゴリーを選択します。
- 連続目標に対するアンサンブル予測値は、基本モデルから予測値の平均値や中央値を使用して結合できます。

デフォルト値は、モデル構築時の仕様に基づき設定されます。結合ルールを変更すると、モデルの精度を再計算し、モデル精度のすべてのビューを更新します。「予測変数の重要度」グラフも更新されます。スコア付けに参照モデルが選択されている場合、このコントロールは無効になります。

「すべての結合規則を表示」。選択した場合、使用可能なすべての結合規則の結果がモデル品質グラフに表示されます。「コンポーネント・モデルの精度」グラフは、各投票方式の参照ラインを示すように更新されます。

モデルの要約: 「モデルの要約」ビューはスナップショットで、アンサンブルの品質とその多様性が一目でわかる要約です。

「品質」。グラフには、参照モデルおよび naive モデルと比較した最終モデルの精度が表示されます。精度は、大きく表示されているものがより適切な形式であることを示し、「最適な」モデルの精度が最も高いことを示します。カテゴリー目標では、精度は予測値が観測値と一致したレコードの割合で示されます。連続目標では、精度は、1 から予測の平均絶対誤差 (予測値から観測値を引いた値の絶対値の平均) を引いた値から、予測値の範囲 (最大予測値から最小予測値を引いた値） となります。

バギング・アンサンブルでは、学習分割全体に構築された標準モデルが参照モデルとなります。ブースティング・アンサンブルでは、最初のコンポーネント・モデルが参照モデルとなります。

モデルが構築されておらず、すべてのレコードが最頻カテゴリーに割り当てられている場合、naive モデルが精度と示します。naive モデルは連続目標では計算されません。

「多様性」。アンサンブルを構築するために使用されたコンポーネント・モデル間の「意見の多様性」がグラフに表示されます。値が大きいほど多様性が大きいことを示します。これは、基本モデル内で予測値にどの程度ばらつきがあるかを示す指標です。ブースティング・アンサンブル・モデルでは多様性は利用できません。また、連続目標では表示されません。

予測変数の重要度: 一般にモデリングの作業では、最も重要な予測フィールドの編集に集中して取り組み、最も重要でない予測フィールドは削除するか無視してしまいたくなるものです。予測値の重要度グラフを使用すると、モデル推定時に各予測値の相対重要度が示されるので便利です。値が相対的であるため、表示されるすべての予測変数の値の合計は 1.0 となります。予測値の重要度は、モデルの精度に関連しません。予測が正確かどうかに関係なく、予測時における各予測の重要度にのみ関連します。

予測変数の重要度は、すべてのアンサンブル・モデルで利用できるわけではありません。予測値セットはコンポーネント・モデルによって異なりますが、重要度は少なくとも 1 つのコンポーネント・モデルで使用された予測値を元に計算されます。

予測値の頻度: 予測値セットは、モデリング方式の選択または予測値の選択により、コンポーネント・モデルごとに異なる場合があります。予測値の頻度の作図は、アンサンブル内のコンポーネント・モデルにおける予測値の分布を示す点の作図です。各点は、予測値を含む 1 個以上のコンポーネント・モデルです。予測値は y 軸に作図され、度数の降順で並べられます。よって、最上位の予測値は最も多くのコンポーネント・モデルで使用されている予測値で、最下位の予測値は最も少ないコンポーネント・モデルで使用されている予測値です。上位 10 個の予測値が表示されます。

最も頻繁に表示される予測値が、一般的に最も重要な予測値です。この作図は、コンポーネント・モデル間で予測値が分散しない方式には適していません。

コンポーネント・モデルの精度: グラフは、コンポーネント・モデルの精度を予測した点を作図したものとなります。各点は、y 軸に作図された精度のレベルを含む 1 個以上のコンポーネント・モデルです。マウス・ポインタを点の上に停止させると、対応するそれぞれのコンポーネント・モデルの情報が表示されます。

基準線: 作図で、アンサンブル、参照モデル、および Naive モデルの色コード化された線が表示されます。スコア付けに使用されるモデルに対応する線の隣には、チェックマークが表示されます。

「対話性」。結合規則を変更すると、グラフは更新されます。

「ブースティング・アンサンブル」。ブースティング・アンサンブルの折れ線グラフが表示されます。

コンポーネント・モデルの詳細: このテーブルには、コンポーネント・モデルの情報が 1 行ずつ表示されます。デフォルトでは、コンポーネント・モデルはモデル番号の昇順に並べられます。行は、任意の列の値の昇順または降順に整列できます。

「モデル」。コンポーネント・モデルが作成された順序を表す番号。

精度: 全体的な精度をパーセント単位で表したもの。

方法モデリングの方法です。

「予測値」。コンポーネント・モデルで使用されている予測値の数。

モデル・サイズ: モデルのサイズはモデリングの方法によって異なります。ツリーの場合はツリー内のノード数、線形モデルの場合は係数の数、ニューラル・ネットワークの場合はシナプスの数になります。

レコード: 学習サンプルの入力レコードの重み付きの数。

データの自動準備:

このビューには、データの自動準備 (ADP) ステップで、どのフィールドが除外されたか、また変換されたフィールドがどのように派生したかについての情報が表示されます。テーブルには変換または除外されたフィールドごとに、フィールド名、分析での役割、および ADP ステップで実行されたアクションがリストされます。フィールドは、フィールド名のアルファベット順 (昇順) に並べ替えられます。

アクションの「外れ値の削除」が表示されている場合、分割値を超える連続型予測フィールドの値 (平均値からの標準偏差が 3) は分割値に設定されていることを示します。

分割モデルのモデル・ナゲット

分割モデルのモデル・ナゲットには、分割によって作成されたすべての個別モデルにアクセスができます。

分割モデル・ナゲットには次のものが含まれます。

- 作成された分割モデル、各モデルの一連の統計を表示したリスト
- モデル全体の情報

分割モデルのリストで、各モデルを開いて詳細に検証することができます。

分割モデル・ビューアー

「モデル」タブにはナゲットに含まれるすべてのモデルが表示され、分割モデルに関する統計をさまざまな形式で提供します。モデル作成ノードに応じて、2 つの一般形式をとります。

ソート項目。モデルを表示する順序を選択します。表示列のいずれかの値に基づいて、昇順または降順でリストをソートできます。また、列の見出しをクリックして、該当する列ごとにリストをソートできます。デフォルトは全体の精度の降順です。

列メニューの表示/非表示 :各列を表示するかまたは表示しないかを選択できるメニューを表示します。

表示 :データ区分を使用している場合、学習データまたは検定データのいずれかの結果を表示することができます。

各分割について、次のような詳細が表示されます。

グラフ :このモデルのデータ分布を示すサムネイル。ナゲットがキャンバス上にある場合、サムネイルをダブルクリックすると、フルサイズのグラフが開きます。

モデル :モデル タイプのアイコン。アイコンをダブルクリックすると、この特定の分割のモデル・ナゲットを開きます。

分割フィールド :さまざまな値を持つ、モデル作成ノードで分割フィールドとして指定されたフィールド。

分割内のレコード数 :この特定の分割にあるレコード数。

使用されるフィールド数 :使用された入力フィールドの数に基づいて、分割モデルをランク付けします。

全体の精度 **(%)** :該当する分割内の全レコード数に関連する分割モデルを正確に予測する、レコードの割合のことです。

分割。列のヘッダーは、分割の作成に使用するフィールドを示します。セルは、分割された値となります。その分割で構築されたモデルをモデル・ビューアーを開くには、分割をダブルクリックします。

精度**:** 全体的な精度をパーセント単位で表したもの。

モデル・サイズ**:** モデルのサイズはモデリングの方法によって異なります。ツリーの場合はツリー内のノード数、線形モデルの場合は係数の数、ニューラル・ネットワークの場合はシナプスの数になります。

レコード**:** 学習サンプルの入力レコードの重み付きの数。

ストリーム内でのモデル・ナゲットの使用

モデル・ナゲットは、ストリーム内に置かれ、新規データのスコアリングや新規ノードの生成に使用できます。データのスコアリングを行うと、モデルの構築で得られた情報を使用して、新規レコードの予測値を作成できます。スコアリングの結果を表示するには、ターミナル・ノード (処理ノードまたは出力ノード) をナゲットに接続し、ターミナル・ノードを実行する必要があります。

モデルによっては、確信度やクラスターの中心からの距離など、予測の品質に関する追加情報を、モデル・ナゲットから得ることができます。新規ノードを生成すると、生成されたモデルの構造を基にして簡単に新規ノードを生成できます。例えば、入力フィールドの選択を行う大部分のモデルでは、モデルが重要と識別した入力フィールドのみを通すフィルター・ノードを生成できます。

注**:** IBM SPSS Modeler の複数のバージョンで実行すると、指定したモデルによって特定のケースに割り当てられるスコアには、若干の違いが生じることがあります。通常、これは、バージョン間でソフトウェアが拡張された結果として起こります。

モデル・ナゲットをデータのスコアリングに使用するには

1. モデル・ナゲットをデータ・ソースまたはそのナゲットにデータを渡すストリームに接続します。
2. 1 つ以上の処理ノードまたは出力ノード (テーブル・ノード や精度分析ノードなど) をモデル・ナゲットに追加するか接続します。
3. モデル・ナゲットから、下流にあるノードの 1 つを実行します。

注 : データのスコアリングに未調整ルール・ノードを使用することはできません。データをアソシエーション・ルール・モデルに基づいてスコアリングする場合は、未調整ルール・ノードを使用してルール・セット・ナゲットを生成し、そのルール・セット・ナゲットを使用してスコアリングを行います。詳しくは、トピック 284 ページの『アソシエーション・モデル・ナゲットからルールセットを生成する』を参照してください。

モデル・ナゲットを処理ノードの生成に使用するには

1. パレットでモデルを参照するか、ストリーム キャンバスでモデルを編集します。
2. モデル・ナゲットのブラウザー・ウィンドウの「ノードの生成」メニューから適切なノードの種類を選択します。使用できるオプションは、モデル・ナゲットの種類によって異なります。特定のモデルから生成できるノードの詳細は、各モデル・ナゲット・タイプを参照してください。

モデル作成ノードの再生成

モデル・ナゲットの変更または更新が必要だが、モデル作成に使用されたストリームが利用できない場合は、最初のモデル作成に使用されたオプションと同じオプションで、モデル作成ノードを再生成できます。

モデルを再構築するには、モデル・パレット内で目的のモデルを右クリックし、「モデル作成ノードの生成」を選択します。

または、モデルの参照時に、「ノードの生成」メニューから 「モデル作成ノードの生成」を選択します。

再生成されたモデル作成ノードは、多くの場合、基のモデルを作成するために使用されたノードと機能的に一致する必要があります。

- ディシジョン・ツリー・モデルの場合、インタラクティブ セッションの間に指定した追加の設定をノードとともに保存することもできます。また、再生成されたモデル作成ノードで、「ツリー ディレクティブを使用」オプションが有効になります。
- ディシジョン・リスト・モデルの場合「保存済みインタラクティブ セッション情報の使用」オプションが有効です。 詳しくは、トピック 160 ページの『ディシジョン・リストのモデル関連のオプション』を参照してください。
- 時系列モデルの場合、「既存のモデルを使用して推定を続行」オプションが有効で、以前のモデルを現在のデータで再生成することができます。詳しくは、トピック時系列のモデル・オプションを参照してください。

PMML としてのモデルのインポートおよびエクスポート

PMML (Predictive Model Markup Language) は、モデルへの入力、データ・マイニングのデータの準備に使用する返還、モデル自体を定義するパラメーターなど、データ・マイニングおよび統計モデルを説明する XML 形式です。IBM SPSS Modeler は PMML をインポートおよびエクスポートし、IBM SPSS Statistics など、この形式をサポートする他のアプリケーションとモデルを共有できるようにします。

PMML の詳細は、データ・マイニング・グループの Web サイト (*http://www.dmg.org*) を参照してください。

モデルをエクスポートするには

PMML エクスポートでは、IBM SPSS Modeler 内で生成されたほとんどの種類のモデルがサポートされます。 詳しくは、トピック 51 ページの『PMML をサポートするモデルの種類』を参照してください。

1. モデル・パレットのモデル・ナゲットを右クリックします(または、キャンバス上のモデル・ナゲットをダブルクリックして、「ファイル」メニューを選択します)。
2. メニューで、「**PMML** をエクスポート」 をクリックします。
3. 「エクスポート」(または「保存」) ダイアログ・ボックスで、対象ディレクトリーとモデルの一意の名前を指定します。

注**:**

「ユーザー オプション」ダイアログ・ボックスで、PMML エクスポートのオプションを変更できます。メイン・メニューで次の各項目をクリックします。

「ツール」 > 「オプション」 > 「ユーザー オプション」

そこで「PMML」タブをクリックします。

PMML として保存されたモデルをインポートするには

IBM SPSS Modeler または別のアプリケーションから PMML としてエクスポートされたモデルは、生成済みモデル・パレットへインポートできます。 詳しくは、トピック『PMML をサポートするモデルの種類』を参照してください。

1. モデル・パレット内で、パレットを右クリックし、メニューから「**PMML** をインポート」を選択します。
2. インポートするファイルを選択し、必要に応じて、変数のラベルに関するオプションを指定します。
3. 「開く」 をクリックします。

モデル内に存在すれば変数ラベルを使用 **:** PMML が、データ・ディクショナリー内の変数に対して、変数名と変数ラベル (*RefID* に対する Referrer ID など) の両方を指定している場合があります。元のエクスポートされた PMML に変数ラベルが存在するときに変数ラベルを使用するには、このオプションを選択します。

変数ラベル・オプションを選択したにもかかわらず、PMML 内に変数ラベルがない場合、変数名は通常のように使用されます。

PMML をサポートするモデルの種類

PMML のエクスポート

IBM SPSS Modeler のモデル**:** IBM SPSS Modeler で作成された次のモデルは、PMML 4.0 としてエクスポートできます。

- C&R Tree
- QUEST
- CHAID
- ニューラル・ネットワーク
- C5.0
- ロジスティック回帰
- 一般化線型
- SVM
- Apriori

- Carma
- K-Means
- Kohonen
- TwoStep
- TwoStep-AS
- GLMM (PMML はすべての GLMM モデルでエクスポートされますが、PMML は固定効果のみを持ちます)
- ディシジョン・リスト
- Cox
- シーケンス (シーケンス PMML モデルのスコアリングはサポートされていません)
- Random Trees
- Tree-AS
- 線型
- Linear-AS
- 線型回帰
- ロジスティック回帰
- GLE
- LSVM
- 異常値検査
- KNN
- アソシエーション・ルール

データベース固有のモデル**:** データベース固有のアルゴリズムを使用して作成されたモデルの場合、PMML エクスポートは、使用できません。Microsoft または Oracle Data Miner の Analysis Services を使用して作成されたモデルをエクスポートすることはできません。

PMML のインポート

IBM SPSS Modeler では、すべての IBM SPSS Statistics 製品の現在のバージョンで作成された PMML モデルをインポートおよびスコアリングできます。IBM SPSS Statistics 17.0 で生成されたモデルまたは変換 PMML と同様に、IBM SPSS Modeler からエクスポートされたモデルもインポートおよびスコアリングできます。基本的には、次の例外を除いて、スコアリング エンジンはすべての PMML をスコアリングできます。

- Apriori、CARMA、異常検出、シーケンス、およびアソシエーション ルールのモデルをインポートすることはできません。
- スコアリングに使用できる場合でも、IBM SPSS Modeler へのインポート後に PMML を参照することはできません。(これには、初めに IBM SPSS Modeler からエクスポートされたモデルは含まれません。この制限を回避するには、モデルを、PMML ではなく、生成されたモデル・ファイル 「`*.gm`」としてエクスポートします。)
- インポート時には制限つき検証が行われますが、モデルのスコアリング試行時には完全検証が実行されます。そのため、インポートは正常に行われますが、スコアリングは失敗したり不正な結果が生成されます。

注**:** サード・パーティーの PMML を IBM SPSS Modeler にインポートした場合、IBM SPSS Modeler は、認識できてスコアリング可能である有効な PMML のスコアリングを試行します。ただし、すべての PMML がスコアリングされることや、PMML を生成したアプリケーションと同じ方法で PMML がスコアリングされることは保証されません。

スコアリング・アダプター向けにモデルを公開

スコアリング・アダプターをインストールしたデータベース・サーバーにモデルを公開できます。スコアリング・アダプタにより、データベースのユーザー定義関数 (UDF) 機能を使用してデータベース内でモデルスコアリングを実行できます。データベース内でスコアリングを実行することにより、スコアリング前にデータを抽出する必要がなくなります。スコアリング・アダプタへ公開することにより、SQL のいくつかの例を生成して UDF を実行します。

スコアリング・アダプターを公開するには

1. モデル・ナゲットをダブルクリックして開きます。
2. モデル・ナゲットのメニューから、次を選択します。

   「ファイル」 > 「**Server Scoring Adapter** 用に公開」
3. ダイアログ・ボックスの関連フィールドを入力して 「**OK**」 をクリックします。

データベース接続**:** モデルに使用するデータベースの接続の詳細を指定します。

公開 **ID:** (Db2 for z/OS データベースのみ) モデルの識別子。同じモデルを再構築し、同じ公開 ID を使用する場合、生成した SQL も同じため、以前生成した SQL を使用するアプリケーションを変更することなくモデルを再構築できます(他のデータベースについては、生成される SQL モデルに対して一意です)。

SQL の例を生成 **:** SQL の例を 「ファイル」 フィールドで指定されたファイルに生成します。

未精製モデル

未調整モデルには、データから抽出された情報が含まれますが、予測の直接の生成には設計されていません。つまり、これらをストリームに追加することはできません。未精製モデルは、モデル生成パレット上では、「磨いていないダイヤモンド」として表示されます。

図 *27.* 未精製モデルのアイコン

未調整ルール・モデルに関する情報を表示するには、モデルを右クリックし、コンテキスト・メニューから「参照」を選択します。IBM SPSS Modeler によって生成された他のモデルと同じように、生成したモデルについての要約やルール情報は各種タブによって提供されます。

ノードの生成**:** 「生成」メニューを使用し、ルールに基づいて新しいノードを作成することができます。

- 条件抽出ノード: 現在選択されているルールを適用するレコードを選択するために、条件抽出ノードを生成します。ルールが選択されていない場合、このオプションは無効になります。
- ルール セット**:**単一の対象フィールドの値を予測するために、ルール セット ノードを生成します。 詳しくは、トピック 284 ページの『アソシエーション・モデル・ナゲットからルールセットを生成する』を参照してください。

第 4 章 モデルのスクリーニング

フィールドとレコードのスクリーニング

モデル作成において関心の対象となる可能性が最も高いフィールドとレコードを検出するために、分析の前段階で、モデル作成ノードをいくつか使用することができます。特徴量選択ノードを使用して、重要度を基準にフィールドをスクリーニングしてランクを付けます。また、異常値検査ノードを使用して、「正常」データの既知のパターンに適合しない、通常とは異なるレコードを検出します。

特徴量選択ノードで、（欠損値の割合などの）諸基準に基づいて入力フィールドをスクリーニングして削除にかけ、指定した目標に相対的な残りの入力フィールドの重要度をランク付けします。例えば、数百の潜在的入力フィールドを含むデータセットがあるとして、患者予後のモデリングにはどれが役に立つのでしょう?

異常値検査ノードで、「正常な」データのパターンに合致しない異常ケースや外れ値を識別します。このノードで、外れ値が既知のパターンに当てはまらなかったり、何を探しているのかはっきりしなかったりする場合でも、外れ値を識別できます。

異常値検査は、クラスター分析を通じて普通でない (通常でない)レコードまたはケースを識別するということに留意してください。このクラスター分析は、特定の対象 (従属) フィールドを考慮せず、また、予測しようとするパターンに関連するフィールドかどうかを無視して、モデル内で選択されたフィールドのセットに基づいて行われます。このため、異常値検査は、特徴量選択や、スクリーニングやフィールドのランク付けのための別の手法と組み合わせて使用できます。例えば、特徴量選択を使用して特定の対象に関連するもっとも重要なフィールドを識別し、その後、異常値検査を使用して、そのようなフィールドにとってもっとも通常でないレコードを特定することができます(別のアプローチとして、ディシジョン・ツリー・モデルを構築し、潜在する異常値として誤って分類されたレコードを検査する方法があります。ただし、この方法は、大規模に繰り返したり自動化したりすることが困難です)。

特徴量選択ノード

場合によっては、数百または数千ものフィールドが入力フィールドとして使用される可能性があり、データ・マイニングの問題となります。その結果、どのフィールドや変数をモデルに含むかを調べるのに、大変な時間と努力を費やすことになるかもしれません。選択範囲を絞り込むために、特徴量選択アルゴリズムを使用して、所定の分析にとって最も重要なフィールドを識別することができます。例えば、いくつかの要素に基づいて患者予後を予測する場合、どの要素が最も重要でしょう?

特徴量選択は、次の 3 つの段階から成り立っています。

- スクリーニング : 重要でなく問題を含んだ入力値とレコードまたはケースを削除します。例えば、欠損値が多すぎる入力フィールドや、使用するには変動が大きすぎたり小さすぎたりする入力フィールドです。
- 順位化。 重要性に基づいて、残った入力フィールドをソートしランクを割り当てます。
- 選択: 例えば、最も重要な入力だけを保持し、その他はすべてフィルタリングまたは除外することにより、機能のサブセットを特定して後続のモデルで使用します。

たくさんの組織があまりにも多くのデータを抱え込んでいる時代に、モデリング・プロセスを単純化し迅速化する過程で特徴量選択を行うことの利点は、少なくありません。フィールドは最も重要な部分であるため、それに機敏に注意を集中することによって、必要な計算量を減らしたり、重要なことなのに小さくて見逃してしまいそうな人や物の関連を簡単に探し出したり、その結果として、単純かつ正確で説明が簡単なモデルを取得したりすることができます。モデルで使用するフィールドの数を減らすことによって、将来、反復して収集するデータ量を減らしたり、スコアリングの回数を減らしたりすることができます。

例: 電話会社は、特別プロモーションに対するこの会社の 5,000 人の顧客からの応答に関する情報を含んでいるデータ・ウェアハウスを持っています。このデータには、顧客の、雇用、収入、電話利用状況の統計などの多くのフィールドが含まれています。3 つの対象フィールドは、顧客がこの 3 つのフィールドに反応したかどうかを示しています。この会社は、このデータを活用して、今後、類似のオファーに対してどの顧客が反応する見込みが最も高いかという予測を立てたいと考えています。

要件: 1 つの対象フィールド (役割が対象 に設定されたフィールド) と、対象に対して相対的なスクリーニングまたはランク付けを行う複数の入力フィールドが必要です。対象フィールドおよび入力フィールドの尺度は連続型 (数値範囲) またはカテゴリー型です。

特徴量選択モデルの設定

「モデル」タブの設定値には、予測フィールドをスクリーニングするための基準を微調整できる設定とともに、標準的なモデル・オプションが含まれています。

モデル名: ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

入力フィールドのスクリーニング

スクリーニングには、入力フィールドと対象の関係について有用な情報を追加しない、入力フィールドまたはケースの削除が含まれます。スクリーニングのオプションは、選択した対象フィールドに関係する予測力を考慮せずに、問題になっているフィールドの属性に基づいています。スクリーニングされたフィールドは、入力フィールドをランク付けるのに使用された計算から除外され、オプションで、フィルターを適用したり、モデル作成に使用されるデータから取り除くことができます。

フィールドは、次の基準に基づいてスクリーニングできます。

- 欠損値の最大パーセンテージ :レコードの総数のパーセントとして示されたレコード数になるまで、多すぎる欠損値フィールドをスクリーニングします。欠損値フィールドの割合が大きいフィールドからは、あまり予測的情報を得ることができません。
- 単一カテゴリー内のレコードの最大パーセンテージ : レコードの総数の割に同じカテゴリーにかたよって多くのレコードを含んでいるフィールドをスクリーニングします。例えば、データベース内の顧客の 95% が同じ車種の車を運転している場合、この情報を含めても、次回から特定の顧客を区別する上で役に立ちません。指定された最大値を超えるフィールドは、スクリーニングされます。このオプションは、カテゴリー型フィールドに対してのみ適用されます。
- レコードのパーセンテージとしての最大カテゴリー数 : レコードの総数に対して多すぎるカテゴリーを減らす目的で、フィールドをスクリーニングします。高いパーセンテージのカテゴリーにただ 1 つのケースが含まれている場合、そのフィールドの使用が限定されている可能性があります。例えば、それぞれの顧客が皆異なる帽子を被っている場合、この情報は、顧客の行動パターンをモデル作成する上で役に立ちそうもありません。このオプションは、カテゴリー型フィールドに対してのみ適用されます。

- 最小変動係数 **:**指定された最小値以下の変動係数で、フィールドをスクリーニングします。この尺度は、入力フィールドの平均に対する入力フィールドの標準偏差の割合です。この値がゼロに近いと、変数の値にあまりばらつきがないと言えます。このオプションは、連続型 (数値範囲型) フィールドに対してのみ適用されます。
- 最小標準偏差 **:**指定された最小値以下の標準偏差で、フィールドをスクリーニングします。このオプションは、連続型 (数値範囲型) フィールドに対してのみ適用されます。

欠損値を含むレコード **:** 対象フィールドに欠損値を含レコードまたはケースを設定しているか、またはすべての入力フィールドに欠損値が含まれている場合、ランク付けに使用されるすべての計算から自動的に除外されます。

特徴量選択のオプション

「オプション」タブで、モデル・ナゲット内の入力フィールドを選択または除外するデフォルトの設定を指定できます。その後、以後のモデル構築作業で使用するフィールドのサブセットを選択するために、ストリームへモデルを追加できます。または、モデルの生成後にモデル・ブラウザー内で追加のフィールドを選択したり選択を解除したりして、このような設定を上書きすることもできます。ただし、デフォルトの設定はそれ以上変更しなくてもモデル・ナゲットに適用できるので、スクリプトを作成する目的に対しては、特に有用です。

詳しくは、トピック 58 ページの『特徴量選択モデルの結果』を参照してください。

使用可能なオプションは次のとおりです。

ランク付けされているすべてのフィールド **:** 「重要」、「境界」、または 「重要ではない」のランクに基づいてフィールドを選択します。各ランクと、レコードにランクを割り当てるために使用される分割値のラベルは、編集できます。

フィールドの上位数 **:** 重要度に基づいて上位 *n* 件のフィールドを選択します。

次より大きな重要度**:** 指定された値よりも高い重要度のすべてのフィールドを選択します。

対象フィールドは、この選択にかかわらず、常に保存されます。

重要度のランク付けオプション

すべてのカテゴリー **:**すべての入力フィールドと対象フィールドがカテゴリー型の場合、重要度には、次の 4 つの測定単位のいずれかでランクを付けることができます。

- **Pearson** のカイ **2** 乗 **:**既存の関係 (リレーションシップ) の強度または方向を示すことなく、対象フィールドと入力フィールドの独立性を検定します。
- 尤度比カイ **2** 乗 **:**Pearson のカイ 2 乗と似ていますが、対象フィールドと入力フィールドの依存性を検定します。
- **Cramer** の **V :** Pearson のカイ 2 乗の統計に基づいたアソシエーションの測定値。アソシエーションがないことを示す 0 の値から、完全なアソシエーションを示す 1 までの範囲の値です。
- ラムダ **:**変数が対象値を予測しようとするときの、誤差 (エラー) の減力を反映したアソシエーションの測定値。1 は、入力フィールドが対象を完全に予測することを示します。値が0 のときは、入力フィールドは対象フィールドの予測で有益な情報を提供しません。

一部のカテゴリー**:** すべての入力ではなく一部の入力がカテゴリー型で、対象もカテゴリー型である場合は、Pearson または尤度比カイ 2 乗のいずれかに基づいて重要度にランクを付けることができます (すべての入力がカテゴリー型でない限り、Cramer の *V* とラムダは使用できません)。

カテゴリー型と連続型の比較 **:** カテゴリー型入力フィールドを連続型の対象フィールドに対してランク付けする、またはその逆 (一方またはその他がカテゴリー型で、両方カテゴリー型でない) の場合に、*F* 統計が使用されます。

両方とも連続型 **:** 連続型の対象値に対する連続型の入力フィールドをランク付けする場合は、相関係数に基づいた *t* 統計が使用されます。

特徴量選択モデル・ナゲット

生成された特徴量選択モデル・ナゲットでは、特徴量選択ノードでランクが付けられたとおりに、選択した対象フィールドに関連する各入力フィールドの重要度が表示されます。ランク付けに先立ってスクリーニングされたすべてのフィールドもリストに表示されます。 詳しくは、トピック 55 ページの『特徴量選択ノード』を参照してください。

特徴量選択モデル・ナゲットが含まれたストリームを実行すると、そのモデルは、「モデル」タブでの選択で示されたように、選択した入力フィールドだけを保存するフィルターとして動作します。例えば、重要度が高いとランクされたすべてのフィールドを選択することも (デフォルトのオプションの 1 つ)、手動で「モデル」タブに表示されたフィールドのサブセットを選択することもできます。対象フィールドも、この選択にかかわらず保存されます。その他のフィールドは、すべて除外されます。

フィルタリングは、フィールド名にだけ基づいています。例えば、*Age* () と *Income* (年収) を選択すると、これらの名前のどちらかと一致するすべてのフィールドが保存されます。このモデルでは、新しいデータに基づいてランク付けを更新しません。単に、選択された名前に基づいてフィールドをフィルタリングするだけです。このため、新規または更新データにモデルを適用する場合は注意が必要です。データが更新されているかどうかが不明な場合は、モデルを再生成することをお勧めします。

特徴量選択モデルの結果

特徴量選択モデル・ナゲットの「モデル」タブには、ウィンドウ枠の上部にすべての入力フィールドのランクと重要度が表示されるので、フィルタリングするフィールドを、左の列のチェック・ボックスを使用して選択できるようになります。ストリームを実行すると、選択されたフィールドのみが保存され、その他のフィールドは破棄されます。デフォルトの選択はモデル構築ノード内で指定されたオプションに基づきますが、必要に応じて追加のフィールドを選択したり、選択を解除したりできます。

下のウィンドウ枠には、欠損値のパーセンテージやモデル作成ノードで指定されたその他の基準に基づいてランク付けから除外された入力フィールドが一覧表示されます。ランク付きのフィールドの場合と同様に、左の列のチェック・ボックスを使用して、これらのフィールドを含めるか、または破棄するかを選択できます。 詳しくは、トピック 56 ページの『特徴量選択モデルの設定』を参照してください。

- ランク、フィールド名、重要度、またはその他の表示された列でリストをソートするには、列見出しをクリックします。または、ツールバーを使用して、「ソート項目」リストから希望の項目を選択し、上向き矢印と下向き矢印を使用してソートの方向を変更します。
- ツールバーを使用してすべてのフィールドにチェックを入れたり外したりできます。また、「フィールドのチェック」ダイアログ・ボックスを利用してランクまたは重要度でフィールドを選択できます。さらに、Shift キーと Ctrl キーを押しながらフィールドをクリックして複数フィールド選択し、スペース・バーを使用して選択されたフィールドのグループのオン/オフを切り替えることもできます。 詳しくは、トピック 59 ページの『重要度別のフィールドの選択』を参照してください。

- 重要度が高い、境界、重要度が低い、として入力フィールドをランク付けするための閾値は、テーブルの下の凡例に表示されます。これらの値は、モデル作成ノード内で指定されます。 詳しくは、トピック 57 ページの『特徴量選択のオプション』を参照してください。

重要度別のフィールドの選択

特徴量選択モデル・ナゲットを使用してデータをスコアリングするときに、ランク付きの、またはスクリーニングされたフィールドのリストから選択されたすべてのフィールド (左の列内のチェック・ボックスで示される) は、保存されます。その他のフィールドは廃棄されます。選択項目を変更するには、ツールバーを使用して、ランクまたは重要度でフィールドを選択できるようにする「フィールドのチェック」ダイアログ・ボックスを利用できます。

マーク済みのすべてのフィールド**:** 重要度が高い、境界、または重要度が低い、としてマークされたすべてのフィールドを選択します。

フィールドの上位数 **:** 重要度に基づいて上位 *n* 件のフィールドを選択できます。

次より大きな重要度 **:** 指定された閾値よりも高い重要度のすべてのフィールドを選択します。

特徴量選択モデルからのフィルターの生成

特徴量選択モデルの結果に基づいて、「特徴量からフィルタを生成」ダイアログ・ボックスを使用して、指定された対象に関連する重要度に基づいたフィールドのサブセットを含めるか除外する、1 つ以上のフィルター・ノードを生成することができます。モデル・ナゲットはフィルターとして使用できますが、このノードには、モデルのコピーや修正なしで、さまざまなフィールド・サブセットを試行できる柔軟性があります。対象フィールドは、含めるか除外するかの選択にかかわらず、フィルターによって常に保存されます。

含める/除外 **:** フィールドを含めるか除外するかを選択できます。例えば、上位 10 個のフィールドを含めたり、重要度が低い、とマークされたすべてのフィールドを除外したりできます。

選択したフィールド **:** テーブル内で現在選択されているすべてのフィールドを含めるか、破棄します。

マーク済みのすべてのフィールド**:** 重要度が高い、境界、または重要度が低い、としてマークされたすべてのフィールドを選択します。

フィールドの上位数 **:** 重要度に基づいて上位 *n* 件のフィールドを選択できます。

次より大きな重要度 **:** 指定された閾値よりも高い重要度のすべてのフィールドを選択します。

異常値検査ノード

異常値検査モデルは、外れ値、つまりデータ内の通常とは異なるケースを識別するのに使用されます。通常と異なるケースに対処するルールを格納するほかのモデル作成の手法とは異なり、異常値検査モデルでは、通常の動作がどのようなものかという情報を格納します。このことで、既知のパターンを確認しなくても外れ値の識別が可能になり、新しいパターンが常に緊急事態になり得る不正検出のようなアプリケーションでは、このモデルが特に役立ちます。異常値検査は、管理抜きの手法です。つまり、開始時に使用する既知の不正が含まれた学習データ・セットが必要ありません。

外れ値を識別する伝統的な手法では、通常一度に 1 つか 2 つの変数を調べますが、異常値検査では、同類のレコードと見なされるクラスターまたはピア・グループを識別するために、大量のフィールドを検査できます。その後各レコードが、異常の可能性を識別するためにピア・グループ内で他のレコードと比較されます。ケースが正常の中心から離れるほど、通常とは異なる可能性が大きくなります。例えば、アルゴリズム

によってレコードが 3 つの異なるクラスターへ一括分類され、いずれかのクラスターの中心からかなり離れたところに収まるレコードには、フラグが設定されます。

各レコードには、異常値の指標が割り当てられます。これは、ケースが属するクラスターの平均に対するグループ偏差指標の割合です。この指標の値が大きいほど、ケースの平均からの偏差が大きくなります。通常の状況では、異常値指標の値が 1 または 1.5 より小さいケースは、偏差が平均とほとんど同じか、わずかに大きいだけなので、異常値とは見なされません。ただし、指標の値が 2 より大きいケースは、偏差が少なくとも平均の 2 倍であるため、異常値の有力な候補になります。

異常値検査は、以後の分析の候補となる通常でないケースやレコードを迅速に検査するために設計された、予備的な手法です。この手法は、異常性が疑わしいものを検査すると見なされるべきです。この手法では、異常性が疑われる ものが検出されます。つまり、さらに詳しい調査によって、その疑いが現実になる場合も、ならない場合もあります。レコードが完全に有効であっても、モデル構築の目的でデータからレコードをスクリーニングすることを選択することもできます。または、アルゴリズムによって偽 (false) の異常値だということが繰り返し判明した場合、このことは、データ収集の過程でのエラーまたは作為である可能性があります。

異常値検査は、クラスター分析を通じて普通でない (通常でない)レコードまたはケースを識別するということに留意してください。このクラスター分析は、特定の対象 (従属) フィールドを考慮せず、また、予測しようとするパターンに関連するフィールドかどうかを無視して、モデル内で選択されたフィールドのセットに基づいて行われます。このため、異常値検査は、特徴量選択や、スクリーニングやフィールドのランク付けのための別の手法と組み合わせて使用できます。例えば、特徴量選択を使用して特定の対象に関連するもっとも重要なフィールドを識別し、その後、異常値検査を使用して、そのようなフィールドにとってもっとも通常でないレコードを特定することができます(別のアプローチとして、ディシジョン・ツリー・モデルを構築し、潜在する異常値として誤って分類されたレコードを検査する方法があります。ただし、この方法は、大規模に繰り返したり自動化したりすることが困難です)。

例: 不正の疑いのあるケースの農業開発補助金のスクリーニングでは、異常値検査を使用して平均からの偏差を発見し、異常で詳しい調査が必要なレコードを強調表示します。特に注目するのは、農場の種類と規模から見て補助金の申請金額が多すぎる (または少なすぎる) と考えられる場合です。

要件: 1 つ以上の入力フィールドが必要です。入力ノードまたはデータ型ノードを使用して、役割が 「入力」 に設定されたフィールドだけを、入力として使用できます。対象フィールド (役割が 「対象」 または 「両方」 に設定されている) は、無視されます。

利点: 既知のルール・セットに従うケースではなく、従わない ケースにフラグを立てておくと、異常値検査モデルは事前に認識されていないパターンの異常なケースを識別することができます。特徴量選択と組み合わせて使用すると、異常値検査により、最も興味あるレコードを識別するために、大量のデータを比較的迅速にスクリーニングすることが可能になります。

異常値検査モデルのオプション

モデル名: ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

異常値の分割基準 : 異常値にフラグを立てるための分割値を決めるのに使用される方法を指定します。次使用可能なオプションは次のとおりです。

- 異常値指数の最低レベル: 異常としてフラグを設定するための最小分割値を指定します。この閾値を満たす、または超えたレコードにフラグが立てられます。

- 学習データの最も異常なレコードのパーセンテージ：学習データの総レコード数に対して指定されたパーセンテージでフラグを立てるレベルで、自動的に閾値を設定します。結果の分割値は、モデル内にパラメーターとして含まれます。このオプションで分割値の設定方法が決定されますが、スコアリング中にフラグが立てられる実際のパーセンテージ (レコード数の) を決めるわけではありません。実際のスコアリング結果は、データに依存して変化します。
- 学習データの最も異常なレコード数：学習データの指定されたレコード数にフラグを立てるレベルで、自動的に閾値を設定します。結果の閾値は、モデル内にパラメーターとして含まれます。このオプションで分割値の設定方法が決定されるのであり、スコアリング中にフラグが立てられる特定のレコード数を決めるわけではありません。実際のスコアリング結果は、データに依存して変化します。

注：分割値の決定方法にかかわらず、この分割値は、各レコードに報告される潜在的な異常性の指標値に影響ありません。モデルの評価またはスコアリング時に、異常としてフラグが立てられる閾値を指定するだけです。後により多くの、またはより少ないレコード数を検査する予定がある場合は、条件抽出ノードを使用して、異常性の指標値 (`$O-AnomalyIndex > X`) に基づいてレコードのサブセットを識別することができます。

報告する異常フィールド数：特定のレコードに異常としてフラグが立てられる理由 (内容) を、報告するフィールド数を指定します。レコードが割り当てられたクラスターのフィールド基準値からの最大の偏差を示すフィールドとして定義された、最も異常なフィールドから報告されます。

異常値検査のエキスパート・オプション

欠損値のオプションと他の設定値を指定するには、「エキスパート」タブでモード を 「エキスパート」に設定します。

調整係数：間隔の計算時に連続型 (数値範囲型) フィールドに指定された相対的な重みとカテゴリー・フィールド間のバランスをとるために使用される値。値が大きくなればなるほど、連続型フィールドの影響が増大します。これは、ゼロ以外の値である必要があります。

ピア・グループ数を自動的に計算する：異常値分析は、学習データに最適なピア・グループ数を選択するために、大量の潜在的なソリューションを迅速に分析するのに使用できます。最小と最大のピア・グループ数を設定して、範囲を広げたり狭めたりすることができます。値が大きいほど広範囲な有力ソリューションを検索できますが、処理時間に応じてコストも増加します。

ピア・グループ数を指定する：モデルに含めるクラスター数がわかっている場合は、このオプションを選択してピア・グループ数を入力します。このオプションを選択すると、一般的にパフォーマンスが向上します。

ノイズ・レベルとノイズ比：この設定は、2 段階のクラスタリング中に外れ値をどのように処理するかを決定します。第 1 の段階では、クラスター特性 (CF) ツリーを使用して、大量の個別レコードから管理可能な数量のクラスターへとデータを圧縮します。ツリーは類似性基準に基づいて構築され、ツリーのノード内のレコードがあまりにも多くなると、そのノードは子ノードに分割されます。第 2 段階では、CF ツリーのターミナル・ノードで階層クラスタリングが始まります。ノイズ処理は、最初のデータ・パスでオンにされ、2 回目のデータ・パスでオフにされます。最初のデータ・パスからのノイズ・クラスター内のケースは、2 回目のデータ・パスで通常のクラスターへ割り当てられます。

- ノイズ・レベル: 0 から 0.5 までの値を指定してください。この設定は、成長フェーズ中に CF ツリーが一杯になった場合にのみ関係します。つまり、リーフ・ノード内でこれ以上ケースを受け取れず、どのリーフ・ノードも分割できないという場合です。

CF ツリーが一杯でノイズ・レベルが 0 に設定されていると、閾値が増やされ、CF ツリーはすべてのケースを伴って再び大きくなります。最終クラスター化の後、クラスターに割り当てられなかった値は、外れ値としてラベル付けされます。外れ値のクラスターには、-1 の識別番号が与えられます。外れ値のクラスターは、クラスター数には数えられません。つまり、*n* 個のクラスターとノイズ処理を指定すると、アルゴリズムによって、*n* 個のクラスターと 1 個のノイズ・クラスターが出力されます。実際的な問題として、この値を増やすと、アルゴリズムは、通常 (普通) でないレコードを独立した外れ値クラスターへ割り当てるのではなく、ツリーに適合させるための自由度を得ることになります。

CF ツリーが一杯でノイズ・レベルが 0 より大きい場合は、疎葉内のデータがすべてノイズ葉に配置された後、CF ツリーが再成長されます。最大の葉にあるケース数に対する疎葉内のケース数の比率がノイズ・レベルより小さい場合、その葉は疎 (まばら) であるとみなされます。ツリーが成長した後、可能な場合は外れ値が CF ツリー内に配置されます。そうではない場合、クラスタリングの第二フェーズのために外れ値は廃棄されます。

- ノイズ率**:** ノイズのバッファリングに使用されるコンポーネントに割り当てられる、メモリーの量を指定します。この値は、0.0〜0.5 の範囲です。特別なケースをツリーの葉に挿入すると生じる気密度が閾値より小さい場合、その葉は分割されません。気密度が閾値を超える場合は葉が分割され、別の小さなクラスターが CF ツリーに追加されます。実際的には、この設定を増やすと、アルゴリズムがより迅速により単純なツリーを作成する方向へ向かう原因となることがあります。

欠損値を代入**:** 連続型フィールドの場合、欠損値の代わりにフィールドの平均値に置き換わります。カテゴリー型フィールドの場合、欠損カテゴリーは結合され、有効なカテゴリーとして処理されます。このオプションが選択解除されている場合は、欠損値のあるすべてのレコードが分析から除外されます。

異常値検査モデル・ナゲット

異常値検査モデル・ナゲットには、異常値検査モデルに捕捉されたすべての情報と、学習データと推定プロセスに関する情報が含まれます。

異常値検査モデル・ナゲットが含まれたストリームを実行すると、多数の新規フィールドが、モデル・ナゲット内の「設定」タブで選択したとおりに、ストリームへ追加されます。詳しくは、トピック 63 ページの『異常値検査モデルの設定』を参照してください。新規フィールドの名前はモデル名を基本にし、次の表にまとめたように、先頭に *$O* が付きます。

表 *6.* 新規フィールド名の生成

フィールド名	説明
$O-Anomaly	レコードが異常かどうかを示すフラグ型フィールド。
$O-AnomalyIndex	レコードの異常性の指標値。
$O-PeerGroup	レコードが割り当てられるピア・グループを指定します。
$O-Field-n	クラスターの基準値からの分散の観点から、*n* 番目のもっとも異常なフィールドの名前。
$O-FieldImpact-n	フィールドの変数偏差指標。この値で、レコードが割り当てられるクラスターの、フィールド基準値からの偏差を測定します。

オプションで、結果を読みやすくするために、異常でないレコードのスコアリングを抑制することができます。 詳しくは、トピック 63 ページの『異常値検査モデルの設定』を参照してください。

異常値検査モデルの詳細

生成された異常値検査モデルの「モデル」タブには、モデル内のピア・グループについての情報が表示されます。

報告されたピア・グループのサイズと統計は、学習データに基づいて推定され、同じデータを実行したとしても、実際のスコアリング結果と若干異なる可能性があることに注意してください。

異常値検査モデルの要約

異常値検査モデル・ナゲットの「要約」タブには、フィールド、構築の設定、および推定プロセスについての情報が表示されます。ピア・グループの数も、異常としてフラグを立てるのに使用される分割値とともに表示されます。

異常値検査モデルの設定

「設定」タブを使用すると、モデル・ナゲットをスコアリングするためのオプションを指定できます。

次で異常レコードを指摘**:** 出力における異常レコードの処理方法を指定します。

- フラグとインデックス**:** モデル内に含まれる分割値を超えたすべてのレコードに *True (*真*)* が設定される、フラグ型フィールドを作成します。各レコードの異常値の指標 (インデックス) も、別のフィールドに報告されます。 詳しくは、トピック 60 ページの『異常値検査モデルのオプション』を参照してください。
- フラグだけ**:** フラグ型フィールドを作成しますが、各レコードの異常指数は報告されません。
- インデックスだけ**:** フラグ型フィールドは作成せずに、異常指標を報告します。

報告すべき異常フィールド数**:** 特定のレコードに異常としてフラグが立てられる理由の指標として報告するフィールドの数を指定します。レコードが割り当てられたクラスターのフィールド基準値からの最大の偏差を示すフィールドとして定義された、最も異常なフィールドから報告されます。

レコードを破棄**:** 下流ノードに潜在する異常値に焦点を絞りやすくするために、すべての「非異常」レコードをストリームから破棄するには、このオプションを選択します。または、すべての「異常」レコードを破棄することも選択できます。これは、以後の分析を、モデルに基づいて異常である可能性があることを示すフラグが立てられていないレコードに限定するためです。

注**:** 同じデータに対してスコアリングと学習を行った場合でも、丸め処理のわずかな違いにより、スコアリング中にフラグが立てられた実際のレコードの数と、モデルの学習中にフラグが立てられたレコードの数の間に、差異が生じる場合があります。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。

- データベースの外部でスコアリング: このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

第 5 章 自動化モデル作成ノード

自動化モデル作成ノードは多くの異なるモデル作成ノードを推定および比較し、単一のモデル作成の実行でさまざまな方法を試用できるようにします。使用するモデル作成アルゴリズム、相互排他的な組み合わせを含む、それぞれ特定のオプションを選択できます。例えば、ニューラル・ネットワークに高速方法、動的方法、剪定方法の中から 1 つ選ぶのではなく、そのすべてを試行できます。ノードは、オプションの可能なすべての組み合わせを検証し、指定する指標に基づいて候補モデルをランク付け、スコアリングまたは詳細分析のブランチに最適なモデルを保存します。

分析の必要性に応じて、3 つの自動化モデル作成ノードから選択できます。

自動分類ノードは、2種類の結果 (yes/no、 churn/don't churn など) を生じる多くの異なるモデルを作成および比較し、与えられた分析への最善のアプローチを選ぶことができるようになります。多くのモデル作成アルゴリズムに対応し、希望する方法、各特定のオプション、そして結果を比較するための基準を選択することができます。このノードで、指定されたオプションに基づいてモデルのセットが生成され、指定された基準に基づいて最善の候補がランク付けされます。

自動数値ノードでは、多くのさまざまな方法を使用し、連続する数値範囲の結果を求めてモデルを推定し比較します。このノードは、自動分類ノードと同じ方法で動作し、1 回のモデル作成のパスで、複数の組み合わせのオプションを使用し試すアルゴリズムを選択することができます。使用できるアルゴリズムには、ニューラル・ネットワーク、C&R Tree、CHAID、線型、一般化線型、サポート・ベクトル・マシン (SVM) が含まれています。モデルは、相関、相対エラー、または使用された変数の数に基づいて比較できます。

自動クラスタリング・ノードは、同様の特性を持つレコードのグループを識別するクラスタリング・モデルを推定し、比較します。ノードは他の自動化モデル作成ノードと同じように動作し、複数の組み合わせのオプションを単一のモデル作成の実行で検証できます。モデルは、クラスター・モデルの有用性をフィルタリングおよびランク付けする基本的な指標を使用して比較し、特定のフィールドの重要度に基づいて指標を提供します。

最良のモデルは単一の複合モデル・ナゲットに保存され、モデルを参照および比較でき、スコアリングに使用するモデルを選択できます。

- 2 値、名義型、数値型対象の場合のみ、複数のスコアリング・モデルを選択し、単一のモデル・アンサンブルにスコアを結合できます。複数のモデルから予測を結合することによって、各モデルの制限を回避でき、モデルの 1 つから取得するより全体的な精度が高い結果が得られます。
- オプションで、結果をドリル・ダウンし、使用するまたは詳細に検証するモデルのモデル作成ノードまたはモデル・ナゲットを生成することができます。

モデルおよび実行時間

データセットおよびモデル数によって、自動化モデル作成ノードの実行には数時間以上かかる場合があります。オプションを選択する際は、作成されるモデルの数に注意してください。実際の実行時で、システムリソースが十分でない場合、可能であればモデル作成の実行を夜間または週末にスケジューリングしてください。

- 必要な場合は、データ区分ノードまたはサンプリング・ノードを使用して、最初の学習パスに含まれるレコード数を減らすことができます。候補のモデルを絞り込んだ後は、データセット全体を復元できます。
- 入力フィールドの数を減らすには、特徴量選択機能を使用します。詳しくは、トピック 55 ページの『特徴量選択ノード』を参照してください。代わりに、最初のモデル作成の実行により、詳細に検証する価値のあるフィールドおよびオプションを識別することができます。例えば、良好なパフォーマンスのモデルがすべて同じ 3 つのフィールドを使用すると考えられる場合、これらのフィールドは保存する価値があるという強い目安となります。
- オプションで、モデルの推定に費やす時間を制限し、モデルをスクリーニングしランクをつけるために使用する評価測定法を指定します。

自動化モデル作成ノードのアルゴリズムの設定

各モデル タイプに対し、デフォルトの設定値を使用するか、各モデルごとのオプションを選択することができます。固有のオプションは、各モデル作成ノードで使用できるオプションと同じですが、1 つずつの設定を選択するのではなく、多くの場合適用に必要な数だけ設定を選択できるという違いがあります。例えば、ニューラル・ネットワーク・モデルを比較する場合、異なる学習方法を複数選択し、ランダム・シードを使用して、または使用せずに各方法を試行することができます。選択したオプションの考えられる組み合わせはすべて使用され、1 回のパスに多くのさまざまなモデルを簡単に生成することができます。ただし、複数の設定を選択すると、モデルの数がすぐに乗算されるので注意してください。

各モデル タイプのオプションを選択するには

1. 自動化モデル作成ノードで、「エキスパート」 タブを選択します。
2. モデル タイプの 「モデル パラメーター」 列をクリックします。
3. ドロップダウン・メニューで、「指定」 を選択します。
4. 「アルゴリズムの設定」 ダイアログの 「オプション」 列からオプションを選択します。

注 : 「アルゴリズムの設定」ダイアログの「エキスパート」タブで高度なオプションを設定できます。

自動化モデル作成ノードの停止規則

自動化モデル作成ノードに指定される停止基準は、ノードによる個々のモデル構築の停止ではなく、ノード全体の実行に関わっています。

実行時間全体を制限 : (ニューラル・ネットワーク、K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net および C&R Tree モデルのみ) 指定された時間が経過すると実行を停止します。その時間までに生成されたすべてのモデルがモデル・ナゲットに含まれますが、それ以上のモデルは作成されません。

有効なモデルが作成されたらすぐに停止 :「破棄」タブ (自動分類ノードまたは自動クラスタリング・ノード) または「モデル」タブ (自動数値ノード) で指定されたすべての基準をモデルが満たしている場合、実行を停止します。 詳しくは、トピック 73 ページの『自動分類ノードの「破棄」オプション』を参照してください。詳しくは、トピック 82 ページの『自動クラスタリング・ノードの「破棄」オプション』を参照してください。

自動分類ノード

自動分類ノードは、さまざまな方法を使用して名義型 (セット型) または 2 値 (yes/no) の対象についてモデルを推定および比較し、単一のモデル作成実行でさまざまな方法を試用することができます。使用するアルゴリズムを選択し、複数の組み合わせのオプションを検証することができます。例えば、SVM に放射基底関数、多項式、Sigmoid、線型の各方法の中から 1 つ選ぶのではなく、そのすべてを試行できます。ノードは、オプションの可能なすべての組み合わせを検証し、指定する指標に基づいて候補モデルをランク付け、スコアリングまたは詳細分析のブランチに最適なモデルを保存します。詳しくは、 65 ページの『第 5 章 自動化モデル作成ノード』を参照してください。

例　小売業には、過去のキャンペーンで特定の顧客に行ったオファーを追跡する履歴データがあります。企業は、それぞれの顧客に合った適切な提案を行うことで、さらに収益を上げることを望んでいます。

要件　測定の尺度が名義型 またはフラグ型 の対象フィールド (役割が「対象」に設定されたもの) と、1 つ以上の入力フィールド (役割が「入力」に設定されたもの) が必要です。フラグ型フィールドの場合、対象フィールド向けに定義された *True* 値は、プロフィット、リフト、および関連統計量の計算時のヒットを表現すると想定されます。入力フィールドは尺度が連続型またはカテゴリー型である場合があり、一部の入力が一部のモデル タイプに適切でないという制限があります。例えば、C&R Tree、CHAID、および QUEST モデルで入力として使用されている順序型フィールドには、(文字列ではなく) 数値型ストレージを含む必要があります。数値型ストレージが含まれない場合は、これらのモデルに無視されます。同様に、連続型入力フィールドが分割される場合があります。用件は、個別のモデル作成ノードを使用している場合と同じです。例えば、ベイズ・ネットワーク・モデルは、ベイズ・ノードから生成された場合も、自動分類ノードから生成された場合も同じように動作します。

度数および重みフィールド

それは、例えばユーザーが構築データセットは母集団のセクションを低く示すことを認識しているため、または 1 つのレコードが多くの同一ケースを示すためです。C&R Tree、CHAID、QUEST、ディシジョン・リスト、ベイズ・ネットワーク・モデルで度数フィールドを使用することができます。重みフィールドは、C&R Tree、CHAID、および C5.0 で使用することができます。その他のモデル タイプでは、これらのフィールドを無視してモデルを構築します。度数および重みフィールドはモデル作成にのみ使用され、モデルの評価またはスコアリングの場合は考慮されません。 詳しくは、 33 ページの『度数フィールドと重みフィールドの使用』を参照してください。

接頭部

自動分類ノードのナゲットにテーブル・ノードを接続する場合は、名前が $ 接頭辞で始まるいくつかの新しい変数がテーブルに存在します。

スコアリング時に生成されるフィールドの名前は、対象フィールドに基づきますが、標準の接頭辞が付加されます。それぞれのモデル タイプでそれぞれの接頭辞を使用します。

例えば、接頭辞 $G、$R、$C は、それぞれ、一般化線型モデル、CHAID モデル、および C5.0 モデルによって生成される予測の接頭辞として使用されます。$X は、通常アンサンブルを使用して生成されます。$XR、$XS、$XF は、対象フィールドがそれぞれ連続型フィールド、カテゴリ型フィールド、フラグ型フィールドの場合に接頭辞として使用されます。

$..C 接頭辞は、カテゴリ型対象またはフラグ型対象の予測確信度で使用されます。例えば、$XFC はアンサンブル フラグ型予測確信度の接頭辞として使用されます。 $RC および $CC は、CHAID モデル、および C5.0 モデルそれぞれに対する確信度の単一予測用の接頭辞です。

サポートするモデル タイプ

サポートするモデル タイプは、ニューラル ネットワーク、C&R Tree、QUEST、CHAID、C5.0、ロジスティック回帰、ディシジョン リスト、ベイズ ネットワーク、判別分析、最近隣、SVM、XGBoost Tree、および XGBoost-AS です。詳しくは、トピック 69 ページの『自動分類ノードのエキスパートに関するオプション』を参照してください。

自動分類ノードの「モデル」オプション

自動分類ノードの「モデル」タブで、モデルの比較に使用される基準に沿って、作成されるモデル数を指定することができます。

モデル名: ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

分割モデルを作成 **:** 分割フィールドとして指定される入力フィールドの各値の個別モデルを作成します。詳しくは、 28 ページの『分割モデルの作成』を参照してください。

モデルのランク付け基準: モデルの比較およびランク付けに使用する基準を指定します。オプションには、全体の精度、ROC 曲線下の領域、プロフィット、リフト、およびフィールド数が含まれます。ここでの選択内容に関係なく、これらの測定基準のすべてが要約レポートで使用できます。

注 : 名義 (セット型) フィールドの場合、ランク付けは全体の精度またはフィールド数に限られています。

対象フィールド向けに定義された *True* 値は、プロフィット、リフト、および関連統計量の計算時のヒットを表現すると想定されます。

- **全体の精度:** 全レコード数に対して、モデルによって正確に予測されるレコード数の割合。
- **ROC 曲線下の領域 (Area under the ROC curve):** ROC 曲線は、モデルのパフォーマンスについての指標を提供します。基準線の上部にある曲線の位置が遠いほど、検定が正確です。
- **利益 (累積):** 指定したコスト、収益、および重みづけの基準に基づいて計算された、累積百分位の全体の利益の合計 (予測の信頼度によってソートされます)。通常、利益は、上位の百分位のゼロ付近から始まり、安定的に増加してから、次に減少します。すぐれたモデルの場合、くっきりとした山形が見られます。これは、山形が発生した場所の百分位とともにレポートされます。情報が得られないモデルの場合、利益曲線が比較的まっすぐで、適用するコストや収益の構造によって上昇または下降したり、平坦になったりする可能性があります。
- **リフト (累積):** サンプル全体に対する、累積分位でのヒット数の比率 (分位は予測の信頼度によってソートされます)。例えば、上位の分位に対するリフト値が 3 の場合は、ヒット率がサンプル全体の 3 倍高いことを示します。すぐれたモデルの場合、リフトは上位の分位では 1.0 を十分に超えて始まり、下部の分位に向けて 1.0 に向かって急激に下降します。リフトが 1.0 付近に留まっているモデルからは情報が得られません。
- **フィールド数:** 使用された入力フィールドの数に基づいて、モデルをランク付けします。

モデルのランク付けに使用する区分: データ区分が使用される場合は、ランクが学習用データセットに基づくか検定セットに基づくかを指定できます。大規模なデータセットの場合、モデルの予備的スクリーニングにデータ区分を使用すると、パフォーマンスが著しく改善される可能性があります。

使用モデル数: 作成されるモデル・ナゲットに表示されるモデルの最大数を指定します。上位にランクされたモデルが指定されたランク付けの基準に従って一覧表示されます。この制限数を増やすと、パフォーマンスが低下するおそれがあります。許容できる最大数は 100 です。

予測変数の重要度を計算: 重要度の適切な測定基準を作成するモデルの場合、モデル推定時に各予測値の相対重要度を示すグラフを表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。予測変数の重要度によってモデルを計算するために必要な時間が長くなる場合があり、多くの異なるモデル全体で比較する場合はお勧めできません。詳細に検証する少数のモデルに対する分析を絞り込んだ場合に、より有用です。 詳しくは、トピック 44 ページの『予測変数の重要度』を参照してください。

利益の基準: 注 : 利益は、各レコードの収益から、そのレコードのコストを引いた値と等しくなります。分位のプロフィットは、その分位の全レコードのプロフィットを合計したものです。プロフィットはヒットだけに適用されることを前提としますが、コストはすべてのレコードに適用されます。

- コスト : 各レコードに関連付けるコストを指定します。「固定」 または 「変数」 を選択することができます。固定コストの場合はコストの値を指定してください。可変コストの場合は、フィールド・ピッカー・ボタンをクリックして、コスト・フィールドとして使用するフィールドを選択してください。(「コスト」は、ROC グラフには使用できません。)
- 収益 : ヒットを表し各レコードに関連付ける収益を指定します。「固定」 または 「変数」 を選択することができます。固定収益の場合は収益値を指定してください。可変収益の場合は、フィールド・ピッカー・ボタンをクリックして、収益フィールドとして使用するフィールドを選択してください。(「収益」は、ROC グラフには使用できません。)
- 重み: データのレコードが複数のユニットからなる場合は、出現頻度の高い重みを使用して結果を調整できます。「固定」 または 「変数」 を選択して、各レコードに関連付ける重みの種類を指定します。重みを固定する場合は、重みの値 (レコードごとのユニット数) を指定してください。重みを変数にする場合は、フィールド・ピッカー・ボタンをクリックして、重みフィールドとして使用するフィールドを選択してください。(「重み」は、ROC グラフには使用できません。)

リフト基準: 注 : リフト計算で使用するパーセンタイルを指定します。なお、結果の比較時にこの値も変更できます。 詳しくは、トピック 82 ページの『自動化モデル・ナゲット』を参照してください。

自動分類ノードのエキスパートに関するオプション

自動分類ノードの「エキスパート」タブで、データ区分を適用し (利用可能な場合)、使用するアルゴリズムを選択し、停止基準を指定することができます。

モデルの選択。 デフォルトでは、作成対象としてすべてのモデルが選択されます。ただし、Analytic Server を使用している場合、モデルを Analytic Server で実行可能なものだけに制限し、それらをプリセットすることを選択できます。これにより、分割モデルが作成されるか、あるいは大規模データ・セットを処理する準備ができます。

注: 自動分類ノード内での Analytic Server モデルのローカル作成はサポートされていません。

使用されたモデル: 左側のチェック・ボックスを使用して、比較に含めるモデル タイプ （アルゴリズム）を選択します。選択したタイプが多ければ多いほど沢山のモデルが作成されるため、処理時間が長くなります。

モデル タイプ: 使用できるアルゴリズムを表示します (下記参照)。

モデル パラメーター**:** 各モデル タイプに対し、デフォルト設定を使用するか、「指定」 を選択してオプションを選択することができます。特定の複数のオプションは別のモデル作成ノードで利用できるものと同じであり、複数オプションまたは組み合わせオプションの違いも選択できます。例えば、ニューラル・ネットワーク・モデルを比較する場合に 6 つの学習方法の 1 つを選択するのではなく、単一パスで 6 モデルを学習させるためにそのすべてを選択することができます。

モデルの数**:** 現在の設定に基づいて各アルゴリズムに対応して作成されるモデルの数を表示します。オプションを組み合わせるとモデルを簡単に追加できるので、特に大きなデータセットを使用する場合はこの数字に細かい注意を払ってください。

単一モデルの構築最大時間を制限 **:** (K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net およびディシジョン・リスト・モデルのみ) モデルの最大時間制限を設定します。例えば、複雑な交互作用を含む特定のモデルの学習で予期外に長い時間を必要とする場合、すべてのモデルの作成を実行しません。

注**:** 対象が名義型 (セット型) の場合、ディシジョン・リスト・オプションは使用できません。

サポート対象のアルゴリズム

サポート・ベクター・マシン (SVM) ノードを使用すると、オーバーフィットすることなく、データを 2 つのグループのいずれかに分類することができます。SVM は、非常に多数の入力フィールドを含むデータセットなど、広範なデータセットを処理することができます。

k が整数である場合、*k* 最近傍 (KNN) ノードは、新しいケースを、予測領域の新しいケースに最も近い *k* 個のオブジェクトのカテゴリーまたは値と関連付けます。類似したケースはお互いに近く、類似していないケースはお互いに離れています。

判別分析によって、ロジスティック回帰より厳密な仮説を立てることができますが、これらの仮説が一致した場合、ロジスティック回帰分析に対する様々な代替あるいは補足になります。

Bayesian network (ベイズ) ノードを使用すると、観測された情報および記録された情報を実際の知識を組み合わせることによって確率モデルを作成し、発生の尤度を確立できます。このノードは、主に分類に使用される Tree Augmented Naïve Bayes (TAN) および Markov Blanket ネットワークに焦点を当てています。

ディシジョン・リスト・ノードは、母集団に関連する与えられた 2 値の結果の高いもしくは低い尤度を示すサブグループまたはセグメントを識別します。例えば、離れる可能性の少ないもしくはキャンペーンに好意的に答える可能性のある顧客を探すことができます。顧客区分を追加し、結果を比較するために他のモデルを並べて表示することによって、ビジネスに関する知識をモデルに導入することができます。ディシジョン・リスト・モデルは、ルールのリストから構成され、各ルールには条件と結果が含まれます。ルールは順番に適用され、一致する最初のルールで、結果が決まります。

ロジスティック回帰は、入力フィールドの値に基づいてレコードを分類する統計手法です。線型と似ていますが、数値範囲ではなくカテゴリー対象フィールドを使用します。

CHAID ノードはディシジョン・ツリーを生成し、カイ二乗統計値を使用して最適な分割を識別します。C&R Tree および QUEST ノードと違って、CHAID は、非 2 分岐ツリーを生成できます。これは、ある分岐が 3 個以上のブランチを持つことを意味します。対象フィールドおよび入力フィールドは、数値範囲 (連続型) またはカテゴリーとなります。Exhaustive CHAID は CHAID の修正版で、可能性のある分割すべてを調べることで、よりよい結果を得られますが、計算時間も長くなります。

QUEST ノードには、ディシジョン・ツリーの構築用に2 分岐の方法が用意されています。これは、大規模な C&R Tree 分析が必要とする処理時間を短縮すると同時に、より多くの分割を可能にする入力値が優先される分類ツリー内の傾向を低減するように設計されています。入力フィールドは、数値範囲 (連続型) にできますが、目標変数はカテゴリーでなければなりません。すべての分割は 2 分岐です。

C&R Tree (分類と回帰ツリー) ノードは、ディシジョン・ツリーを生成し、将来の観測値を予測または分類できるようにします。この方法は再帰的なデータ区分を使用して学習レコードを複数のセグメントに分割し、各ステップで不純性を最小限に抑えます。ツリーのノードが「純粋」であると考えられるのは、ノード中にあるケースの 100% が、対象フィールドのある特定のカテゴリーに分類される場合です。対象フィールドおよび入力フィールドは、数値範囲またはカテゴリー (名義型、順序型、フラグ) が使用できます。すべての分岐は 2 分割です (2 つのサブグループのみ)。

C5.0 ノードは、ディシジョン・ツリーとルール・セットのどちらかを構築します。このモデルは、各レベルで最大の情報の対応をもたらすフィールドに基づいてサンプルを分割します。対象フィールドは、カテゴリーでなければなりません。複数の分割を 2 つ以上のサブグループに分割できます。

ニューラル・ネット・ノードは、人間の脳が情報を処理する方法を単純化したモデルを使用します。ニューラル・ネットワーク・ノードは、連係する多数の単純な処理単位をシミュレートします。処理単位は、ニューロンを抽象化したものと表現できます。ニューラル・ネットワークは強力な一般関数推定法であり、学習させたり、適用するには、最低限の統計学および数学の知識しか必要ありません。

線型モデルは、対象と 1 つまたは複数の予測値との線型の関係に基づいて連続型対象を予測します。

線型サポート・ベクター・マシン (LSVM) ノードを使用すると、オーバーフィットすることなく、データを 2 つのグループのいずれかに分類することができます。LSVM は線型であり、極めて多数のレコードを含むデータセットなど、広範なデータセットを処理することができます。

Random Trees ノードは、既存の C&R Tree ノードと似ていますが、ビッグデータを処理して単一のツリーを作成することを目的に設計されており、結果のモデルが SPSS Modeler バージョン 17 で追加された出力ビューアに表示されます。Random Trees ノードは、将来の観測値を予測または分類するために使用するディシジョン ツリーを生成します。この方法では、再帰的なデータ分岐を使用して、各ステップで不純性を最小限に抑えることで、学習レコードがセグメントに分割されます。ツリー内のノードは、ノード内のケースの 100% が対象フィールドの特定のカテゴリーに分類される場合に、純粋 と見なされます。対象フィールドおよび入力フィールドは、数値範囲またはカテゴリー (名義型、順序型、フラグ) が使用できます。すべての分岐は 2 分割です (2 つのサブグループのみ)。

Tree-AS ノードは既存の CHAID ノードに似ていますが、Tree-AS ノードはビッグデータを処理して 1 つのツリーを作成することを目的に設計されており、結果モデルが SPSS Modeler バージョン 17 で追加された出力ビューアーに表示されます。このノードは、カイ 2 乗統計量 (CHAID) を使用して最適な分割を特定することで、ディシジョン・ツリーを生成します。CHAID をこのように使用することで、非 2 分岐ツリーを生成できます。これは、3 個以上のブランチを持つ分岐が存在することを意味します。対象フィールドおよび入力フィールドは、数値範囲 (連続型) またはカテゴリーとなります。Exhaustive CHAID は CHAID の修正版で、可能性のある分割すべてを調べることで、よりよい結果を得られますが、計算時間も長くなります。

XGBoost Tree© は、ツリー モデルを基本モデルとして使用する勾配ブースティング・アルゴリズムの高度な実装です。ブースティング・アルゴリズムでは、弱い分類子に繰り返し学習させ、それを最終的な強い分類子に追加します。XGBoost ツリーは柔軟性が極めて高く、多くのユーザーを圧倒するほどの多数のパラメータが用意されています。このため、SPSS Modeler の XGBoost ツリー・ノードでは、コア・フィーチャーおよびよく使用されるパラメータが公開されています。このノードは Python で実装されています。

XGBoost© は、勾配ブースティング・アルゴリズムの高度な実装です。ブースティング・アルゴリズムでは、弱い分類子に繰り返し学習させ、それを最終的な強い分類子に追加します。XGBoost は柔軟性が極めて高く、多くのユーザーを圧倒するほどの多数のパラメータが用意されています。このため、SPSS Modeler の XGBoost-AS ノードでは、コア・フィーチャーおよびよく使用されるパラメータが公開されています。XGBoost-AS ノードは Spark で実装されています。

注: Analytic Server で Tree-AS を実行することを選択すると、データ区分ノードが上流にある場合は、モデルの作成に失敗します。この場合、Analytic Server 上の他のモデル作成ノードで自動分類を機能させるには、Tree-AS モデル タイプを選択解除します。

誤分類コスト

状況によっては、特定の誤りコストが他の誤りコストに比べて高いことがあります。例えば、信用リスクの高い申請者を低リスクに分類した場合 (ある種の誤分類) のコストは、リスクの低い申請者を高リスクに分類した場合 (別種の誤分類) よりも高くなります。誤分類コストでは、さまざまな予測の誤りに対し、相対的な重要度を指定できます。

誤分類コストは、基本的には、特定の結果に対して適用される重みです。これらの重みは、モデルに組み込まれ、(コストの高い誤りを防ぐための手段として) 実際に予測値に影響する場合があります。

C5.0 モデルを例外として、誤分類コストは、モデルのスコアリング時には適用されず、自動分類ノード、評価グラフ、または精度分析ノードを使用してモデルをランク付けまたは比較する場合には考慮されません。コストを含むモデルは、コストを含まないモデルに比べてエラーが少なく、全体の精度の項目で高くランク付けされません。ただし、コストが少ない エラーにより組み込まれたバイアスがあるため、実際の問題でパフォーマンスが優れる場合があります。

コスト行列には、可能な各予測カテゴリーや実際のカテゴリーの組み合わせのコストが表示されます。デフォルトでは、すべての誤分類コストが 1.0 に設定されています。コストの値を自分で入力するには、「誤分類コストを使用」 を選択して、コスト行列に独自の値を入力します。

誤分類コストを変更するには、目的の予測値と実際の値の組み合わせに対応するセルを選択して、セルの内容を削除してから、適切なコストを入力してください。コストは自動的に対称的にはなりません。例えば *A* を *B* として誤分類した場合のコストを 2.0 に設定しても、*B* を *A* として誤分類した場合のコストは、変更しない限りデフォルト値 (1.0) のまま変わりません。

自動分類ノードの「破棄」オプション

自動分類ノードの「破棄」タブで、一定の基準を満たさないモデルを自動的に破棄できるようにします。このようなモデルは要約レポートに表示されません。

全体的な精度、モデル内で使用される変数の数の最大閾値を指定できます。また、フラグ型対照の場合、リフト、プロフィット、曲線下の領域の最小閾値を指定できます。リフトとプロフィットは、「モデル」タブで指定されたとおりに決定します。 詳しくは、トピック 68 ページの『自動分類ノードの「モデル」オプション』を参照してください。

オプションで、すべての指定基準を満たすモデルが生成された最初の時点で実行を停止するように、ノードを構成できます。 詳しくは、トピック 66 ページの『自動化モデル作成ノードの停止規則』を参照してください。

自動分類ノードの設定に関するオプション

自動分類ノードの「設定」タブを使用すると、ナゲットに使用可能なスコアリング時間のオプションを事前に設定することができます。

アンサンブル モデルにより生成されたフィールドを除外: 出力から、アンサンブル・ノードに使用する個々のモデルで生成されたすべての追加フィールドを削除します。すべての入力モデルの結合スコアにのみ関心がある場合、このチェック・ボックスを選択します。例えば精度分析ノードまたは評価ノードを使用して結合スコアのと各入力モデルの制度を比較する場合、このオプションが選択解除されていることを確認します。

自動数値ノード

自動数値ノードは、さまざまな方法を使用して 連続型数値範囲の結果についてモデルを推定および比較し、単一のモデル作成実行でさまざまな方法を試用することができます。使用するアルゴリズムを選択し、複数の組み合わせのオプションを検証することができます。例えば、最も良好なパフォーマンスを確認するニューラル・ネットワーク、線型、C&R Tree、CHAID モデルを使用して住宅価格を予測したり、ステップワイズ法、変数増加法、および変数減少法のさまざまな組み合わせを試すこともできます。ノードは、オプションの可能なすべての組み合わせを検証し、指定する指標に基づいて候補モデルをランク付け、スコアリングまたは詳細分析のブランチに最適なモデルを保存します。 詳しくは、トピック 65 ページの『第 5 章 自動化モデル作成ノード』を参照してください。

例　自治体は、固定資産税を正確に見積もり、すべての資産を調査することなく、必要に応じて特定の資産の価格を調整したいと考えています。自動数値ノードを使用して、アナリストは、建築の種類、近傍、大きさおよびその他の既知の要素に基づいて資産の価値を予測する多くのモデルを生成および比較することができます。

要件　1 つの対象フィールド (役割が出力)、少なくとも1 つの入力フィールド (役割が入力)。対象フィールドは、年齢 または 収入 など、連続型 (数値範囲型) フィールドである必要があります。入力フィールドは連続型またはカテゴリー型である場合があり、一部の入力が一部のモデル タイプに適切でないという制限があります。例えば、C&R Tree モデルは入力値としてカテゴリー文字列フィールドを使用できますが、線型モデルではこのフィールドは使用できず、指定されていても無視されます。用件は、個々のモデル作成ノードを使用する場合と同じです。例えば、CHAID モデルは CHAID ノードを使用する場合も自動数値ノードを使用する場合も同じように動作します。

度数および重みフィールド

それは、例えばユーザーが構築データセットは母集団のセクションを低く示すことを認識しているため、または 1 つのレコードが多くの同一ケースを示すためです。度数フィールドは、C&R Tree および CHAID アルゴリズムによって使用できます。重みフィールドは、C&R Tree、CHAID、回帰および GenLin アルゴリズムで使用することができます。その他のモデル タイプでは、これらのフィールドを無視してモデルを構築します。度数および重みフィールドはモデル作成にのみ使用され、モデルの評価またはスコアリングの場合は考慮されません。 詳しくは、トピック 33 ページの『度数フィールドと重みフィールドの使用』を参照してください。

接頭部

自動数値ノードのナゲットにテーブル・ノードを接続する場合は、名前が $ 接頭辞で始まるいくつかの新しい変数がテーブルに存在します。

スコアリング時に生成されるフィールドの名前は、対象フィールドに基づきますが、標準の接頭辞が付加されます。それぞれのモデル タイプでそれぞれの接頭辞を使用します。

例えば、接頭辞 $G、$R、$C は、それぞれ、一般化線型モデル、CHAID モデル、および C5.0 モデルによって生成される予測の接頭辞として使用されます。$X は、通常アンサンブルを使用して生成されます。$XR、$XS、$XF は、対象フィールドがそれぞれ連続型フィールド、カテゴリ型フィールド、フラグ型フィールドの場合に接頭辞として使用されます。

$..E 接頭辞は、連続型対象の予測確信度で使用されます。例えば、$XRE はアンサンブル連続型予測確信度の接頭辞として使用されます。 $GE は、一般化線型モデルに対する確信度の単一予測用の接頭辞です。

サポートするモデル タイプ

サポートするモデル タイプは、ニューラル ネットワーク、C&R Tree、CHAID、回帰、GenLin、最近隣、SVM、XGBoost Linear、GLE、および XGBoost-AS です。詳しくは、 76 ページの『自動数値ノード

の「エキスパート」オプション』を参照してください。

自動数値ノードの「モデル」オプション

自動数値ノードの「モデル」タブで、モデルの比較に使用される基準に沿って、保存されるモデル数を指定することができます。

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

分割モデルを作成 **:** 分割フィールドとして指定される入力フィールドの各値の個別モデルを作成します。詳しくは、 28 ページの『分割モデルの作成』を参照してください。

モデルのランク付け基準 **:** モデルの比較に使用する基準を指定します。

- 相関**:** 各レコードの観察値およびモデルに予測された値の間の Pearson 相関。相関は、強い関係を示す 1 に近い値を持つ 2 つの変数の間の線型連関の測定です。(完全に負の関係の -1 と完全に正の関係の +1 の間の相関値。0 の値は線型関係がないことを示し、負の相関を持つモデルは最も低くランク付けします。)
- フィールド数**:** モデルの予測として使用されるフィールド数。少ないフィールドを使用するモデルを選択すると、データの準備を合理化し、パフォーマンスが向上する場合があります、
- 相対誤差**:** 相対誤差は、モデルで予測された観察値からの平均値からの観察値の分散に対する観察値の偏差です。実際的には、ヌルまたは対象フィールドの平均値を予測として返す定数項モデルに相対して、モデルのパフォーマンスがいかに良好かを比較します。良好なモデルの場合、この値は 1 より小さく、モデルがヌル・モデルより正確であることを示します。1 より大きな相対エラーを含むモデルは、ヌル・モデルより正確ではないため役に立ちません。線型モデルの場合、相対エラーは相関の 2 乗と等しく、追加する新しい情報はありません。非線型モデルの場合、相対エラーは相関と関連せず、モデルのパフォーマンスを評価する追加の測定を提供します。

モデルのランク付けに使用する区分**:** データ区分が使用される場合は、ランクが学習データ区分に基づくか検定データ区分に基づくかを指定できます。大規模なデータセットの場合、モデルの予備的スクリーニングにデータ区分を使用すると、パフォーマンスが著しく改善される可能性があります。

使用モデル数**:** ノードによって作成されるモデル・ナゲットに表示されるモデルの最大数を指定します。上位にランクされたモデルが指定されたランク付けの基準に従って一覧表示されます。モデルの最大数が大きくなると、より多くのモデルの結果を比較できますが、パフォーマンスの速度は低下します。許容できる最大数は 100 です。

予測変数の重要度を計算**:** 重要度の適切な測定基準を作成するモデルの場合、モデル推定時に各予測値の相対重要度を示すグラフを表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。予測変数の重要度によってモデルを計算するために必要な時間が長くなる場合があり、多くの異なるモデル全体で比較する場合はお勧めできません。詳細に検証する少数のモデルに対する分析を絞り込んだ場合に、より有用です。 詳しくは、トピック 44 ページの『予測変数の重要度』を参照してください。

次の場合はモデルを維持しない**:** 相関、相対エラー、および使用されるフィールド数の閾値を指定します。これらの基準のいずれにも一致しないモデルは破棄され、要約レポートには表示されません。

- 小さい相関**:** 要約レポートに含まれるモデルの最小相関 (絶対値)。

- 使用されるフィールド数が大きい:含まれるモデルに使用されるフィールド数の最大数。
- 相対誤差が次より大きい: 含まれるモデルの最大相対エラー。

オプションで、すべての指定基準を満たすモデルが生成された最初の時点で実行を停止するように、ノードを構成できます。 詳しくは、トピック 66 ページの『自動化モデル作成ノードの停止規則』を参照してください。

自動数値ノードの「エキスパート」オプション

自動数値ノードの「エキスパート」タブで、アルゴリズムおよび停止規則を使用し指定するオプションを選択することができます。

モデルの選択。 デフォルトでは、作成対象としてすべてのモデルが選択されます。ただし、Analytic Server を使用している場合、モデルを Analytic Server で実行可能なものだけに制限し、それらをプリセットすることを選択できます。これにより、分割モデルが作成されるか、あるいは大規模データ・セットを処理する準備ができます。

注: 自動数値ノード内での Analytic Server モデルのローカル作成はサポートされていません。

使用されたモデル: 左側のチェック・ボックスを使用して、比較に含めるモデル タイプ （アルゴリズム）を選択します。選択したタイプが多ければ多いほど沢山のモデルが作成されるため、処理時間が長くなります。

モデル タイプ: 使用できるアルゴリズムを表示します (下記参照)。

モデル パラメーター: 各モデル タイプに対し、デフォルト設定を使用するか、「指定」 を選択してオプションを選択することができます。特定の複数のオプションは別のモデル作成ノードで利用できるものと同じであり、複数オプションまたは組み合わせオプションの違いも選択できます。例えば、ニューラル・ネットワーク・モデルを比較する場合に 6 つの学習方法の 1 つを選択するのではなく、単一パスで 6 モデルを学習させるためにそのすべてを選択することができます。

モデルの数: 現在の設定に基づいて各アルゴリズムに対応して作成されるモデルの数を表示します。オプションを組み合わせるとモデルを簡単に追加できるので、特に大きなデータセットを使用する場合はこの数字に細かい注意を払ってください。

単一モデルの構築最大時間を制限 : (K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net およびディシジョン・リスト・モデルのみ) モデルの最大時間制限を設定します。例えば、複雑な交互作用を含む特定のモデルの学習で予期外に長い時間を必要とする場合、すべてのモデルの作成を実行しません。

サポート対象のアルゴリズム

ニューラル・ネット・ノードは、人間の脳が情報を処理する方法を単純化したモデルを使用します。ニューラル・ネットワーク・ノードは、連係する多数の単純な処理単位をシミュレートします。処理単位は、ニューロンを抽象化したものと表現できます。ニューラル・ネットワークは強力な一般関数推定法であり、学習させたり、適用するには、最低限の統計学および数学の知識しか必要ありません。

C&R Tree (分類と回帰ツリー) ノードは、ディシジョン・ツリーを生成し、将来の観測値を予測または分類できるようにします。この方法は再帰的なデータ区分を使用して学習レコードを複数のセグメントに分割し、各ステップで不純性を最小限に抑えます。ツリーのノードが「純粋」であると考えられるのは、ノード中にあるケースの 100% が、対象フィールドのある特定のカテゴリーに分類される場合です。対象フィールドおよび入力フィールドは、数値範囲またはカテゴリー (名義型、順序型、フラグ) が使用できます。すべての分岐は 2 分割です (2 つのサブグループのみ)。

CHAID ノードはディシジョン・ツリーを生成し、カイ二乗統計値を使用して最適な分割を識別します。C&R Tree および QUEST ノードと違って、CHAID は、非 2 分岐ツリーを生成できます。これは、ある分岐が 3 個以上のブランチを持つことを意味します。対象フィールドおよび入力フィールドは、数値範囲 (連続型) またはカテゴリーとなります。Exhaustive CHAID は CHAID の修正版で、可能性のある分割すべてを調べることで、よりよい結果を得られますが、計算時間も長くなります。

線型は、データを要約する一般的な統計手法であり、予測された出力値と実際の出力値の違いを最小限にする直線または面を当てはめることにより予測を行います。

一般化線型モデルは、指定したリンク関数によって従属変数が因子および共変量と線型関係になるよう、一般線型モデルを拡張したものです。さらにこのモデルでは、非正規分布の従属変数を使用することができます。 線型、ロジスティック回帰、カウント・データに関するログ線型モデル、そして区間打切り生存モデルなど、統計モデルの機能性が数多く含まれています。

k が整数である場合、*k* 最近傍 (KNN) ノードは、新しいケースを、予測領域の新しいケースに最も近い *k* 個のオブジェクトのカテゴリーまたは値と関連付けます。類似したケースはお互いに近く、類似していないケースはお互いに離れています。

サポート・ベクター・マシン (SVM) ノードを使用すると、オーバーフィットすることなく、データを 2 つのグループのいずれかに分類することができます。SVM は、非常に多数の入力フィールドを含むデータセットなど、広範なデータセットを処理することができます。

線型モデルは、対象と 1 つまたは複数の予測値との線型の関係に基づいて連続型対象を予測します。

線型サポート・ベクター・マシン (LSVM) ノードを使用すると、オーバーフィットすることなく、データを 2 つのグループのいずれかに分類することができます。LSVM は線型であり、極めて多数のレコードを含むデータセットなど、広範なデータセットを処理することができます。

Random Trees ノードは、既存の C&R Tree ノードと似ていますが、ビッグデータを処理して単一のツリーを作成することを目的に設計されており、結果のモデルが SPSS Modeler バージョン 17 で追加された出力ビューアに表示されます。Random Trees ノードは、将来の観測値を予測または分類するために使用するディシジョン ツリーを生成します。この方法では、再帰的なデータ分岐を使用して、各ステップで不純性を最小限に抑えることで、学習レコードがセグメントに分割されます。ツリー内のノードは、ノード内のケースの 100% が対象フィールドの特定のカテゴリーに分類される場合に、純粋 と見なされます。対象フィールドおよび入力フィールドは、数値範囲またはカテゴリー (名義型、順序型、フラグ) が使用できます。すべての分岐は 2 分割です (2 つのサブグループのみ)。

Tree-AS ノードは既存の CHAID ノードに似ていますが、Tree-AS ノードはビッグデータを処理して 1 つのツリーを作成することを目的に設計されており、結果モデルが SPSS Modeler バージョン 17 で追加された出力ビューアーに表示されます。このノードは、カイ 2 乗統計量 (CHAID) を使用して最適な分割を特定することで、ディシジョン・ツリーを生成します。CHAID をこのように使用することで、非 2 分岐ツリーを生成できます。これは、3 個以上のブランチを持つ分岐が存在することを意味します。対象フィールドおよび入力フィールドは、数値範囲 (連続型) またはカテゴリーとなります。Exhaustive CHAID は CHAID の修正版で、可能性のある分割すべてを調べることで、よりよい結果を得られますが、計算時間も長くなります。

XGBoost Linear© は、線型モデルを基本モデルとして使用する勾配ブースティング・アルゴリズムの高度な実装です。ブースティング・アルゴリズムでは、弱い分類子に繰り返し学習させ、それを最終的な強い分類子に追加します。SPSS Modeler の XGBoost Linear ノードは Python で実装されています。

GLE は、対象を非正規分布とできるように線型モデルを拡張したものであり、指定されたリンク関数を介して因子および共変量に線形に関連し、観測が相関できるようになりました。一般化線型混合モデルには、単純な線型から、非正規分布の縦断的データを取り扱う複雑なマルチレベル・モデルまで、さまざまなモデルがあります。

XGBoost© は、勾配ブースティング・アルゴリズムの高度な実装です。ブースティング・アルゴリズムでは、弱い分類子に繰り返し学習させ、それを最終的な強い分類子に追加します。XGBoost は柔軟性が極めて高く、多くのユーザーを圧倒するほどの多数のパラメータが用意されています。このため、SPSS Modeler の XGBoost-AS ノードでは、コア・フィーチャーおよびよく使用されるパラメータが公開されています。XGBoost-AS ノードは Spark で実装されています。

自動数値ノードの設定に関するオプション

自動数値ノードの「設定」タブを使用すると、ナゲットに使用可能なスコアリング時間のオプションを事前に設定することができます。

アンサンブル モデルにより生成されたフィールドを除外: 出力から、アンサンブル・ノードに使用する個々のモデルで生成されたすべての追加フィールドを削除します。すべての入力モデルの結合スコアにのみ関心がある場合、このチェック・ボックスを選択します。例えば精度分析ノードまたは評価ノードを使用して結合スコアのと各入力モデルの制度を比較する場合、このオプションが選択解除されていることを確認します。

標準誤差を計算 : 対象フィールドが連続型 (数値範囲) の場合、標準誤差の計算がデフォルトで実施され、測定された値または推定された値と真の値との差異を計算し、それらの推定がどれほど近いかを示します。

自動クラスタリング・ノード

自動クラスタリング・ノードは、同様の特性を持つレコードのグループを識別するクラスタリング・モデルを推定し、比較します。ノードは他の自動化モデル作成ノードと同じように動作し、複数の組み合わせのオプションを単一のモデル作成の実行で検証できます。モデルは、クラスター・モデルの有用性をフィルタリングおよびランク付けする基本的な指標を使用して比較し、特定のフィールドの重要度に基づいて指標を提供します。

クラスタリング・モデルは、後続の分析で入力として使用できるグループを識別するために使用されます。例えば、収入など人口統計的な特性に基づいて、または過去に購入したサービスに基づいて顧客のグループを対象に設定する場合があります。検出するグループ数、グループの定義に使用する機能がわからない場合があるため、グループおよびそれらの特性に関する以前の情報を使用せずに実行することができます。対象フィールドを使用せず、真または偽として評価できる特定の予測を返さないため、クラスタリング・モデルは、教師なし学習モデルとも呼ばれます。クラスタリング・モデルの価値は、データのグループ構成を把握し、それらのグループについて役に立つ説明を提供できるかどうかで決まります。 詳しくは、 247 ページの『第 11 章 クラスタリング・モデル』を参照してください。

要件: 重要な特性を定義する 1 つまたは複数のフィールド。真または偽として評価できる特定の予測を行わないため、クラスター・モデルは、対象フィールドを他のモデルと同じ方法で使用しません。代わりに、関連するケースのグループを識別するために使用します。例えば、クラスター・モデルを使用して、特定の顧客が解約するか、またはオファーに反応するかを予測することはできません。ただし、クラスター・モデルを使用して、これらのことを実行する傾向に基づいて、顧客をグループに割り当てることができます。重みフィールドおよび度数フィールドは使用しません。

評価フィールド: 対象フィールドが使用されていない場合、オプションで、モデルを比較する際に使用する評価フィールドを 1 つまたは複数指定できます。クラスター・モデルの有用性は、クラスターがこれらのフィールドをどれだけ良く (または悪く) 識別しているかを測定することによって評価できます。

サポートするモデル タイプ

サポートするモデル タイプは、TwoStep、K-Means、Kohonen、One-Class SVM、および K-Means-AS です。

自動クラスタリング・ノードの「モデル」オプション

自動クラスタリング・ノードの「モデル」タブで、モデルの比較に使用される基準に沿って、保存されるモデル数を指定することができます。

モデル名: ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

モデルのランク付け基準 : モデルの比較およびランク付けに使用する基準を指定します。

- シルエット : クラスターの結束性および分割を測定するインデックス。詳細は以下の「シルエット・ランク付け指標」を参照してください。
- クラスター数: モデル内におけるクラスターの数。

- 最小クラスターのサイズ：最小クラスターのサイズです。
- 最大クラスターのサイズ：最大クラスターのサイズです。
- 最小クラスター/最大クラスター：最大クラスターに対する最小クラスターのサイズの比率。
- 重要度:「フィールド」タブの 「評価」フィールドの重要度。「評価」 フィールドが指定されている場合にのみ計算することができます。

モデルのランク付けに使用する区分: データ区分が使用される場合は、ランクが学習用データセットに基づくか検定セットに基づくかを指定できます。大規模なデータセットの場合、モデルの予備的スクリーニングにデータ区分を使用すると、パフォーマンスが著しく改善される可能性があります。

保存するモデル数：作成されるナゲットに表示されるモデルの最大数を指定します。上位にランクされたモデルが指定されたランク付けの基準に従って一覧表示されます。この制限数を増やすと、パフォーマンスが低下するおそれがあります。許容できる最大数は 100 です。

シルエットランク付け指標

デフォルトのランク付け指標、シルエットは、デフォルト値 0 です。それは、0 より小さい値 (負の数) は、割り当てられたクラスターのケースとポイント間の平均距離が、別のクラスターのポイントへの最小平均距離より大きいためです。そのため、負のシルエットを持つモデルは破棄されます。

実際、ランク付け指標は変更されたシルエット係数で、クラスター結合の概念 (密に結合するクラスターを含むモデルを選択) とクラスター分割の概念 (分割されたクラスターを含むモデルを選択) を結合します。平均シルエット係数は、各ケースに対する次の計算のすべてのケースの平均です。

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

ここで、A は、ケースからケースが所属するクラスターの重心への距離で、B は、ケースから他のすべてのクラスターの重心への距離です。

シルエット係数 (およびその平均) は、-1 (非常に悪いモデルを示す) から 1 (非常に良いモデルを示す) です。平均は、全体のケースのレベル (全体のシルエットを作成) またはクラスターのレベル (クラスター・シルエットを作成) のレベルで計算できます。距離は、ユークリッド距離を使用して計算できます。

自動クラスタリング・ノードの「エキスパート」オプション

自動クラスタリング・ノードの「エキスパート」タブで、データ区分を適用し (利用可能な場合)、使用するアルゴリズムを選択し、停止基準を指定することができます。

モデルの選択。 デフォルトでは、作成対象としてすべてのモデルが選択されます。ただし、Analytic Server を使用している場合、モデルを Analytic Server で実行可能なものだけに制限し、それらをプリセットすることを選択できます。これにより、分割モデルが作成されるか、あるいは大規模データ・セットを処理する準備ができます。

注: 自動クラスタリング・ノード内での Analytic Server モデルのローカル作成はサポートされていません。

使用されたモデル: 左側のチェック・ボックスを使用して、比較に含めるモデル タイプ （アルゴリズム）を選択します。選択したタイプが多ければ多いほど沢山のモデルが作成されるため、処理時間が長くなります。

モデル タイプ: 使用できるアルゴリズムを表示します (下記参照)。

モデル パラメーター**:** 各モデル タイプに対し、デフォルト設定を使用するか、「指定」 を選択してオプションを選択することができます。特定の複数のオプションは別のモデル作成ノードで利用できるものと同じであり、複数オプションまたは組み合わせオプションの違いも選択できます。例えば、ニューラル・ネットワーク・モデルを比較する場合に 6 つの学習方法の 1 つを選択するのではなく、単一パスで 6 モデルを学習させるためにそのすべてを選択することができます。

モデルの数**:** 現在の設定に基づいて各アルゴリズムに対応して作成されるモデルの数を表示します。オプションを組み合わせるとモデルを簡単に追加できるので、特に大きなデータセットを使用する場合はこの数字に細かい注意を払ってください。

単一モデルの構築最大時間を制限 **:** (K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net およびディシジョン・リスト・モデルのみ) モデルの最大時間制限を設定します。例えば、複雑な交互作用を含む特定のモデルの学習で予期外に長い時間を必要とする場合、すべてのモデルの作成を実行しません。

サポート対象のアルゴリズム

K-Means ノードで、データ・セットが異なるグループ (つまりクラスター) へ、クラスタリングされます。この方法で、固定数のクラスターを定義し、クラスターにレコードを繰り返し割り当てて、これ以上調整してもモデルが改善されなくなるまで、クラスターの中心を調整します。*K*-means では、結果を予測するのではなく、入力フィールドのセット内のパターンを明らかにするために、「教師なし学習」として知られるプロセスが使用されます。

Kohonen ノードは、ニューラル・ネットワークの一種であり、データ・セットをクラスター化して異なるグループを形成する目的で使用できます。ネットワークの学習が完了すると、類似のレコードは出力マップで互い近くに表示され、違いの大きいレコードほど離れたところに表示されます。強度の高いユニットを識別するために生成されたモデル内で、各ユニットが獲得した観察の数値を調べることができます。これは、適切なクラスター数についてのヒントになる場合があります。

TwoStep ノードで、2 段階のクラスター化手法が使用されます。最初のステップでは、データを 1 度通過させて、未処理の入力データを管理可能な一連のサブクラスターに圧縮します。2 番目のステップでは、階層クラスター化手法を使用して、サブクラスターをより大きなクラスターに結合させていきます。TwoStep には、学習データに最適なクラスター数を自動的に推定するという利点があります。また、フィールド・タイプの混在や大規模データ・セットも効率よく処理できます。

One-Class SVM ノードでは、教師なし学習アルゴリズムを使用します。このノードは、新規性検知の目的で使用できます。このノードは、与えられたサンプル・セットのソフト境界を検知し、新規ポイントがこのセットに属するか、属さないかを分類します。SPSS Modeler の One-Class SVM モデル作成ノードは Python で実装されており、`scikit-learn`© Python ライブラリーを必要とします。

K-Means は、最も一般的に使用されるクラスタリング アルゴリズムの 1 つです。このアルゴリズムは、データ ポイントをクラスタリングして、事前定義された数のクラスタを作成します。SPSS Modeler の K-Means-AS ノードは Spark で実装されています。 K-Means アルゴリズムについて詳しくは、https://spark.apache.org/docs/2.2.0/ml-clustering.html を参照してください。K-Means-AS ノードでは、カテゴリ変数の場合にワン ホット エンコーディングが自動的に実行されることに留意してください。

自動クラスタリング・ノードの「破棄」オプション

自動クラスタリング・ノードの「破棄」タブで、一定の基準を満たさないモデルを自動的に破棄できるようにします。このようなモデルはモデル・ナゲットに表示されません。

モデルで使用する最小のシルエット値、クラスター数、クラスターのサイズ、評価フィールドの重要度を指定できます。シルエットとクラスターの数とサイズは、モデル作成ノードで指定されてとおりに指定されます。 詳しくは、トピック 79 ページの『自動クラスタリング・ノードの「モデル」オプション』を参照してください。

オプションで、すべての指定基準を満たすモデルが生成された最初の時点で実行を停止するように、ノードを構成できます。 詳しくは、トピック 66 ページの『自動化モデル作成ノードの停止規則』を参照してください。

自動化モデル・ナゲット

自動化モデル作成ノードを実行する場合、ノードはすべての組み合わせのオプションについて候補のモデルを推定し、指定した指標に基づいて候補モデルをランク付け、複合自動化モデル・ナゲットに最良のモデルを保存します。このナゲットには実際、ノードで生成した 1 つまたは複数のモデルのセットが含まれ、スコアリングで使用するモデルを参照および選択することができます。各モデルのモデル タイプおよび構築時間が、モデルのタイプに応じ、その他の指標とともに一覧表示されます。最も興味あるモデルをすぐに識別できるように、これらの列の 1 つを基準にテーブルをソートできます。

- 各モデル・ナゲットを表示するには、ナゲットのアイコンをダブルクリックします。そこから、そのモデルのモデル作成ノードをストリーム・キャンバスに、またはモデル・ナゲットのコピーをモデル・パレットに生成できます。
- 以下に説明しているように、サムネイル・グラフによって、各モデルを迅速に視覚的に評価できます。サムネイルをダブルクリックすると、フルサイズのグラフを生成できます。フルサイズの作図は最大 1000 ポイント表示し、データセットにより多くの作図が含まれている場合はサンプルに基づきます。(散布図の場合のみ、グラフが表示されるごとに再生成されるため、 無作為サンプルまたはデータ区分 (「ランダム シードの設定」が選択されていない場合) の更新など、上流データの変更は散布図を再描画するごとに反映されます。)
- ツールバーを使用して、「モデル」タブの特定の列を表示したり隠したり、テーブルのソートに使用される列を変更したりできます。(列見出しをクリックして、ソートを変更することもできます。)
- 未使用モデルを永久に削除するには「削除」ボタンを使用します。
- 列を並べ替える場合、列見出しをクリックして、列を希望の場所にドラッグします。
- データ区分が使用される場合は、適用されるデータ区分の学習または検定結果を表示するように選択できます。

下記に説明しているように、特定の列は比較されるモデルの種類によって異なります。

2 値の対象

- 2 項モデルの場合、サムネイル・グラフは予測値と重ねて実際の値の分布を表示し、各カテゴリーで正確に予測されたレコードの数を迅速に視覚的に表示します。
- ランク付けの基準は、自動分類モデル作成ノードのオプションと一致します。 詳しくは、トピック 68 ページの『自動分類ノードの「モデル」オプション』を参照してください。
- 最大プロフィットについては、最大値が発生したパーセンタイルについても報告されます。
- 累積リフトについては、ツールバーを使用して選択されているパーセンタイルを変更できます。

名義型対象

- 名義型 (セット型) モデルの場合、サムネイル・グラフは予測値と重ねて実際の値の分布を表示し、各カテゴリーで正確に予測されたレコードの数を迅速に視覚的に表示します。
- ランク付けの基準は、自動分類モデル作成ノードのオプションと一致します。 詳しくは、トピック 68 ページの『自動分類ノードの「モデル」オプション』を参照してください。

連続型対象

- 連続型 (数値範囲型) モデルの場合、各モデルの観察値に対する予測値を表示し、それらの間の相関を迅速に視覚的に表示します。良好なモデルの場合、ポイントはグラフ全体に無作為に散在するのではなく、対角線に沿ってクラスタリングします。
- ランク付けの基準は、自動数値モデル作成ノードのオプションと一致します。 詳しくは、トピック 75 ページの『自動数値ノードの「モデル」オプション』を参照してください。

クラスターの対象

- クラスター・モデルの場合、グラフは各モデルのクラスターに対してカウントし、クラスター分布をすばやく視覚的に表示します。
- ランク付けの基準は、自動クラスタリング・モデル作成ノードのオプションと一致します。 詳しくは、トピック 79 ページの『自動クラスタリング・ノードの「モデル」オプション』を参照してください。

スコアリング用のモデル選択

「使用**?**」 列を使用すると、スコアリングに使用するモデルを選択できます。

- 2 値の対象、名義型対象および数値型対象の場合、複数のスコアリング・モデルを選択し、単一のアンサンブル・モデル・ナゲットにスコアを結合できます。複数のモデルから予測を結合することによって、各モデルの制限を回避でき、モデルの 1 つから取得するより全体的な精度が高い結果が得られます。
- クラスター・モデルの場合、スコアリング・モデルは一度に 1 つだけ選択できます。デフォルトでは、最上位にランクされたモデルが最初に選択されます。

ノードとモデルの生成

複合自動化モデル・ナゲットのコピー、またはモデル・ナゲットが構築された自動化モデル作成ノードのコピーを生成できます。例えば、自動化モデル・ナゲットが構築された元のストリームがない場合、役に立ちます。また、自動化モデル・ナゲットで表示されたどのモデルについても、モデル・ナゲットまたはモデル作成ノードを生成することができます。

自動化モデル作成ナゲット

「生成」メニューから 「モデルをパレットに」 を選択し、自動化モデル・ナゲットをモデル・パレットへ追加します。生成されたモデルは、保存できるためストリームを再実行しなくてもそのまま使用できます。

または、「生成」メニューから 「モデル作成ノードの生成」 を選択し、モデル作成ノードをストリーム・キャンバスへ追加できます。このノードは、モデル作成全体を繰り返し実行しなくても、選択したモデルを再度推定するのに使用できます。

個々のモデル作成ナゲット

1. 「モデル」 メニューで、必要な個々のモデルをダブルクリックします。ナゲットのコピーが新しいダイアログで開きます。

2. 新しいダイアログの「生成」メニューから 「モデルをパレットに」 を選択し、個々のモデル作成ナゲットをモデル・パレットへ追加します。
3. または、新しいダイアログの「生成」メニューから 「モデル作成ノードの生成」 を選択し、モデル作成ノードをストリーム・キャンバスへ追加できます。

評価グラフの生成

2 項モデルの場合のみ、各モデルのパフォーマンスを視覚的に評価し比較する方法を提供する評価グラフを生成できます。評価グラフは、自動数値ノードまたは自動クラスタリング・ノードで生成されたモデルには使用できません。

1. 自動分類の結果のブラウザー内の 「使用?」 列で、評価するモデルを選択します。
2. 「生成」メニューから 「評価グラフ」 を選択します。 「評価グラフ」ダイアログ・ボックスが表示されます。
3. 希望のグラフ タイプとその他のオプションを選択します。

評価グラフ

自動化モデル・ナゲットの「モデル」タブで、表示される各モデルの個々のグラフを表示するようドリル・ダウンします。自動分類ナゲットおよび自動数値ナゲットの場合、「グラフ」タブには結合されたすべてのモデルの結果を反映するグラフおよび予測変数の重要度を表示します。 詳しくは、トピック 44 ページの『予測変数の重要度』を参照してください。

自動分類の場合、棒グラフが表示され、線グラフ (散布図とも呼ばれる) が自動数値について表示されます。

第 6 章 ディシジョン ツリー

ディシジョン・ツリー・モデル

ディシジョン・ツリー・モデルを使用すると、ディシジョン・ルールのセットに基づいて将来の観測値を予測または分類する、分類システムを開発できます。ローンのリスクの高低、購読者と非購読者、投票者と非投票者、バクテリアの種類などの、興味があるいくつかのクラスに分割できるデータがある場合、そのデータを使用して最大限の精度で、新旧のケースを分類するためのルールを作成できます。例えば、年齢やその他の要素に基づいて、クレジットのリスクや、購入の意志を分類するツリーを作成できます。

この方法は、ルール算出としても知られており、いくつかの利点があります。まず、ツリーを参照するときに、モデルが使用する判断の過程が非常に明快です。これは、内部ロジックの理解が困難な他の「ブラック・ボックス」的なモデル作成技法とは対照的です。

2 番目に、プロセスが、決定において実際に問題になる属性だけを自動的にルールに取り込むという点です。ツリーの精度に関係のない属性は無視されます。これにより、データに関する非常に有益な情報が得られます。また、この機能を使用することにより、ニューラル・ネットワークなどの別の手法で学習する前に、関連するフィールドが残るようにデータを減らすことができます。

ディシジョン・ツリー・モデルは、一連の If-Then ルール (ルールセット) に変換できます。多くの場合、このルールを使用すると情報をさらにわかりやすく表示できます。ディシジョン・ツリーによる表示は、データの属性が問題と関連したサブセットにデータを分割または区分する方法を調べる場合に役立ちます。Tree-AS ノードの出力は、他のディシジョン・ツリー・ノードとは異なります。これは、ルールのリストがナゲットに直接含まれており、ルール・セットを作成する必要がないためです。ルール・セットによる表示は、あるグループと結果の関連を調べる場合に役立ちます。例えば、次のルールを使用すると、購入価値のある車のグループのプロファイルを得ることができます。

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

ツリー構築アルゴリズム

分類とセグメント化分析の実行には、いくつかのアルゴリズムを使用できます。これらすべてのアルゴリズムは、基本的には同じ処理を行います。つまり、ユーザーのデータセットのすべてのフィールドを検査して、データをサブグループに分割することで、最も適切な分類または予測が得られるデータを探し出します。このプロセスは再帰的に適用され、ツリーが完了するまで、サブグループは小さい単位に繰り返し分割されます (完了は、何らかの停止基準により定義されます)。ツリー構築で使用される対象フィールドまたは入力フィールドは、使用するアルゴリズムによって、連続型 (数値範囲) またはカテゴリー型で使用できます。連続型目標が使用される場合、回帰ツリーが生成され、カテゴリー目標が使用される場合、分類ツリーが生成されます。

C&R Tree (分類と回帰ツリー) ノードは、ディシジョン・ツリーを生成し、将来の観測値を予測または分類できるようにします。この方法は再帰的なデータ区分を使用して学習レコードを複数のセグメントに分割し、各ステップで不純性を最小限に抑えます。ツリーのノードが「純粋」であると考えられるのは、ノード中にあるケースの 100% が、対象フィールドのある特定のカテゴリーに分類される場合です。対象フィールドおよび入力フィールドは、数値範囲またはカテゴリー (名義型、順序型、フラグ) が使用できます。すべての分岐は 2 分割です (2 つのサブグループのみ)。

CHAID ノードはディシジョン・ツリーを生成し、カイ二乗統計値を使用して最適な分割を識別します。C&R Tree および QUEST ノードと違って、CHAID は、非 2 分岐ツリーを生成できます。これは、ある分岐が 3 個以上のブランチを持つことを意味します。対象フィールドおよび入力フィールドは、数値範囲 (連続型) またはカテゴリーとなります。Exhaustive CHAID は CHAID の修正版で、可能性のある分割すべてを調べることで、よりよい結果を得られますが、計算時間も長くなります。

QUEST ノードには、ディシジョン・ツリーの構築用に2 分岐の方法が用意されています。これは、大規模な C&R Tree 分析が必要とする処理時間を短縮すると同時に、より多くの分割を可能にする入力値が優先される分類ツリー内の傾向を低減するように設計されています。入力フィールドは、数値範囲 (連続型) にできますが、目標変数はカテゴリーでなければなりません。すべての分割は 2 分岐です。

C5.0 ノードは、ディシジョン・ツリーとルール・セットのどちらかを構築します。このモデルは、各レベルで最大の情報の対応をもたらすフィールドに基づいてサンプルを分割します。対象フィールドは、カテゴリーでなければなりません。複数の分割を 2 つ以上のサブグループに分割できます。

Tree-AS ノードは既存の CHAID ノードに似ていますが、Tree-AS ノードはビッグデータを処理して 1 つのツリーを作成することを目的に設計されており、結果モデルが SPSS Modeler バージョン 17 で追加された出力ビューアーに表示されます。このノードは、カイ 2 乗統計量 (CHAID) を使用して最適な分割を特定することで、ディシジョン・ツリーを生成します。CHAID をこのように使用することで、非 2 分岐ツリーを生成できます。これは、3 個以上のブランチを持つ分岐が存在することを意味します。対象フィールドおよび入力フィールドは、数値範囲 (連続型) またはカテゴリーとなります。Exhaustive CHAID は CHAID の修正版で、可能性のある分割すべてを調べることで、よりよい結果を得られますが、計算時間も長くなります。

Random Trees ノードは、既存の C&R Tree ノードと似ていますが、ビッグデータを処理して単一のツリーを作成することを目的に設計されており、結果のモデルが SPSS Modeler バージョン 17 で追加された出力ビューアに表示されます。Random Trees ノードは、将来の観測値を予測または分類するために使用するディシジョン ツリーを生成します。この方法では、再帰的なデータ分岐を使用して、各ステップで不純性を最小限に抑えることで、学習レコードがセグメントに分割されます。ツリー内のノードは、ノード内のケースの 100% が対象フィールドの特定のカテゴリーに分類される場合に、純粋 と見なされます。対象フィールドおよび入力フィールドは、数値範囲またはカテゴリー (名義型、順序型、フラグ) が使用できます。すべての分岐は 2 分割です (2 つのサブグループのみ)。

ツリー・ベースの分析の一般的な使用方法

次にいくつかのツリー・ベースの分析の一般的な使用方法を示します。

セグメンテーション: 特定のクラスのメンバーである可能性が高い人物を特定します。

層化: 高、中、低の各リスクを持つグループなど、複数のカテゴリーのどれか 1 つにケースを割り当てます。

予測: ルールを作成し、そのルールを使用して将来のイベントを予測します。また、予測は、予測属性を連続した値に関連付けようとする試みであるとも言えます。

データの分解と変数のスクリーニング: 形式的なパラメトリック・モデルの構築で使用するために、大規模な変数のセットから有用な予測変数のサブセットを選択します。

交互作用識別: 特定のサブグループにのみ関連する関係を特定し、形式的なパラメトリック・モデル内でそれらの関係を指定します。

カテゴリーの結合と連続変数のバンド化: 情報の損失を最小限に抑えながら、グループ予測カテゴリーと連続変数を再コード化します。

インタラクティブ・ツリー・ビルダー

ツリー・モデルは、自動的に生成することができます。この場合、各レベルの最適な分割が、アルゴリズムによって決定されます。また、インタラクティブ・ツリー・ビルダーを使用して、ユーザーが手動でツリー・モデルを生成することもできます。この方法では、ユーザーのビジネス知識を活かして、ツリーを詳細化または簡素化してから、モデル・ナゲットを保存できます。

1. ストリームを作成して、C&R Tree、CHAID、QUEST のディシジョン・ツリー・ノードから 1 つを追加します。

   注: インタラクティブなツリーの作成は、Tree-AS または C5.0 ツリーではサポートされていません。
2. ノードを開き、「フィールド」タブで、対象フィールドと予測フィールドを選択して、必要に応じて追加のモデル・オプションを指定します。細かい指示については、各ツリー構築ノードのドキュメンテーションを参照してください。
3. 「作成オプション」タブの「目的」パネルで、「インタラクティブ セッションの起動」 を選択します。
4. 「実行」 をクリックして、ツリー・ビルダーを起動します。

現在のツリーが表示されます。また、ルート・ノードから開始されます。1 つ以上のモデルを生成する前に、レベルごとにツリーを編集したり、ゲイン、リスク、および関連する情報にアクセスできます。

コメント

- C&R Tree、CHAID、および QUEST ノードを使用する場合、モデル中で使用される順序型フィールドは、(文字列でなく) 数値ストレージを持っていなければなりません。必要な場合、データ分類ノードを使用して変換できます。
- 必要に応じて、データ区分フィールドを使用して、データを学習およびテスト用のサンプルに分割できます。
- ツリー・ビルダーを使用する代わりに、ストリームの実行時に、他の IBM SPSS Modeler のモデルと同じように、モデル作成ノードから直接モデルを生成できます。 詳しくは、トピック 99 ページの『ツリー・モデルの直接作成』を参照してください。

ツリーの成長と剪定

ツリー・ビルダーの「ビューアー」タブは、現在のツリーを表示します。また、ルート・ノードから開始されます。

1. ツリーを成長させるには、次のメニュー項目を選択してください。

   「ツリー」 > 「ツリーを成長」

システムは、1 つ以上の停止条件が成立するまで、それぞれのブランチを再帰的に分割することで、ツリーを成長させます。各分割時に、使用しているモデル作成方法に基づいて最適な予測フィールドが自動的に選択されます。

2. 代わりに、「ツリーを **1** レベル成長」 を選択すると 1 レベルだけ追加します。
3. 特定のノードの下に枝葉を追加するには、ノードを選択して、「枝葉を成長」 を選択します。
4. 分割の予測フィールドを選択するには、目的のノードを選択して、「ユーザー設定の分割で枝葉を成長」 を選択します。 詳しくは、トピック『ユーザー設定の分割の定義』を参照してください。
5. 枝葉を剪定するには、ノードを選択し、「枝葉を削除」 を選択して、選択したノードを整理します。
6. ツリーから最下位のレベルを削除するには、「**1** レベル削除」 を選択します。
7. C&R Tree および QUEST ツリーの場合のみ、「ツリーの成長と剪定」を選択して、末端ノードの個数に基づいてリスクの予測を調整するコスト・複雑性アルゴリズムに基づいて剪定できます。通常、この方法の方がツリーがより単純になります。 詳しくは、トピック 100 ページの『C&R Tree ノード』を参照してください。

「ビューアー」タブの分割ルールの読み込み

「分割」タブに分割ルールが表示されると、大かっこは隣接する値が範囲内含まれ、かっこは隣接する値が範囲から除外されていることを示します。式 (23,37] は 23 を除き 37 を含む 23 ～ 37 の範囲を示します。「モデル」タブで同じ条件は次のように示されます。

```
Age > 23 and Age <= 37
```

ツリーの成長の中止 **:** (例えば、予想よりも処理に時間がかかるなどの理由で) ツリー成長処理を中止するには、ツールバーの「実行の中止」ボタンをクリックします。

図 *28.* 「実行の中止」ボタン

このボタンは、ツリーの成長時にのみ使用できます。このボタンは、既に追加されているノードはそのままにして、変更を保存せずに、また、ウィンドウも開いたままで、現在の成長処理をボタンが押された時点で停止します。ツリー・ビルダーは、開いたまま残るので、必要に応じて、モデルを生成したり、ディレクティブを更新したり、適当な形式で出力をエクスポートできます。

ユーザー設定の分割の定義

「分割の定義」ダイアログ・ボックスを使用すると、予測フィールドを選択したり、各分割の条件を指定できます。

1. ツリー・ビルダーで、「ビューアー」タブ にあるノードを選択するか、次のメニュー項目を選択してください

   「ツリー」 > 「ユーザー設定の分割で枝葉を成長」
2. ドロップダウン・リストから必要な予測フィールドを選択するか、または、「予測値」 ボタンをクリックして各予測値の詳細を表示します。 詳しくは、トピック 89 ページの『予測フィールドの詳細の表示』を参照してください。
3. 各分割について、デフォルトの条件をそのまま使用するか、または 「ユーザー設定」 を選択して、適切に分岐の条件を指定します。

- 連続型 (数値範囲) 予測フィールドには、「範囲型値の編集」 を使用して、それぞれの新規ノードに一致する値の範囲を指定できます。
- カテゴリー予測フィールドでは、 「セット型の編集」 または 「順序型の値を編集」 フィールドを使用して、新しいノードにマップする特定の値 (または、順序型予測値の場合、値の範囲) を指定できます

4. 「成長」 を選択して、選択した予測フィールドを使用してブランチを再度成長させます。

ツリーは、一般に、停止ルールに関わらず、任意の予測フィールドを使用して分割できます。唯一の例外は、ノードが純粋 (つまり、ケースの 100 % が同じ目標クラスに含まれ、従って分割には何も含まれない場合) であるか、または、選択された予測フィールドが定数 (何も分割できるものがない) の場合です。

欠損値の行き先： CHAID ツリーの場合のみ、与えられた予測フィールドで欠損値が利用可能な場合、ユーザー選択の分割の定義時に、欠損値を特定の子ノードに割り当てるオプションがあります。(C&R Tree および QUEST ツリーでは、欠損値は、そのアルゴリズムで定義されたように、代理変数を使用して処理されます。 詳しくは、トピック『分割の詳細と代理変数』を参照してください。)

予測フィールドの詳細の表示

「予測フィールドの選択」ダイアログ・ボックスは、利用可能な予測値 (または、「競合値」とも呼ばれます) に関する統計値を表示します。この値は現在の分割に使用できます。

- CHAID および Exhaustive CHAID では、各カテゴリー予測フィールドについてのカイ 2 乗統計値の一覧が表示されます。予測フィールドが数値範囲の場合、*F*統計値が表示されます。カイ 2 乗統計値は、分岐フィールドに対する対象フィールドの独立性を示す測定値です。カイ 2 乗統計値が大きい場合、一般的に低い確率と相関があります。つまり、2 つのフィールドが互いに独立である可能性が低いことを意味します。これは、分割が適切であることを示唆します。自由度も投入されます。2 方向分割に比べると、3 方向分割は統計値が大きく、確率が小さくなる傾向があるという事実を考慮に入れているからです。
- C&R Tree および QUEST では、各予測フィールドの改善度が表示されます。改善度が大きいほど、予測フィールドが使用された場合、親ノードと子ノード間の不純度の減少も大きくなります。(純粋なノードとは、すべてのケースが単一の対象カテゴリーに含まれるノードのことで、ツリー内で不純度が低下するにつれて、モデルのデータへの適合度が改善されます。)つまり、高い改善度は一般的に、この種類のツリーにとって有用であることを意味します。使用される不純度測定法は、ツリー構築ノードで指定されます。

分割の詳細と代理変数

「ビューアー」タブでは、任意のノードを選択でき、またツールバーの右側にある「分割情報」ボタンを選択すると、そのノードの分割についての詳細を表示できます。関連のある統計値と一緒に、使用される分割ルールが表示されます。C&R Tree カテゴリー・ツリーでは、改善度と関連付けが表示されます。関連付けは、代理変数と主分岐フィールドの間にある対応の測定値です。「最善の」代理変数とは一般に、分岐フィールドに最もよく似ているものです。C&R Tree および QUEST ツリーでは、主予測フィールドの代わりに使用される代理変数の一覧も表示されます。

選択されたノードの分割を編集するには、代理変数パネルの左側にあるアイコンをクリックして「分割の定義」ダイアログ・ボックスを開きます。(ショートカットとして、アイコンをクリックして、主分割フィールドとして選択する前に、リストから代理変数を選択できます。)

代理変数**:** 適用可能な場合、選択されているノードで、主分岐フィールド用の代理変数が表示されます。代理変数は、あるレコードで主予測値が欠損値の場合に、代わりに使用されるフィールドです。ツリー構築ノードでは、ある分割で使用できる代理変数の個数の最大値を指定します。ただし、実際の個数は、学習用デ

ータに依存します。一般に、欠損値データが多いほど、使用される代理変数も多くなります。他のディシジョン・ツリー・モデルの場合、このタブには何も表示されません。

注: 代理変数をモデルに含めるには、代理変数を学習フェーズ中に識別する必要があります。学習用サンプルに欠損値がない場合、代理変数は識別されません。また、テストまたはスコアリング中に出現した、欠損値を持つレコードは、自動的に最大のレコード数を持つ子ノードに分類されます。テストまたはスコアリング中に欠損値が予測される場合は、その値が学習用サンプルでも欠損値であることを確認してください。代理変数は、CHAID ツリーでは使用できません。

CHAID ツリーでは、代理変数は使用されませんが、ユーザー指定の分割を定義するときにそれらを特定の子ノードに割り当てるオプションが使用できます。 詳しくは、トピック 88 ページの『ユーザー設定の分割の定義』を参照してください。

ツリー・ビューのカスタマイズ

ツリー・ビルダーの「ビューアー」タブは、現在のツリーを表示します。デフォルトでは、ツリーのすべてのブランチが展開されていますが、ブランチは展開も省略も可能で、また必要に応じて他の設定をカスタマイズすることもできます。

- 親ノードの右下隅にあるマイナス記号 (-) をクリックすると、その子ノードがすべて非表示になります。親ノードの右下隅にあるプラス記号 (+) をクリックすると、その子ノードが表示されます。
- 「表示」メニューまたはツールバーを使用すると、ツリーの方向を変更できます (上から下、左から右、または右から左)。
- メイン・ツールバーの「フィールド・ラベルと値ラベルを表示」ボタンをクリックして、フィールドおよび値のラベルの表示を切り替えます。
- 拡大鏡ボタンを使用すると、ビューをズーム・インまたはズームアウトでき、また、ツールバーの右端にあるツリー・マップ ボタンをクリックすると、完全なツリーのダイアグラムを表示できます。
- データ区分フィールドが使用されている場合、学習用およびテスト用の各データ区分間でツリー・ビューを交換できます (「表示 > データ区分」)。テスト用サンプルが表示されているとき、ツリーは表示できますが編集はできません。(現在のデータ区分は、ウィンドウの右下隅のステータスバーに表示されます。)
- 分割情報ボタン (ツールバー右端の「i」ボタン) をクリックすると、現在の分割の詳細が表示されます。 詳しくは、トピック 89 ページの『分割の詳細と代理変数』を参照してください。
- 各ノード内で統計値、グラフ、またはその両方を表示します (次を参照)。

統計値とグラフの表示

ノードの統計 : カテゴリー目標変数フィールドに対して、各ノードのテーブルには各カテゴリーのレコード数と割合、およびノードが表すサンプル全体の割合が表示されています。連続型 (数値範囲) 対象フィールドの場合、テーブルには対象フィールドの平均値、標準偏差、レコード数、および予測フィールドが表示されます。

ノードのグラフ : カテゴリー対象フィールドに対して、対象フィールドの各カテゴリーの割合を表す棒グラフです。テーブルの各行の先頭に表示される色は、ノードのグラフに表示される各目的変数のカテゴリーに対応しています。連続型 (数値範囲) 対象フィールドの場合、グラフにはノード中のレコードに対する対象フィールドのヒストグラムが表示されます。

ゲイン

「ゲイン」タブは、ツリー中のすべてのターミナル・ノードの統計値を表示します。ゲインは、あるノードの平均値 (割合) が、全体の平均値からどの程度、離れているかの尺度を提供します。一般的に、この差が大きいほど、そのツリーは、判断の材料として、より有用です。例えば、あるノードのインデックス値、言い換えると「リフト」値が 148 % である場合、そのノードにあるレコードは、データセット全体と比較して約 1.5 倍の割合で対象カテゴリーに含まれる確率が高いことを示しています。

オーバーフィット防止セットが指定される C&R Tree および QUEST ノードの場合、2 つのセットの統計が表示されます。

- ツリー成長セット - オーバーフィット防止セットが除外されている学習サンプル
- オーバーフィット防止セット

その他の C&R Tree および QUEST インタラクティブ・ツリーの場合、およびすべての CHAID インタラクティブ・ツリーの場合、ツリー成長セットの統計だけが表示されます。

「ゲイン」タブでは、次のことが実行できます。

- ノード・バイ・ノード、累積、分位統計値の表示。
- ゲインまたはプロフィットの表示。
- テーブルとグラフ間でのビューの交換。
- 対象カテゴリーの選択 (カテゴリー・ターゲットのみ)。
- インデックスのパーセンテージに基づいて、テーブルを昇順または降順にソート。複数のデータ区分の統計値が表示されている場合、ソートは、テスト用のサンプルではなく、常に学習用のサンプルに適用されます。

一般に、ゲイン・テーブルで行われた選択は、ツリー・ビューでも更新されます。逆もまた同様です。例えば、テーブルで行を選択した場合、ツリーでも対応するノードが選択されます。

分類ゲイン

分類ツリー (カテゴリー目標変数を持つもの) では、ゲイン・インデックスの割合 (パーセント) から、各ノードで与えられた対象カテゴリーの割合が、どの程度全体の割合から離れているかが解ります。

ノード・バイ・ノード統計値

このビューでは、各ターミナル・ノードごとに 1 行を表示します。例えば、ダイレクト・メール キャンペーンに対する全体の応答は 10% ですが、レコードの 20 % が肯定的な応答ノード X に含まれるとします。そのノードのインデックス割合は 200 %となり、このグループの回答者は、人口全体と比較した場合、2 倍の割合でその製品を購入する可能性があります。

オーバーフィット防止セットが指定される C&R Tree および QUEST ノードの場合、2 つのセットの統計が表示されます。

- ツリー成長セット - オーバーフィット防止セットが除外されている学習サンプル
- オーバーフィット防止セット

その他の C&R Tree および QUEST インタラクティブ・ツリーの場合、およびすべての CHAID インタラクティブ・ツリーの場合、ツリー成長セットの統計だけが表示されます。

ノード **:** (「ビューアー」タブで表示されているように) 現在ノードの ID です。

ノード **:n:**そのノードにあるレコードの総数です。

ノード **(%):** このノードに含まれるデータセット中のすべてのレコードの割合です。

ゲイン**:n:**このノードに含まれる、選択された対象カテゴリーを持つレコード数です。つまり、この対象カテゴリーに含まれるデータセットのすべてのレコードのうち、何個がこのノードにあるか、ということです。

ゲイン **(%):** すべてのデータセットにまたがって、このノードに含まれる対象カテゴリー含まれるデータセット中のすべてのレコードの割合です。

回答 **(%):** 現在のノードにあるレコードが対象カテゴリーに含まれる割合です。この意味での回答は、「ヒット」とも呼ばれることがあります。

インデックス **(%):** データセット全体の回答 % の割合として表された現在のノードの回答 % です。例えば、あるノードのインデックス値が 300% である場合、そのノードにあるレコードは、データセット全体と比較して 3 倍の割合で対象カテゴリーに含まれる確率が高いことを示しています。

累積統計値

累積ビューでは、テーブルは行あたり 1 ノードを表示しますが、統計値が累積の場合、インデックスのパーセンテージにより昇順または降順でソートされます。例えば、降順のソートが適用されている場合、最上位のインデックス割合 (パーセント) を持つノードが最初に表示されます。またそれに続く各行の統計値は、その行とその上の累積値です。

累積インデックス割合 (パーセント) は、より低い回答割合が追加されるにつれて、行から行へと減少します。最終行の累積インデックスは、常に 100 %ですが、これは、この時点ですべてのデータセットが含まれているためです。

分位

このビューでは、テーブル内の各行は、ノードではなく分位を表します。分位は、4 分位 (4 分の1)、5 分位 (5 分の1)、10 分位 (10 分の1)、20 分位 (20分の1)、または 100 分位 (100 分の1) があります。割合 (パーセント) を構成するのに複数のノードが必要な場合、複数のノードを単一でリストを作成できます (例えば、4 分位が表示されているのに最上位の 2 ノードが すべてのケースの 50 % よりも少ないケースしか含んでいない場合)。テーブルの残りは、累積値で、累積ビューと同じ方法で解釈できます。

利益と ROI の分類

分類ツリーでは、利益と ROI (return on investment) についてのゲイン統計値も表示できます。「プロフィットの定義」ダイアログ・ボックスを使用すると、各カテゴリーの歳入と支出を指定できます。

1. 「ゲイン」タブから、ツールバーの「プロフィット」ボタン ($/$ のラベル) をクリックすると、このダイアログ・ボックスにアクセスできます。
2. 対象フィールドの各カテゴリーに歳入と支出の値を入力します。

例えば、それぞれの顧客にダイレクト・メールを送るのに $0.48 かかるとして、肯定的な応答からの歳入が 3 カ月間の購読で、$9.95 とすると、その結果、各 *no* の応答には $0.48 のコストがかかり、また、各 *yes* では、$9.47 を稼ぎます (9.95-0.48 として計算)。

ゲイン・テーブルで、プロフィットは、ターミナル・ノードにある各レコードで歳入から支出を引いた合計として計算されます。**ROI** は、ノードにある合計支出で全プロフィットを割った値です。

コメント

- 統計値をより結果に適合するように表示するための方法であるため、プロフィット値は、ゲイン・テーブルに表示されている平均プロフィットと ROI 値にのみ影響します。ツリー・モデルの基本構造には影響しません。プロフィットを、誤分類コストと混同しないでください。誤分類コストは、ツリー構築ノードで指定され、コスト的な誤りを防ぐための方法としてモデルを因子分析します。
- プロフィット指定は、あるインタラクティブ・ツリー・ビルディングのセッションから、その次のセッションに引き継がれません。

回帰ゲイン

回帰ツリーでは、ノード・バイ・ノード、ノード・バイ・ノード (累積)、および分位ビューから選択できます。平均値がテーブルに表示されます。グラフは、4 分位でのみ利用可能です。

ゲイン・グラフ

グラフは、テーブルの代わりに「ゲイン」タブで表示できます。

1. 「ゲイン」タブで、「4 分位」アイコンを選択します (ツールバーの左から 3 番目です)。(グラフは、ノード・バイ・ノードまたは累積統計値では利用できません。)
2. グラフ アイコンを選択します。
3. 必要に応じて、ドロップダウン・リストから、表示単位 (100 分位、10 分位など) 選択します。
4. 表示方法を変更するには、「ゲイン」、「回答」、または「リフト」を選択します。

ゲイン・グラフ

ゲイン・グラフは、テーブルの 「ゲイン (%)」 列にある値を作図します。ゲインは、次の式を使用して、各増分のツリー中の全ヒット数に対する相対的な割合として定義されています。

(増加中のヒット数 / 全ヒット数) x 100%

このグラフは、ツリー内のすべてのヒットの与えられた割合を捕獲するために、網をどれだけ広げたかを効果的に説明しています。対角線は、モデルが使用されなかった場合に、すべてのサンプルで期待される回答を作図したものです。この場合、1 人が別の人と全く同じように応答するため、回答割合は定数です。売り上げを 2 倍にするには、2 倍の人に質問する必要があります。曲線は、ゲインに基づいてより高位割合にランクされている人だけを含めることで、回答をどの程度、改善できるのかを示しています。例えば、上位の 50 % を含めると、70% を上回る肯定的な応答を網羅できます。カーブが急になるほど、ゲインも高くなります。

リフト・グラフ

リフト・グラフは、テーブルの 「インデックス(%)」 列にある値を作図します。このグラフでは、次の式を使用して、各分位でヒットしたレコードの割合 (パーセント) が、学習データ内の全ヒットの割合と比較されます。

(増加中のヒット数 / 増加中のレコード数) / (全ヒット数 / 全レコード数)

回答グラフ

回答グラフは、テーブルの「回答 %」列にある値を作図します。回答は、ヒットが増加しているレコードの割合で、次の式を使用します。

(増加中の応答数 / 増加中のレコード数) x 100%

ゲインに基づく選択

「ゲインに基づく選択」ダイアログ・ボックスを使用すると、指定されたルールまたは閾値に基づいて、最高 (または最悪) のゲインを持つターミナル・ノードを自動的に選択できます。それから、その選択に基づいて条件抽出ノードを生成できます。

1. 「ゲイン」タブでは、ノード・バイ・ノードまたは累積表示を選択し、また、選択の基準にしたい対象カテゴリーを選択します。(選択は、現在のテーブル表示に基づいており、分位では利用できません。)
2. 「ゲイン」タブで、次のメニュー項目を選択してください。

   「編集」 > 「ターミナル ノードの選択」 > 「ゲインに基づく選択」

   選択条件**:** 一致するノードまたは 一致しないノードを選択することができます。例えば、上位 100 件のレコードを除くすべての レコードを選択することができます。

   ゲイン情報の一致**:** 現在の目標カテゴリーのゲイン統計値に基づいて一致するノードで、次が含まれます。

   - ゲイン、回答、またはリフト (インデックス) が指定された閾値と一致するノード。例えば、回答が 50 % 以上。
   - 目標カテゴリーのゲインに基づく上位 *n* 個のノード。
   - 指定されたレコード数までの上位ノード。
   - 学習用データの指定された割合 (パーセント) までの上位ノード。
3. 「ビューアー」タブで選択を更新するには、「**OK**」をクリックします。
4. 「ビューアー」タブの現在の選択に基づいて、新しい条件抽出ノードを作成するには、「生成」メニューから「条件抽出ノード」を選択します。 詳しくは、トピック 98 ページの『フィルター・ノードおよび条件抽出ノードの生成』を参照してください。

注 : 実際には、レコードまたは割合 (パーセント) ではなくノードを選択していることに注意してください。選択した基準への完全な一致は、必ずしも達成できないかもしれません。システムは、指定されたレベルまでの 完全なノードを選択します。例えば、上位 12 ケースを選択して、10 個を最初のノードに、2 個を 2番目のノードに持っているとき、最初のノードだけが選択されます。

リスク

リスクは、あるレベルで誤分離の機会があることを示しています。「リスク」タブは、ポイント・リスク推定、および (カテゴリー出力の場合) 誤分類表を表示します。

- 数値予測値の場合、リスクは、個々のターミナル・ノードでの分散のプールされた予測値です。
- カテゴリー 予測値では、リスクは、誤って分類されたケースの割合で、任意の事前確率または誤分類コストで調整されます。

ツリー・モデルと結果の保存

インタラクティブ・ツリー・ビルディング セッションの結果は、次の方法を含むさまざまな方法で保存またはエクスポートできます。

- 現在のツリーに基づいてモデルを生成する (「生成 > モデルの生成」)。
- 現在のツリーを生長させるのに使用するディレクティブを保存します。次にツリー構築ノードを実行するときに、現在のツリーは自動的に再成長し、定義した任意のユーザー指定の分割を含みます。
- モデル、ゲイン、およびリスク情報のエクスポート。 詳しくは、トピック 97 ページの『モデル、ゲイン、およびリスク情報のエクスポート』を参照してください。

ツリー・ビルダーと生成されたモデルのどちらからでも、次を実行できます。

- 現在のツリーに基づいて、フィルター・ノードまたは条件抽出ノードを生成する。詳しくは、 98 ページの『フィルター・ノードおよび条件抽出ノードの生成』を参照してください。
- ツリー構造を、ツリーのターミナル・ブランチを定義するルールのセットとして表す、新しいルール・セット・ノードを作成する。詳しくは、 98 ページの『ディシジョン・ツリーからのルールセットの生成』を参照してください。
- さらに、ツリー・モデル・ナゲットについてのみ、モデルを PMML 形式でエクスポートできます。詳しくは、 41 ページの『モデル・パレット』を参照してください。モデルがユーザー定義の分割を含んでいる場合、その情報は、エクスポートされた PMML には保存されません。(分割は保存されますが、アルゴリズムによる選択ではなく、ユーザー定義であるという情報は保存されません。)
- 現在のツリーの選択した部分に基づいてグラフを生成する。なお、ストリーム内のその他のノードに接続している場合のナゲットにのみ生成できます。詳しくは、 130 ページの『グラフの生成』を参照してください。

注 : インタラクティブ・ツリー自体は保存できません。作業内容を失わないようにするには、ツリー・ビルダー・ウィンドウを閉じる前にモデルを生成するか、ディレクティブを更新します。

ツリー・ビルダーからのモデル生成

現在のツリーに基づいてモデルを生成するには、ツリー・ビルダーのメニューから次のメニュー項目を選択してください。

「生成」 > 「モデル」

「新規モデルの生成」ダイアログ・ボックスで、以下のオプションから選択できます。

モデル名**:** ユーザー指定の名前を指定するか、モデル作成ノードの名前に基づいて自動的に名前を生成できます。

ノードの生成先**:** ノードを、「キャンバス」、「**GM** パレット」、または「両方」に追加することができます。

ツリー・ディレクティブを含める **:** 生成されたモデルの現在のツリーからディレクティブを含める場合、このボックスをオンにします。こうすることによって、必要に応じてツリーを再生成できます。 詳しくは、トピック『ツリー成長ディレクティブ』を参照してください。

ツリー成長ディレクティブ

C&R Tree、CHAID、および QUEST モデルの場合、ツリー・ディレクティブはツリーの成長するための条件を、1 回 1 レベルに指定します。毎回、そのノードからインタラクティブ・ツリー・ビルダーが起動されるたびに、ディレクティブが適用されます。

- ディレクティブは、前のインタラクティブ　セッションで生成されたツリーを再生成する場合に、最も安全な方法です。詳しくは、トピック 97 ページの『ツリー・ディレクティブの更新』を参照してください。手動でディレクティブを編集することもできますが、慎重に行ってください。
- ディレクティブは、記述しているツリーの構造に極めて特有のものです。このため、元になっているデータやモデル作成オプションの変更は、以前に有効だったディレクティブのセットをエラーにする場合があります。例えば、CHAID アルゴリズムは、更新されたデータに基づいて 2 方向の分割を 3 方向の分割に変更しますが、以前の 2 方向の分割に基づくディレクティブはすべて失敗します。

注：(ツリー・ビルダーを使用しないで) モデルを直接生成するように選択した場合、すべてのツリー・ディレクティブは無視されます。

ディレクティブの編集

1. 保存されているディレクティブを編集するには、ツリー構築ノードを開いて、「作成オプション」タブの「目的」パネルを選択します。
2. コントロールを有効にするために「インタラクティブ セッションの起動」を選択し、次に「ツリー ディレクティブを使用」を選択し、さらに「ディレクティブ」をクリックします。

ディレクティブ シンタックス

ディレクティブは、ルート・ノードから始めて、ツリーの成長するための条件を指定します。例えば、ツリーを 1 レベル成長させるには次のようにします。

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

予測フィールドが指定されていないため、アルゴリズムが最適な分割を選択します。

最初に分割されるのは、常にルート・ノードです (Index 0) また、両方の子のインデックス値を指定する必要があります (この場合、1 および 2 です)。初めてノード 2 を作成したルートを成長させた場合を除いて、Grow Node Index 2 Children 3 4 の指定は無効です。

ツリーを成長させるには次のようにします。

```
Grow Tree
```

ツリーを成長または剪定するには次のようにします (C&R Tree のみ)。

```
Grow_And_Prune Tree
```

連続型予測フィールドにユーザー指定の分割を指定するには次のようにします。

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

2 個の値で名義型予測フィールドを分割するには次のようにします。

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

複数値で名義型予測フィールドを分割するには次のようにします。

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

順序型予測フィールドで分割するには次のようにします。

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

注：カスタム分割を指定する場合、フィールド名と値 (EDUCATE、GENDER、CHILDS など) は大文字と小文字が区別されます。

CHAID ツリーのディレクティブ

CHAID ツリーのディレクティブは、データやモデルの変更に特に敏感です。これは、C&R Tree および QUEST と違って、2 進分割の使用に制約されないからです。例えば、次のシンタックスは、完全に正しいように見えますが、アルゴリズムがルート・ノードを 3 つ以上の子に分割しようとすると失敗します。

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

CHAID では、ノード 0 は 3 または 4 個の子を持つことが可能なので、シンタックスの 2行目が失敗します。

スクリプトでのディレクティブの使用

三重引用符記号を使用すると、スクリプト中にディレクティブを埋め込むことができます。

ツリー・ディレクティブの更新

インタラクティブ・ツリー構築セッションからの作業を保存するには、現在のツリーを生成するために使用したディレクティブを保存できます。モデル・ナゲットを保存するのと異なり、それ以上は編集できません。これにより、あとで編集するために現在の状態でツリーを再生成できます。

ディレクティブを更新するには、ツリー・ビルダーのメニューから次のメニュー項目を選択してください。

「ファイル」 > 「ディレクティブを更新」

ディレクティブは、ツリーを生成するのに使用したモデル作成ノード (C&R Tree、QUEST、または CHAID のいずれか) に保存され、現在のツリーを再生するために使用できます。 詳しくは、トピック 95 ページの『ツリー成長ディレクティブ』を参照してください。

モデル、ゲイン、およびリスク情報のエクスポート

ツリー・ビルダーから、モデル、ゲイン、およびリスク統計を、目的に応じて、テキスト、HTML、またはイメージの各形式でエクスポートできます。

1. ツリー・ビルダー・ウィンドウで、エクスポートしたいタブまたはビューを選択します。
2. メニューから次の項目を選択します。

   「ファイル」 > 「エクスポート」
3. 目的に応じて、テキスト、**HTML**、グラフを選択し、サブメニューからエクスポートしたい項目を選択します。

適用可能な場合、現在の選択に基づいてエクスポートされます。

テキストまたは **HTML** 形式のエクスポート **:** 学習またはテスト用データ区分が定義されている場合、そのゲインまたはリスク統計値をエクスポートできます。エクスポートは、「ゲイン」タブの現在の選択に基づいています。例えば、ノード・バイ・ノード、累積、または 4 分位統計値を選択できます。

グラフィックのエクスポート**:**「ビューアー」タブに表示されている現在のツリーをエクスポートすることも、また、定義されている場合、学習またはテスト用データ区分のゲイン・グラフをエクスポートすることもできます。利用可能な形式には、*.JPEG*、*.PNG*、およびビットマップ (*.BMP*) が含まれます。ゲインの場合、エクスポートは、「ゲイン」タブの現在の選択に基づきます (グラフが表示されているときのみ利用可能です)。

フィルター・ノードおよび条件抽出ノードの生成

ツリー・ビルダー・ウィンドウから、またはモデル・ナゲットのディシジョン・ツリー・モデルの参照中に、次のメニュー項目を選択してください。

「生成」 > 「フィルター ノード」

または

> 「条件抽出ノード」

フィルター・ノード: 現在のツリーにより使用されていないフィールドをフィルタリングするノードを生成します。これは、アルゴリズムにより重要であると選択されているフィールドのみを含むように、データセットを刈り込むための簡単な方法です。このディシジョン・ツリー・ノードの上流にデータ型ノードがある場合、フィルター・モデル・ナゲットは役割が出力のフィールドをすべて通過させます。

条件抽出ノード: 現在のノードに含まれるすべてのレコードを選択するノードを生成します。このオプションには、「ビューアー」タブ で、1 つ以上のブランチが選択されている必要があります。

モデル・ナゲットはストリーム・キャンバスに配置されます。

ディシジョン・ツリーからのルールセットの生成

ツリー構造を、ツリーのターミナル・ブランチを定義するルールのセットとして表す、ルール・セット・モデル・ナゲットを生成できます。ルール・セットは、より単純なモデルでありながら、ディシジョン・ツリー全体からの重要な情報のほとんどを保持できます。最大の違いは、ルール・セットでは、特定のレコードに複数のルールが当てはまることもあれば、当てはまるルールがまったくないこともある点です。例えば、すべてのルールについて *no* という結果が予測された後に、すべてのルールについて *yes* という結果が予測される場合があります。複数のルールが当てはまる場合は、各ルールに対して、そのルールに関連付けられた確信度に基づいて重み付けされた「票」が割り当てられ、最終的な予測は、対象レコードに当てはまるすべてのルールの重み付きの票を組み合わせて決定されます。適用するルールがない場合、デフォルトの予測がレコードに割り当てられます。

注**:** ルール・セットをスコアリングするときに、ツリーに対するスコアリングと比べて、スコアリングに差異が生じる場合があります。これは、ツリー内の各ターミナル・ブランチが個別にスコアリングされるためです。この違いが顕著になる可能性のある領域として、データ内に欠落値がある場合が挙げられます。

ルール・セットは、カテゴリー対象フィールドを持つツリーからのみ生成できます (回帰ツリーは使用できません)。

ツリー・ビルダー・ウィンドウで、またはディシジョン・ツリー・モデル・ナゲットを参照するときに、メニューから次の項目を選択します。

「生成」 > 「ルール セット」

ルール セット名**:** 新規ルール・セット・モデル・ナゲットの名前を指定します。

ノードの生成先**:** 新規ルール・セット・モデル・ナゲットの場所を制御します。「キャンバス」、「**GM** パレット」、または 「両方」 を選択します。

最小インスタンス数**:** ルールセット・モデル・ナゲットに保持するルールのインスタンス数 (そのルールが当てはまるレコード数) の最小値を指定します。指定した値よりもサポートが少ないルールは、新規ルール・セットに含まれません。

最低確信度**:** ルール・セット・モデル・ナゲットに保持するルールの最低確信度を指定します。指定した値よりも確信度が低いルールは、新規ルール・セットに含まれません。

ツリー・モデルの直接作成

インタラクティブ・ツリー・ビルダーを使用する代わりに、ストリームの実行時に、ノードから直接ディシジョン・ツリー・モデルを作成できます。この方法は、ほとんどのモデル構築ノードで使用できます。インタラクティブ・ツリー・ビルダーでサポートされていない C5.0 ツリーおよび Tree-AS モデルでは、これが使用可能な唯一の方法です。

1. ストリームを作成し、いずれかのディシジョン・ツリー・ノード (C&R Tree、CHAID、QUEST、C5.0、または Tree-AS) を追加します。
2. C&R Tree、QUEST または CHAID の場合、「作成オプション」タブの「目的」パネルで、主な目的のいずれかを選択します。「単一ツリーの構築」を選択する場合は、「モード」が「モデルの生成」に設定されていることを確認してください。

   C5.0 の場合、「Model」タブで、「出力形式」 を 「ディシジョン ツリー」 に設定します。

   Tree-AS の場合は、「作成オプション」タブの「基本」パネルで、「ツリー成長アルゴリズム」タイプを選択します。
3. 対象フィールドと予測値フィールドを選択して、必要に応じて追加のモデル・オプションを指定します。細かい指示については、各ツリー構築ノードのドキュメンテーションを参照してください。
4. ストリームを実行してモデルを生成します。

ツリーの作成に関するコメント

- この方法を使用してツリーを生成するときは、ツリー成長ディレクティブは無視されます。
- インタラクティブか直接かに関わらず、ディシジョン・ツリーの生成方法はどちらも、最終的には同じようなモデルを生成します。単に、ユーザーがどの程度、ツリーの生成を制御できるのかという問題です。

ディシジョン・ツリー・ノード

IBM SPSS Modeler のディシジョン・ツリー・ノードでは、次のツリー構築アルゴリズムを使用できます。

- C&R Tree
- QUEST
- CHAID
- C5.0
- Tree-AS
- Random Trees

詳しくは、トピック 85 ページの『ディシジョン・ツリー・モデル』を参照してください。

アルゴリズムは、データをより小さいサブグループに再帰的に分割してディシジョン・ツリーを構築できるという点では類似していますが、大きく異なる点がいくつかあります。

入力フィールド： 入力フィールド (予測値) は、連続型、カテゴリー型、フラグ型、名義型、または順序型のいずれかになります。

対象フィールド： 指定できる対象フィールドは 1 つだけです。C&R Tree、CHAID、Tree-AS、および Random Trees の場合、対象は連続型、カテゴリー型、フラグ型、名義型、または順序型です。QUEST の場合、カテゴリー型、フラグ型、または名義型となります。C5.0 の場合、対象はフラグ型、名義型または順序型となります。

分割の種類**:** C&R Tree、QUEST、および Random Trees では、2 進分割のみがサポートされます (つまり、ツリーの各ノードは 3 つ以上のブランチには分割できません)。一方、CHAID、C5.0、および Tree-AS は一度に 3 つ以上のブランチへの分割をサポートしています。

分割に使用する方法： アルゴリズムは、分割の指定に使用する基準によって異なります。C&R Tree がカテゴリー型出力を予測する場合、分散計測が使用されます (デフォルトでは Gini 係数ですが、変更できます)。連続型対象フィールドの場合、最小 2 乗偏差 (LSD) 法が使用されます。CHAID および Tree-AS ではカイ 2 乗検定、QUEST ではカテゴリー型予測フィールドにカイ 2 乗検定、連続型入力フィールドに分散分析を使用します。C5.0 の場合、情報理論測定、情報ゲイン率が使用されます。

欠損値の処理]すべてのアルゴリズムでは、予測フィールドの欠損値を許可しますが、それらの処理にはさまざまな方法を使用します。C&R Tree と QUEST は、必要に応じて代理の予測フィールドを使用し、学習時にツリー全体の欠損値を持つレコードの処理を進めます。CHAID は欠損値に別のカテゴリーを作成し、それらをツリー構築に使用できるようにします。C5.0 では分割方法を使用し、分割が欠損値を持つフィールドに基づくノードから、レコードの一部をツリーの各ブランチに渡します。

剪定： C&R Tree、QUEST および C5.0 には、ツリーを完全に成長させ、ツリーの精度に大きく貢献しない下位レベルの分割を削除してツリーを剪定するオプションがあります。ただし、すべてのディシジョン・ツリー・アルゴリズムを使用して、最小サブグループ・サイズを制御し、ブランチのデータ・レコード数が少なくならないようにすることができます。

インタラクティブ ツリー構築**:** C&R Tree、 QUEST および CHAID には、インタラクティブ セッションを起動するオプションがあります。このオプションを使用して、モデルを作成する前に、一度に 1 レベルずつツリーを構築、分割を編集、そしてツリーを剪定することができます。C5.0、Tree-AS、および Random Treesにはインタラクティブ オプションがありません。

事前確率： C&R Tree および QUEST では、カテゴリー型対象フィールドを予測する際に、カテゴリーの事前確率を指定できるようになります。事前確率は、学習データを取り出す母集団内の各対象カテゴリーの全体的な相対頻度の見積もりです。つまり、予測値を知る「前に」、可能性のある各対象値に対して行われる確率の予測です。CHAID、C5.0、Tree-AS、および Random Treesでは、事前確率を指定できません。

ルール セット**:** カテゴリー型対象フィールドのあるモデルの場合、ディシジョン・ツリー・ノードでは、ルール・セットの形式でモデルを作成できます。Tree-AS および Random Treesでは使用できません。この場合、複雑なディシジョン・ツリーに比べて解釈が容易になります。C&R Tree、QUEST および CHAID の場合、インタラクティブ セッションでルール・セットを生成できます。C5.0 の場合、このオプションはモデル作成ノードで指定できます。また、すべてのディシジョン・ツリー・モデルを使用して、モデル・ナゲットからルール・セットを生成できます。 詳しくは、トピック 98 ページの『ディシジョン・ツリーからのルールセットの生成』を参照してください。

C&R Tree ノード

C&R Tree ノードには、ツリーベースの分類と予測の方法があります。この方法では、C5.0 と同様に、帰納的な分岐が行われ、学習レコードが同じような出力フィールド値を持つセグメントに分割されます。まず、入力フィールドが検証されます。分割による不純度の減少が測定され、最適な分割が検出されます。次に、分割によって 2 つのサブグループが定義されます。停止基準が起動されるまで、2 つのサブグループへの分割が繰り返されます。すべての分割は 2 分割 (2 つのサブグループのみ) です。

剪定

C&R Tree は、最初にツリーを成長させるオプションを提供しており、その後、ターミナル・ノードの数に基づいていリスク予測フィールドを調整する、コスト-複雑性アルゴリズムに基づいて剪定します。この方法では、より複雑な基準に基づいて剪定前にツリーを大きく成長させることができ、より優れた交差検証特性を持つより小さいツリーが得られる結果になるかもしれません。ターミナル・ノード数の増加は、一般的に現在の (学習) データのリスクを低減しますが、モデルが事前に検討していないデータに対して一般化されるときには、実際のリスクがより高くなることがあります。極端な場合、学習セットの各レコードに別々のターミナル・ノードを持っているとします。すべてのレコードは、そのノード自体に含まれますが、検討されていない (テスト用) データの分類リスクはほとんど確実に 0 より大きくなるため、リスク予測フィールドは 0 % になります。コスト複雑性測定で、これを補正します。

例: あるケーブル テレビ会社が、ケーブル経由のインタラクティブ ニュース・サービスをどの顧客が購入するかを判断するためのマーケティング調査を委託しました。調査データを使用して、対象フィールドを購読する意図とするストリームを作成し、予測値フィールドに、年齢、性別、教育レベル、収入カテゴリー、毎日テレビ視聴に費やす時間、および子供の数を含めます。C&R Tree ノードをストリームに適用することで、キャンペーンで最高の回答率を得るために、回答を予測し分類することができます。

要件: C&R Tree モデルを学習するには、1 つ以上の入力 フィールドと 1 つの対象 フィールドが必要です。対象フィールドおよび入力フィールドは、連続型 (数値範囲) またはカテゴリーとなります。両方 またはなし が設定されているフィールドは無視されます。モデルで使用されるフィールドは、その型を完全にインスタンス化している必要があり、モデルで使用される順序型 (順序セット) フィールドは、数値ストレージ (文字列不可) である必要があります。必要な場合、データ分類ノードを使用して変換できます。

利点: C&R Tree モデルは、欠損データや大量のフィールドなどの問題が存在する場合に非常に強力です。通常、推定に長い学習時間を必要としません。また、C&R Tree モデルから派生したルールは非常に解釈しやすいので、他のモデルよりわかりやすいという利点があります。C5.0 とは異なり、C&R では、出力フィールドとして連続型フィールドもカテゴリー型フィールドも使用できます。

CHAID ノード

CHAID (Chi-squared Automatic Interaction Detection) は、最適な分割を識別するために、カイ 2 乗統計を使用してディシジョン・ツリーを構築する分類方法です。

CHAID は、最初に、個々の入力フィールドと結果の間のクロス集計を検査し、カイ 2 乗独立性検定を使用して有意確率を検定します。これらの関係の 1 つ以上が統計的に有意である場合、CHAID は、最も有意な入力フィールドを選択します (最小の p 値)。入力フィールドが 3 つ以上のカテゴリーを持っている場合、それらは比較され、結果中で違いが見あたらないカテゴリーは、一緒に折りたたまれます。これは、最も有意差が小さいように見えるカテゴリーのペアを連続的に結合することで行われます。指定された検定レベルで、すべての残りのカテゴリーが異なるとき、カテゴリーのマージ プロセスは停止します。名義型入力フィールドでは、すべてのカテゴリーはマージできます。順序セットでは、隣接するカテゴリーだけがマージできます。

Exhaustive CHAID は、CHAID の修正版で、各予測フィールドですべての可能性のある分割を調べることで、よりよい結果を得られますが、計算時間も長くなります。

要件**:** 入力フィールドは、連続型またはカテゴリー型です。ノードは、各レベルで 2 個以上のサブグループに分割できます。このモデルで使用される順序フィールドは、数値ストレージを持っていなければなりません (文字列不可)。必要な場合、データ分類ノードを使用して変換できます。

利点: C&R Tree および QUEST ノードと異なり、CHAID は、非 2 分岐ツリーを生成できます。これは、ある分岐が 3 個以上のブランチを持てることを意味します。そのため、2 分成長法よりも、幅の広いツリーを生成する傾向があります。CHAID は、入力フィールドのすべてのタイプで動作し、ケースの重み付け変数と度数変数の両方を受け付けます。

QUEST ノード

QUEST (Quick、 Unbiased、Efficient Statistical Tree) は、2 分岐ディシジョン・ツリーの構築用の分類方法です。その開発時における主な同期は、多くの変数または多くのケースを持つ大規模な C&R Tree 分析で必要な処理時間を低減することでした。QUEST の 第2 の目標は、分類ツリー法に見られる、より多くの分割を可能にする入力フィールド、つまり連続型 (数値範囲) 入力フィールド変数や多くのカテゴリーを持つ予測フィールド変数を好む、という傾向を低減することでした。

- QUEST は、ノードで入力フィールド変数を評価するために、有意度検定に基づいて、ルールのシーケンスを使用します。選択用に、シングル テストと同じくらい小さいテストを各入力フィールドについてノードで実行する必要がある場合があります。C&R Tree と異なり、一部の分割は検査されません。また、C&R Tree および CHAID、選択用に入力フィールドを評価するときに、カテゴリーの組み合わせをテストしません。これが分析スピードを速くしています。
- 分割は、対象カテゴリーごとに形成されたグループで選択された入力フィールドを使用して 2 次判別分析を実行することにより決定されます。この方法は、徹底的な検索 (C&R Tree) が最適な分割を決定する場合勝る速度の改善が得られます。

要件: 入力フィールドは、連続型 (数値範囲) にできますが、目標変数はカテゴリーでなければなりません。すべての分割は 2 分岐です。重みフィールドは使用できません。このモデルで使用される順序型 (順序セット) フィールドは、数値ストレージを持っていなければなりません (文字列不可)。必要な場合、データ分類ノードを使用して変換できます。

利点: CHAID と同様に (ただし、C&R Tree とは異なり)、QUEST は統計的な検定を使用して、入力フィールドを使用するかどうかを決定します。また、入力の選択と分割についての問題を切り離し、それぞれに異なる基準を適用します。また、入力フィールド選択と分割についての問題を切り離し、それぞれの異なる基準を適用できます。これは CHAID で制約します。その場合、統計的検定が、変数の選択を決定し、また分割を生成します。同様に、C&R Tree は、入力フィールドの選択と分割の決定に、不純度-変更測度を使用します。

ディシジョン・ツリー・ノードのフィールド・オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

定義済みの役割を使用: このオプションを選択すると、上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) が使用されます。

カスタム・フィールド割り当ての使用: 対象、予測、およびその他の役割を手動で割り当てるには、このオプションを選択します。

フィールド: 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

目標: 1 つのフィールドを予測の対象として選択します。

予測値 **(**入力**):** 1 つ以上のフィールドを予測の入力として選択します。

分析の重み付け**:** (CHAID、C&R Tree、および Trees-AS のみ) フィールドをケースの重みとして使用するには、ここでフィールドを指定します。ケースの重みを使用して、出力フィールドのレベル間の分散における相違を処理します。 詳しくは、トピック 33 ページの『度数フィールドと重みフィールドの使用』を参照してください。

ディシジョン・ツリー・ノードの作成オプション

「作成オプション」タブで、モデルを構築するすべてのオプションを設定します。もちろん、「実行」 ボタンをクリックすると、すべてデフォルト・オプションのモデルが構築されますが、通常は、それぞれの目的で作成をカスタマイズする必要があります。

このタブに含まれる数種類のペインを使用して、モデルに固有のカスタマイズを設定します。

ディシジョン・ツリー・ノード **-** 目的

C&R Tree、QUEST、および CHAID ノードの場合は、「作成オプション」タブの「目的」ペインで、新しいモデルを作成するか、既存のモデルを更新するかを選択できます。ノードの主な目的を、標準モデルの構築、精度または安定性を拡張したモデルの構築、非常に大きなデータセットで使用するモデルの構築に設定することができます。

実行する作業

新規モデルの作成 **:** (デフォルト) このモデル作成ノードを含むストリームを実行するごとに、まったく新しいモデルを作成します。

既存モデルの学習を継続**:** デフォルトでは、モデル作成ノードが実行されるごとに、まったく新しいモデルが作成されます。このオプションを選択すると、ノードによって正常に生成された最後のモデルで学習が続行されます。元のデータにアクセスすることなく既存のモデルを更新またはリフレッシュできます。また、新規レコードまたは更新されたレコードのみ がストリームに適用されるため、パフォーマンスが大幅に向上します。以前のモデルの詳細はモデル作成ノードで保存されるので、以前のモデル・ナゲットがストリームまたは「モデル」パレットでもう使用できない場合でも、このオプションを使用することができます。

注**:** 「単一ツリーの構築」 (C&R Tree、CHAID、および QUEST の場合)、「標準モデルの作成」 (ニューラル・ネットワークおよび線形の場合)、または「非常に大きいデータ セットのモデルを作成」を選択した場合にのみ、このオプションは有効です。

主な目的

- 単一ツリーの構築**:** 標準のディシジョン・ツリー・モデルを 1 つ作成します。一般的に、他の目的オプションを使用して作成されたモデルに比べ、標準モデルがより解釈しやすく、スコアリングの速度が速くなる場合があります。

  注**:** 分割モデルの場合、このオプションを「既存モデルの学習を継続」とともに使用するには、Analytic Server に接続する必要があります。

  モード**:** モデルを構築するために使用する方法を指定します。「モデルの生成」は、ストリームの実行時に自動的にモデルを生成します。「インタラクティブ セッションの起動」は、ツリー・ビルダーを起動します。ツリー・ビルダーを使用すると、モデル・ナゲットを作成する前に、目的に応じて、一度に 1 レベル単位でツリーを構築したり、分割を編集したり、剪定することができます。

ツリー ディレクティブを使用**:** ノードからインタラクティブ・ツリーを生成する場合に、ディレクティブを適用することを指定するには、このオプションを選択します。例えば、1 番目と 2 番目のレベルの分割を指定した場合、これらは、ツリー・ビルダーの起動時に自動的に適用されます。後で、ツリーを再生成できるように、インタラクティブ・ツリー構築セッションからディレクトリーを保存することもできます。 詳しくは、トピック 97 ページの『ツリー・ディレクティブの更新』を参照してください。

- モデル精度の強化 **(**ブースティング**):** ブースティングを使用して、アンサンブルモデルを構築します。ブースティングという特別な方法を使用してモデルの精度を高める場合に、このオプションを選択します。ブースティングは、複数のモデルを順番に作成して行われます。最初のモデルは、通常の方法で作成されます。それから、2 番目のモデルが、最初のモデルで誤分類されたレコードに焦点を当てる方法で構築されます。さらに、2 番目のモデルの誤差に焦点を当てて 3 番目のモデルが作成されます。以下同様に作成されていきます。最後に、モデルのセット全体がケースに適用され、重み付き票決を使用して別々の予測が 1 つの全体予測にまとめられて、ケースが分類されます。ブースティングにより、ディシジョン・ツリー・モデルの精度を大幅に改善することができますが、学習時間は長くなります。
- モデルの安定性を拡張 **(**バグ**):** バギング (ブートストラップ集計) を使用して、アンサンブルモデルを構築します。バギング (ブートストラップ集計) という特別な方法を使用してモデルの安定性を高め、オーバーフィットを避ける場合に、このオプションを選択します。このオプションを選択すると複数のモデルを作成してそれらを結合し、より信頼できる予測を取得します。このオプションを使用して取得されたモデルは、標準モデルと比べて作成およびスコアリングに時間がかかります。
- 非常に大きなデータ セットのモデルを作成**:** 特に大きいデータセットを扱う場合、このオプションを選択して、他の目的オプションを使用してモデルを作成します。このオプションでは、データをより小さいデータ・ブロックに分割し、各ブロックでモデルを作成します。最も正確なモデルが自動的に選択され、単一のモデル・ナゲットに結合されます。この画面で 「既存モデルの学習を継続」 オプションを選択すると、増分モデル更新を実行できます。

  注**:** 非常に大きいデータセットに対してこのオプションを使用するには、IBM SPSS Modeler Server に接続する必要があります。

ディシジョン・ツリー・ノード **-** 基本

ディシジョン・ツリーを構築する方法について、基本オプションを指定します。

ツリー成長アルゴリズム**:** (CHAID および Tree-AS のみ) 使用する **CHAID** アルゴリズムの種類を選択します。**Exhaustive CHAID** は、CHAID の修正版で、各予測フィールドですべての可能性のある分割を調べることで、よりよい結果を得られますが、計算時間も長くなります。

最大ツリー深度**:** ルート・ノード下の最大レベル数 (サンプルが再帰的に分割される回数) を指定します。デフォルトは 5 です。「ユーザー設定」 を選択して値を入力し、異なるレベルを指定します。

剪定 **(C&R Tree** および **QUEST** のみ**)**

オーバーフィットしないようツリーを剪定**:** 剪定では、ツリーの精度にほとんど影響を及ぼさない下位レベルの分割が削除されます。剪定によりツリーを簡素化し、理解しやすくすることができます。また、一般化を改善できる場合もあります。ツリーを剪定せずに完全な状態で使用したい場合は、このオプションを解除してください。

- リスク **(**標準誤差**)** の最大差を設定**:** より許容度の高い剪定ルールを指定することができます。標準誤差ルールを使用した場合、最も単純なツリーが選択されます。そのリスク推定値は、リスクが最も小さいサブツリーのものに近く (ただし、サブツリーのよりも大きい) なります。値は、剪定ツリーとリスク推

定の観点からリスクが最小のツリーの間で許容されるリスク推定の標準誤差数の差を示します。例えば、2 を指定すると、リスク推定 (2 × 標準誤差数) が完全なツリーよりも大きいツリーが選択されます。

最大代理変数**:** 代理変数は、欠損値を処理するための方法です。このアルゴリズムでは、選択した分割フィールドに最も似ている入力フィールドがツリーの分割ごとに検出されます。それらのフィールドが、その分割の代理変数となります。レコードを分類するときに分割フィールドに欠損値があると、代理変数フィールドの値を使用して分割が実行されます。設定値を大きくすると、欠損値をより柔軟に処理できるようになります。ただし、メモリー使用量が増えるので、学習時間が長くなることがあります。

ディシジョン・ツリー・ノード **-** 停止規則

ツリーの構成に関するオプションです。停止ルールは、ツリーの各ブランチの分割をいつ停止するかを指定します。ブランチの最小サイズを設定すると、分割によって非常に小さいサブグループが作成されるのを防止できます。「親枝葉の最小レコード」を指定すると、分割するノード (親) に含まれるレコード数が指定された値よりも小さい場合に、分割を中止します。「子枝葉の最小レコード」を指定すると、分割により作成されるブランチ (子) に含まれるレコード数が指定された値よりも小さい場合に、分割を中止します。

- **100** 分率を使用**:** サイズを学習データ全体の割合で指定します。
- 絶対値を使用**:** サイズをレコード数の絶対値で指定します。

ディシジョン・ツリー・ノード **-** アンサンブル

これらの設定は、「目的」で、ブースティング、バギング、または非常に大きなデータ・セットが要求される場合に起きるアンサンブルの動作を決定します。選択された目的に適用されないオプションは無視されます。

バギングおよび非常に大きなデータ・セット: アンサンブルをスコアリングする場合、基本モデルの予測値を結合するために使用するルールで、アンサンブル・スコア値を計算します。

- カテゴリー型対象のデフォルト結合ルール。 カテゴリー型対象のアンサンブル予測値は、票決、最も高い確率、または最も高い平均確率を使用して結合できます。「票決」は、基本モデルで最も頻繁であり、最も確率が高いカテゴリーを選択します。「高確率」は、すべての基本モデルで最も高い単独の確率に達したカテゴリーを選択します。「最高平均確率」は、基本モデルでカテゴリーの確率が平均化された場合の、最も値の高いカテゴリーを選択します。
- 連続型対象のデフォルトの結合規則: 連続型対象のアンサンブル予測値は、基本モデルの予測値の平均または中央地を使用して結合できます。

モデルの精度を上げることが目的である場合、結合規則の選択が無視されることに注意してください。ブースティングは常に、カテゴリー対象のスコアリングには重み付き多数決を使用し、連続型対象のスコアリングには重み付き中央値を使用します。

ブースティングおよびバギング: バギングの場合は、ブートストラップ数となります。正の整数でなければなりません。

C&R Tree および QUEST ノード - コストと事前確率

誤分類コスト

状況によっては、特定の誤りコストが他の誤りコストに比べて高いことがあります。例えば、信用リスクの高い申請者を低リスクに分類した場合 (ある種の誤分類) のコストは、リスクの低い申請者を高リスクに分類した場合 (別種の誤分類) よりも高くなります。誤分類コストでは、さまざまな予測の誤りに対し、相対的な重要度を指定できます。

誤分類コストは、基本的には、特定の結果に対して適用される重みです。これらの重みは、モデルに組み込まれ、(コストの高い誤りを防ぐための手段として) 実際に予測値に影響する場合があります。

C5.0 モデルを例外として、誤分類コストは、モデルのスコアリング時には適用されず、自動分類ノード、評価グラフ、または精度分析ノードを使用してモデルをランク付けまたは比較する場合には考慮されません。コストを含むモデルは、コストを含まないモデルに比べてエラーが少なく、全体の精度の項目で高くランク付けされません。ただし、コストが少ない エラーにより組み込まれたバイアスがあるため、実際の問題でパフォーマンスが優れる場合があります。

コスト行列には、可能な各予測カテゴリーや実際のカテゴリーの組み合わせのコストが表示されます。デフォルトでは、すべての誤分類コストが 1.0 に設定されています。コストの値を自分で入力するには、「誤分類コストを使用」 を選択して、コスト行列に独自の値を入力します。

誤分類コストを変更するには、目的の予測値と実際の値の組み合わせに対応するセルを選択して、セルの内容を削除してから、適切なコストを入力してください。コストは自動的に対称的にはなりません。例えば *A* を *B* として誤分類した場合のコストを 2.0 に設定しても、*B* を *A* として誤分類した場合のコストは、変更しない限りデフォルト値 (1.0) のまま変わりません。

事前確率

これらのオプションで、カテゴリー対象フィールドを予測する際に、カテゴリーの事前確率を指定できるようになります。事前確率は、学習データを取り出す母集団内の各対象カテゴリーの全体的な相対頻度の見積もりです。つまり、予測値を知る前に、可能性のある各対象値に対して行われる確率の予測です。事前確率を設定する方法は 3 つあります。

- 学習データに基づく**:** これがデフォルトです。事前確率は、学習データ内のカテゴリーの相対度数に基づいて決定されます。
- すべてのクラスで同じ**:** すべてのカテゴリーの事前確率を $1/k$ として定義します (k は対象カテゴリーの数)。
- カスタム**:** 独自の事前確率を指定することもできます。事前確率の開始値が、すべてのクラスで同じに設定されます。各カテゴリーの確率を、ユーザー定義値に調整することができます。特定のカテゴリーの確率を調整するには、そのカテゴリーに対応するテーブル内で確率セルを選択し、セルの内容を削除してから、適切な値を入力してください。

すべてのカテゴリーの事前確率の合計は、1.0 である必要があります (確率の制約)。合計が 1.0 にならない場合、値を自動的に正規化するオプションと警告が表示されます。この自動調整によって、確率の制約を強制しながら、カテゴリー間の比率が維持されます。この調整は、任意の時点で 「正規化」 ボタンをクリックして行うことができます。すべてのカテゴリーで値を均等化するためテーブルをリセットするには、「均等化」 ボタンをクリックします。

誤分類コストを使用して事前確率を調整**:** このオプションにより、(「コスト」タブで指定した) 誤分類コストに基づいて事前確率を調整することができます。これによって、Twoing 不純度測定を使用するツリーに

対して、コスト情報をツリー成長過程に直接取り入れることができます。(このオプションを選択しなかった場合、コスト情報は Twoing 手法に基づいて、レコードの分類とツリーのリスク予測値の算出を行う場合にだけ利用されます。)

CHAID ノード - コスト

状況によっては、特定の誤りコストが他の誤りコストに比べて高いことがあります。例えば、信用リスクの高い申請者を低リスクに分類した場合 (ある種の誤分類) のコストは、リスクの低い申請者を高リスクに分類した場合 (別種の誤分類) よりも高くなります。誤分類コストでは、さまざまな予測の誤りに対し、相対的な重要度を指定できます。

誤分類コストは、基本的には、特定の結果に対して適用される重みです。これらの重みは、モデルに組み込まれ、(コストの高い誤りを防ぐための手段として) 実際に予測値に影響する場合があります。

C5.0 モデルを例外として、誤分類コストは、モデルのスコアリング時には適用されず、自動分類ノード、評価グラフ、または精度分析ノードを使用してモデルをランク付けまたは比較する場合には考慮されません。コストを含むモデルは、コストを含まないモデルに比べてエラーが少なく、全体の精度の項目で高くランク付けされません。ただし、コストが少ない エラーにより組み込まれたバイアスがあるため、実際の問題でパフォーマンスが優れる場合があります。

コスト行列には、可能な各予測カテゴリーや実際のカテゴリーの組み合わせのコストが表示されます。デフォルトでは、すべての誤分類コストが 1.0 に設定されています。コストの値を自分で入力するには、「誤分類コストを使用」 を選択して、コスト行列に独自の値を入力します。

誤分類コストを変更するには、目的の予測値と実際の値の組み合わせに対応するセルを選択して、セルの内容を削除してから、適切なコストを入力してください。コストは自動的に対称的にはなりません。例えば *A* を *B* として誤分類した場合のコストを 2.0 に設定しても、*B* を *A* として誤分類した場合のコストは、変更しない限りデフォルト値 (1.0) のまま変わりません。

C&R Tree ノード: アドバンス

拡張オプションを使用すると、ツリー構築プロセスを微調整できます。

不純度の最小変化 : ツリーに新しい分割を作成する際の不純度の最小変化を指定します。不純度とは、ツリーで定義されたサブグループにおいて、広範囲にわたる出力フィールド値が含まれている程度のことです。カテゴリー変数目標値について、ノードが「純粋」であると考えられるのは、ノード中にあるケースの 100% が、対象フィールドのある特定のカテゴリーに分類される場合です。ツリー構築の目標は、似かよった出力値を持つサブグループを作成することです。つまり、それぞれのノード内における不純度を最小にすることです。ブランチが適切に分割されて不純度が指定値を下回ると、分割は実行されません。

カテゴリー対象の不純度の測度 : カテゴリー変数目標値フィールド用に、ツリーの不純度の測定に使用する方法を指定します。(連続した目標値の場合、このオプションは無視されます。また、最小 **2** 乗偏差不純度測定が常に使用されます。)

- 「**Gini**」は、ブランチの所属カテゴリーの確率に基づく一般的な不純度測定法です。
- 「**Twoing**」は、2 分割を強調する不純度測定法です。分割によってほぼ等サイズのブランチが作成されます。
- 「順序尺度による **Twoing**」は、順序目標変数にのみ適応可能であるため、隣接する目標クラスだけがグループ化できる新しい制約を追加します。このオプションが名義変数目標値用に選択されている場合、デフォルトにより標準 Twoing 測定法が使用されます。

オーバーフィット防止セット。 アルゴリズムは、レコードをモデル作成セットとオーバーフィット防止セットに内部的に分割します。オーバーフィット防止セットは学習時のエラーの追跡に使用されるデータ・レコードの独立したセットで、メソッドがデータ内の偶然変動のモデル作成を行わないようにします。レコードの割合を指定します。デフォルトは 30 です。

結果の再現: ランダム・シードを設定すると、分析を反復することができます。整数を指定、または「生成」をクリックすると、1 ～ 2147483647 の擬似無作為の整数を作成します。

QUEST ノード - アドバンス

拡張オプションを使用すると、ツリー構築プロセスを微調整できます。

分割の有意水準： ノードを分割するための有意水準 (α) を指定します。この値は 0～1 です。値が低いほど、生成されるツリーのノード数が少なくなる傾向があります。

オーバーフィット防止セット。 アルゴリズムは、レコードをモデル作成セットとオーバーフィット防止セットに内部的に分割します。オーバーフィット防止セットは学習時のエラーの追跡に使用されるデータ・レコードの独立したセットで、メソッドがデータ内の偶然変動のモデル作成を行わないようにします。レコードの割合を指定します。デフォルトは 30 です。

結果の再現: ランダム・シードを設定すると、分析を反復することができます。整数を指定、または「生成」をクリックすると、1 ～ 2147483647 の擬似無作為の整数を作成します。

CHAID ノード - アドバンス

拡張オプションを使用すると、ツリー構築プロセスを微調整できます。

分割の有意水準： ノードを分割するための有意水準 (α) を指定します。この値は 0～1 です。値が低いほど、生成されるツリーのノード数が少なくなる傾向があります。

結合の有意水準。 カテゴリーを結合するための有意水準 (α) を指定します。値は、0 より大きく 1 以下でなければなりません。カテゴリーを結合しないようにするには、値を 1 に指定します。連続型対象の場合、最終的なツリーの変数のカテゴリー数は、指定した区間数に一致します。このオプションは、Exhaustive CHAID で利用できません。

Bonferroni メソッドを使用して有意確率値を調整。 予測フィールドの様々なカテゴリーの組み合わせをテストするときに、有意確率の値を調整します。値は、テスト数に基づいて調整されます。テスト数は、カテゴリー数および予測フィールドの測定の尺度と調節関係があります。false-positive エラー率をより制御しやすくなるため、一般にはこの方法が望ましいと言えます。このオプションを無効にすると、本当の差を見つけるための分析能力が向上しますが、false-positive 率が犠牲になります。特に、小さいサンプルの場合にこのオプションをオフにすることをお勧めします。

ノード内の結合したカテゴリーの再分割を許可。CHAID アルゴリズムは、モデルを記述する最も単純なツリーを生成する目的で、カテゴリーの結合を試みます。選択した場合、このオプションは、より良い結果が得られる場合に、マージされたカテゴリーを再分割できるようにします。

カテゴリー対象のカイ **2** 乗 **:** カテゴリー対象では、カイ 2 乗統計値を計算するための方法を指定できます。

- **Pearson :** この手法は、計算は速くなりますが、サンプルが小さい場合には注意して使用する必要があります。
- 尤度比 **:** この方法は、Pearson より強固ですが、計算により長い時間がかかります。小さいサンプルに適した方法です。連続型対象では、この方法が常に使用されます。

期待されるセル度数の最小変化**:** (名義モデルおよび行効果順序モデルの両方のために) セル度数を予測する場合、反復手順 (イプシロン) を使用して、特定の分割のカイ 2 乗検定に使用する最適な予測値に収束させます。ϵは、繰り返しを続けるにはどの程度の変更が発生するのかを決定します。最後の反復での変更が指定された値より小さい場合、反復処理は停止します。アルゴリズムが収束しないという問題がある場合、この値を増やすか、または収束するまでの反復数の最大値を増やします。

収束のための最大反復回数。収束が起きたかどうかに関わらず、停止するまでの最大反復回数を指定します。

オーバーフィット防止セット。 (このオプションは、インタラクティブ・ツリー・ビルダーを使用する場合にのみ使用可能です。) アルゴリズムは、レコードをモデル作成セットとオーバーフィット防止セットに内部的に分割します。オーバーフィット防止セットは学習時のエラーの追跡に使用されるデータ・レコードの独立したセットで、メソッドがデータ内の偶然変動のモデル作成を行わないようにします。レコードの割合を指定します。デフォルトは 30 です。

結果の再現: ランダム・シードを設定すると分析を反復することができます。整数を指定するか、「生成」をクリックします。「生成」をクリックすると、1 から 2147483647 までの整数の疑似乱数が作成されます。

ディシジョン・ツリー・ノードのモデル・オプション

「モデル・オプション」タブで、モデルの名前を指定するか、自動的に名前を生成するかを選択できます。フラグ型対象の調整なしおよび調整済み傾向スコアのほか、予測変数の重要度情報を取得するよう選択することもできます。

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

モデル評価

予測変数の重要度を計算**:** 重要度の適切な測定基準を作成するモデルの場合、モデル推定時に各予測値の相対重要度を示すグラフを表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。モデルによっては、特に大きなデータセットを使用する場合、予測変数の重要度の計算に時間がかかることがあります。そのため、一部のモデルではデフォルトでオフになっています。予測変数の重要度は、ディシジョン・リスト・モデルには使用できません。 詳しくは、 44 ページの『予測変数の重要度』を参照してください。

傾向スコア

傾向スコアは、モデル作成ノードで、またはモデル・ナゲットの「設定」タブで有効にできます。この機能は、選択された対象がフラグ型フィールドである場合にのみ使用できます。 詳しくは、トピック 36 ページの『傾向スコア』を参照してください。

未調整傾向スコアを計算**:** 生の傾向スコアは学習データだけに基づいたモデルから得られます。モデルが *true* 値 (応答する) を予測する場合、傾向は P と同じになります。ここで P は、予測値の確率です。モデルが false 値を予測する場合、傾向は (1 - P) と算出されます。

- モデルを構築する際にこのオプションを選択すると、傾向スコアはそのモデル・ナゲット内でデフォルトで有効になります。ただし、モデル作成ノードで選択したかどうかにかかわらず、モデル・ナゲット内でいつでも生の傾向スコアを有効にできます。

- モデルをスコアリングする際、生の傾向スコアは、標準の接頭辞に *RP* が追加されてフィールドに追加されます。例えば、予測値が *$R-churn* という名前のフィールドにある場合は、傾向スコア フィールドの名前は *$RRP-churn* となります。

調整済み傾向スコアを計算：生の傾向スコアはモデルによって与えられた推定値のみに基づいて算出されますが、これはオーバフィットしている可能性があり、極端に楽観的な傾向が推定されることがあります。調整済み傾向スコアは、テスト・データ区分や検証データ区分に対するモデルの成果を調べて、傾向を調整することによって、より的確な推定を行うものです。

- この設定では、ストリームに有効なデータ区分フィールドが存在している必要があります。
- 生の傾向スコアと違い、調整済み傾向スコアは、モデルを構築するときに計算されなければなりません。そうでなければ、モデル・ナゲットをスコアリングするときにそれらを使用することはできません。
- モデルをスコアリングする際、調整済み傾向スコアは、標準の接頭辞に *AP* が追加されてフィールドに追加されます。例えば、予測値が *$R-churn* という名前のフィールドにある場合は、傾向スコア フィールドの名前は *$RAP-churn* となります。調整済み傾向スコアは、ロジスティック回帰モデルには使用できません。
- 調整済み傾向スコアを計算する場合、計算に使用するテスト・データ区分または検証データ区分はバランス化されていてはいけません。そのため、上流のバランス・ノードで 「学習データのみをバランス」オプションを必ず選択します。さらに、複雑なサンプルが上流にとられた場合は、それによって調整済み傾向スコアが無効になります。
- 調整済み傾向スコアは、「ブーストされた」ツリーまたはルールセット・モデルには使用できません。詳しくは、トピック 130 ページの『ブーストされた C5.0 モデル』を参照してください。

準拠：調整済み傾向スコアが計算されるには、ストリームにデータ区分フィールドが存在していなければなりません。この計算にテスト・データ区分または検証データ区分を使用するかどうかを指定できます。最適な結果を得るには、テスト・データ区分または検証データ区分に、少なくとも、その区分が元のモデルを学習するのに使用したのと同じ数のレコードを含める必要があります。

C5.0 ノード

この機能は SPSS Modeler Professional および SPSS Modeler Premium で使用できます。

このノードでは、C5.0 アルゴリズムを使用して、ディシジョン ツリーまたはルール セットを作成します。C5.0 モデルは、最大の情報の対応をもたらすフィールドに基づいてサンプルを分割します。最初の分割によって定義された各サブサンプルは、異なるフィールドに基づいて再度分割されます。サブサンプルをこれ以上分割できなくなるまで、この過程が繰り返されます。最終的に、最下位レベルの分割が再検証され、モデルの値にほとんど寄与しないレベルが削除 (剪定) されます。

注：C5.0 ノードは、カテゴリー対象のみ予測できます。カテゴリー型 (名義型または順序型) フィールドを含むデータを分析する場合、ノードはリリース 11.0 以前の C5.0 バージョンよりもカテゴリーをグループ化します。

C5.0 では、2 種類のモデルを生成できます。ディシジョン ツリーは、アルゴリズムによって検出された分割の詳細を表しています。各ターミナル (「葉」ノード) は、学習データの特定のサブセットを表します。学習データの各ケースは、ディシジョン・ツリーの 1 つのターミナル・ノードだけに属します。つまり A ディシジョン・ツリーに存在する特定のデータ・レコードに対しては、1 つの予測だけが可能です。

これとは対照的に、ルール セットは、各レコードに対して予測を試みる複数のルールをセットにしたものです。ルール・セットは、ディシジョン・ツリーから派生したもので、ディシジョン・ツリーで検索された

情報を単純化または凝縮したものと言うことができます。ルール・セットは、より単純なモデルでありながら、ディシジョン・ツリー全体からの重要な情報のほとんどを保持できます。ルール・セットとディシジョン・ツリーでは機能が異なるため、属性も異なります。最大の違いは、ルール・セットでは、特定のレコードに複数のルールが適用されることもあれば、ルールがまったく適用されないこともある点です。複数のルールを適用する場合、各ルールに対して、そのルールに関連付けられた確信度に基づいて重み付けされた「票決」が行われ、最終的な予測は、対象レコードに適用するすべてのルールの重み付き票を組み合わせて決定されます。適用するルールがない場合、デフォルトの予測がレコードに割り当てられます。

例: ある医学研究者が、同じ病気に悩む患者に関するデータを収集しています。治療過程において、それぞれの患者に対して 5 種類の薬品の中のいずれかで効果がありました。他のノードとともに、C5.0 モデルを使用して、同じ疾病に苦しむ将来の患者のために適切な薬剤を見つけることができます。

要件: C5.0 モデルを学習するには、1 つのカテゴリー型 (名義型または順序型)「対象」フィールドと、任意のタイプの 1 つ以上の「入力」フィールドが必要です。両方 またはなし が設定されているフィールドは無視されます。モデルで使用するフィールド・タイプは、完全にインスタンス化する必要があります。重みフィールドも指定できます。

利点: C5.0 モデルは、欠損データや大量の入力フィールドがあるような状況で役立ちます。通常、推定に長い学習時間を必要としません。また、C5.0 モデルから派生したルールは非常に解釈しやすいので、他のモデルよりわかりやすいという利点があります。さらに、C5.0 では、分類の精度を向上するための強力なブースティング手法を利用できます。

注**:** C5.0 モデルの構築の速度は、並行処理を有効にすると有利になる可能性があります。

C5.0 ノードの「モデル」オプション

モデル名**:** 作成するモデルの名前を指定します。

- 自動: このオプションを選択すると、対象フィールド名に基づいてモデル名が自動的に生成されます。これはデフォルトです。
- カスタム**:** このノードで作成されたモデル・ナゲットに対して、独自の名前を指定する場合に選択します。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

分割モデルを作成 **:** 分割フィールドとして指定される入力フィールドの各値の個別モデルを作成します。詳しくは、 28 ページの『分割モデルの作成』を参照してください。

出力形式**:** 「ディシジョン ツリー」と「ルール セット」のどちらのモデル・ナゲットを生成するかを指定します。

シンボル値のグループ化**:** このオプションを選択すると、出力フィールドに関して同様のパターンを持つシンボル値の結合が試行されます。このオプションを選択していない場合は、親ノードの分割に使用されたシンボル値フィールドの各値に子ノードが生成されます。例えば、C5.0 が *COLOR* フィールドを分割する場合 (値は *RED*、*GREEN*、および *BLUE*)、デフォルトでは 3 方向の分割が作成されます。ただし、このオプションを選択し、*COLOR = RED* のレコードが *COLOR = BLUE* のレコードと大変似ているような場合、*GREEN* が片方のグループ、そして *BLUE* と *RED*が別のグループにあるような双方向の分割が作成されます。

ブースティングを使用**:** C5.0 アルゴリズムには、ブースティングと呼ばれる、モデルの精度を向上させる特殊な方法があります。この方法では、複数のモデルが順番に作成されます。最初のモデルは、通常の方法

で作成されます。それから、2 番目のモデルが、最初のモデルで誤分類されたレコードに焦点を当てる方法で構築されます。さらに、2 番目のモデルの誤差に焦点を当てて 3 番目のモデルが作成されます。以下同様に作成されていきます。最後に、モデルのセット全体がケースに適用され、重み付き票決を使用して別々の予測が 1 つの全体予測にまとめられて、ケースが分類されます。ブースティングにより、C5.0 モデルの精度を大幅に改善することができますが、学習時間は長くなります。「繰り返し回数」オプションを選択すると、ブースティング・モデルに使用するモデルの数を制御できます。この機能は、Freund & Schapire の研究に基づきながら、ノイズのあるデータを効率的に処理するために独自の改良が加えられています。

交差検証**:** このオプションを選択すると、学習データのサブセットで作成された一連のモデルを使用して、完全なデータセットで作成されたモデルの精度が推定されます。これは、データセットが小さすぎて従来の学習セットと検定セットに分割できない場合に役立ちます。交差検証モデルは、精度の推定の計算後に破棄されます。交差検証に使用する分割数またはモデル数を指定することができます。前のバージョンの IBM SPSS Modeler では、モデルを構築し、交差検証を行う作業は、2 つのそれぞれ別な操作として行われていました。今回のバージョンでは、モデルの構築を個別の手順として行う必要はありません。モデルの構築と交差検証は同時に行われます。

モード**:** 「シンプル」を選択すると、大部分の C5.0 パラメーターが自動的に設定されます。「エキスパート」学習により、学習パラメーターを直接制御できるようになりました。

単純な「モード」オプション

優先**:** デフォルトでは、できる限り精度の高いツリーの作成が試行されます。場合によっては、これがオーバーフィットにつながり、モデルを新しいデータに適用するときに性能が低下することがあります。このような問題を防ぐには、「一般化」を選択し、アルゴリズム設定を調整してください。

注 : 「一般化」を選択して作成したモデルが、他のモデルより一般化を適切に行えるという保証はありません。一般化が重要な場合には、提供されている検定サンプルに照らし合せてモデルを検証してください。

予測されるノイズ **(%):** 学習セット中の予測されるノイズまたは誤データの比率を指定します。

エキスパート・モード・オプション

剪定度 **:** ディシジョン・ツリーまたはルール・セットの剪定の程度を示します。値を大きくすると、より小さく簡単なツリーが生成されます。値を小さくすると、より精度の高いツリーが生成されます。この設定はローカル剪定にだけ適用されます (後述する「グローバル剪定を使用」を参照)。

子枝の最小レコード**:** サブグループのサイズを使用して、ツリーのブランチにおける分割数を制限できます。作成された子ブランチのうち 2 つ以上に、学習セットからのレコードが指定した数以上ある場合に、ツリーのブランチが分割されます。デフォルト値は 2 です。値を大きくすると、データにノイズがある場合の過度な学習が防止されます。

グローバル剪定を使用**:** ツリーは、2 段階で剪定されます。まず、ローカル剪定を実施して、サブツリーの調査と分ブランチの折りたたみを行い、モデルの精度を高めます。次に、ツリーを全体的に捉えて、弱いサブツリーを閉じるグローバル剪定が行われます。デフォルトでは、グローバル剪定が行われます。グローバル剪定を省略するには、このオプションの選択を解除してください。

属性による選別**:** このオプションを選択した場合、C5.0 はモデルの作成を開始する前に、予測フィールドの有用性を調査します。不適切と判明した予測フィールドは、モデルの構築処理から除外されます。このオプションは、予測値フィールドが多いモデルの場合に、オーバーフィットを防止するために役立ちます。

注**:** C5.0 モデルの構築の速度は、並行処理を有効にすると有利になる可能性があります。

Tree-AS ノード

Tree-AS ノードは、分散環境内のデータと共に使用できます。このノードでは、CHAID モデルまたは Exhaustive CHAID モデルのいずれかを使用したディシジョン・ツリーの構築を選択できます。

CHAID (Chi-squared Automatic Interaction Detection) は、最適な分割を識別するために、カイ 2 乗統計を使用してディシジョン・ツリーを構築する分類方法です。

CHAID は、最初に、個々の入力フィールドと結果の間のクロス集計を検査し、カイ 2 乗独立性検定を使用して有意確率を検定します。これらの関係の 1 つ以上が統計的に有意である場合、CHAID は、最も有意な入力フィールドを選択します (最小の p 値)。入力フィールドが 3 つ以上のカテゴリーを持っている場合、それらは比較され、結果中で違いが見あたらないカテゴリーは、一緒に折りたたまれます。これは、最も有意差が小さいように見えるカテゴリーのペアを連続的に結合することで行われます。指定された検定レベルで、すべての残りのカテゴリーが異なるとき、カテゴリーのマージ プロセスは停止します。名義型入力フィールドでは、すべてのカテゴリーはマージできます。順序セットでは、隣接するカテゴリーだけがマージできます。

Exhaustive CHAID は、CHAID の修正版で、各予測フィールドですべての可能性のある分割を調べることで、よりよい結果を得られますが、計算時間も長くなります。

要件: 入力フィールドは、連続型またはカテゴリー型です。ノードは、各レベルで 2 個以上のサブグループに分割できます。このモデルで使用される順序フィールドは、数値ストレージを持っていなければなりません (文字列不可)。必要な場合、データ分類ノードを使用して変換します。

利点: CHAID は、非 2 分岐ツリーを生成できます。これは、ある分岐が 3 個以上のブランチを持つことを意味します。そのため、2 分成長法よりも、幅の広いツリーを生成する傾向があります。CHAID は、入力フィールドのすべてのタイプで動作し、ケースの重み付け変数と度数変数の両方を受け付けます。

Tree-AS ノードのフィールド・オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

定義済みの役割を使用: このオプションを選択すると、上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) が使用されます。

カスタム・フィールド割り当ての使用: 対象、予測、およびその他の役割を手動で割り当てるには、このオプションを選択します。

フィールド: 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

目標: 1 つのフィールドを予測の対象として選択します。

予測値: 1 つ以上のフィールドを予測の入力として選択します。

分析の重み付け: フィールドをケースの重みとして使用するには、ここでフィールドを指定します。ケースの重みを使用して、出力フィールドのレベル間の分散における相違を処理します。 詳しくは、 33 ページの『度数フィールドと重みフィールドの使用』を参照してください。

Tree-AS ノードの作成オプション

「作成オプション」タブで、モデルを構築するすべてのオプションを設定します。もちろん、「実行」 ボタンをクリックすると、すべてデフォルト・オプションのモデルが構築されますが、通常は、それぞれの目的で作成をカスタマイズする必要があります。

このタブに含まれる数種類のペインを使用して、モデルに固有のカスタマイズを設定します。

Tree-AS ノード - 基本

ディシジョン・ツリーを構築する方法について、基本オプションを指定します。

ツリー成長アルゴリズム**:** 使用する **CHAID** アルゴリズムの種類を選択します。**Exhaustive CHAID** は、CHAID の修正版で、各予測フィールドですべての可能性のある分割を調べることで、よりよい結果を得られますが、計算時間も長くなります。

最大ツリー深度**:** ルート・ノード下の最大レベル数 (サンプルが再帰的に分割される回数) を指定します。デフォルトは 5 です。レベル (ノード と呼ばれることもあります) の最大数は 50,000 です。

データ分割**:** 連続型データを使用する場合は、入力をビン分割する必要があります。前のノードでこれを行うことができます。ただし、Tree-AS ノードは、すべての連続型入力を自動的にビン分割します。Tree-AS ノードを使用してデータを自動的にビン分割する場合は、入力が分割される「ビン数」を選択します。データは、等しい度数でビンに分割されます。使用可能なオプションは 2、4、5、10、20、25、50、または 100 です。

Tree-AS ノード - 成長

成長オプションを使用して、ツリー構築プロセスを微調整します。

p 値から効果サイズに切り替えるレコードしきい値**:** ツリーの構築時にモデルが「**p** 値の設定」の使用から「効果サイズの設定」に切り替えるレコード数を指定します。デフォルトは 1,000,000 です。

分割の有意水準**:** ノードを分割するための有意水準 (*α*) を指定します。この値は 0.01 から 0.99 までです。値が低いほど、生成されるツリーのノード数が少なくなる傾向があります。

結合の有意水準**:** カテゴリーを結合するための有意水準 (*α*) を指定します。この値は 0.01 から 0.99 までです。連続型対象の場合、最終的なツリーの変数のカテゴリー数は、指定した区間数に一致します。このオプションは、Exhaustive CHAID で利用できません。

Bonferroni メソッドを使用して有意確率値を調整**:** 予測フィールドのさまざまなカテゴリーの組み合わせをテストするときに、有意確率値を調整します。値は、テスト数に基づいて調整されます。テスト数は、カテゴリー数および予測フィールドの測定の尺度と調節関係があります。false-positive エラー率をより制御しやすくなるため、一般にはこの方法が望ましいと言えます。このオプションを無効にすると、真の差を検出するための分析能力が向上しますが、その代償として false-positive 率が増加します。特に、小さいサンプルの場合にこのオプションをオフにすることをお勧めします。

効果サイズしきい値 **(**連続型対象のみ**):** 連続型対象を使用するときに、ノードの分割時およびカテゴリーの結合時に使用する効果サイズしきい値を設定します。この値は 0.01 から 0.99 までです。

効果サイズしきい値 **(**カテゴリー型対象のみ**):** カテゴリー型対象を使用するときに、ノードの分割時およびカテゴリーの結合時に使用する効果サイズしきい値を設定します。この値は 0.01 から 0.99 までです。

ノード内の結合したカテゴリーの再分割を許可**:** CHAID アルゴリズムは、モデルを記述する最も単純なツリーを生成する目的で、カテゴリーの結合を試みます。選択した場合、このオプションは、より良い結果が得られる場合に、マージされたカテゴリーを再分割できるようにします。

葉ノードのグループ化の有意水準**:** 葉ノードのグループが形成される方法または異常な葉ノードが識別される方法を決定する有意水準を指定します。

カテゴリー対象のカイ **2** 乗**:** カテゴリー対象では、カイ 2 乗統計値を計算するための方法を指定できます。

- **Pearson:** この手法は、計算は速くなりますが、サンプルが小さい場合には注意して使用する必要があります。
- 尤度比 この方法は、Pearson より強固ですが、計算により長い時間がかかります。小さいサンプルに適した方法です。連続型対象では、この方法が常に使用されます。

Tree-AS ノード - 停止規則

ツリーの構成に関するオプションです。停止ルールは、ツリーの各ブランチの分割をいつ停止するかを指定します。ブランチの最小サイズを設定すると、分割によって非常に小さいサブグループが作成されるのを防止できます。「親枝葉の最小レコード」を指定すると、分割するノード (親) に含まれるレコード数が指定された値よりも小さい場合に、分割を中止します。「子枝葉の最小レコード」を指定すると、分割により作成されるブランチ (子) に含まれるレコード数が指定された値よりも小さい場合に、分割を中止します。

- **100** 分率を使用**:** サイズを学習データ全体の割合で指定します。
- 絶対値を使用**:** サイズをレコード数の絶対値で指定します。

期待されるセル度数の最小変化**:** (名義モデルおよび行効果順序モデルの両方のために) セル度数を予測する場合、反復手順 (イプシロン) を使用して、特定の分割のカイ 2 乗検定に使用する最適な予測値に収束させます。最後の反復での変更が指定された値より小さい場合、反復処理は停止します。アルゴリズムが収束しないという問題がある場合、この値を増やすか、または収束するまでの反復数の最大値を増やします。

収束のための最大反復回数**:** 収束が起きたかどうかに関わらず、停止するまでの最大反復回数を指定します。

Tree-AS ノード - コスト

状況によっては、特定の誤りコストが他の誤りコストに比べて高いことがあります。例えば、信用リスクの高い申請者を低リスクに分類した場合 (ある種の誤分類) のコストは、リスクの低い申請者を高リスクに分類した場合 (別種の誤分類) よりも高くなります。誤分類コストでは、さまざまな予測の誤りに対し、相対的な重要度を指定できます。

誤分類コストは、基本的には、特定の結果に対して適用される重みです。これらの重みは、モデルに組み込まれ、(コストの高い誤りを防ぐための手段として) 実際に予測値に影響する場合があります。

コストを含むモデルは、コストを含まないモデルに比べてエラーが多く、全体の精度に関して低くランク付けされる可能性があります。ただし、コストが少ないエラーを優先するという組み込まれたバイアスがあるため、実際面ではパフォーマンスが優れている場合があります。

コスト行列には、可能な各予測カテゴリーや実際のカテゴリーの組み合わせのコストが表示されます。デフォルトでは、すべての誤分類コストが 1.0 に設定されています。コストの値を自分で入力するには、「誤分類コストを使用」 を選択して、コスト行列に独自の値を入力します。

誤分類コストを変更するには、目的の予測値と実際の値の組み合わせに対応するセルを選択して、セルの内容を削除してから、適切なコストを入力してください。コストは自動的に対称的にはなりません。例えば *A* を *B* として誤分類した場合のコストを 2.0 に設定しても、*B* を *A* として誤分類した場合のコストは、変更しない限りデフォルト値 (1.0) のまま変わりません。

順序型対象の場合にのみ、「順序型対象のデフォルトのコスト増加」を選択し、コスト行列内のデフォルト値を設定することができます。使用可能なオプションを次のリストに示します。

- 増加なし **(No increase)** - すべての正しい予測でデフォルト値の 1.0 を使用します。
- 線型 - 後続の不正確な予測が検出されるごとに、コストは 1 ずつ増加します。
- 平方 **(Square)** - 後続の不正確な予測が検出されるごとに、コストは線型値の平方になります。この場合、値は 1、4、9 などになります。
- カスタム - テーブル内のいずれかの値を手動で編集すると、ドロップダウン・オプションは、「カスタム」に自動的に変更されます。ドロップダウン選択を他のいずれかのオプションに変更すると、編集された値は、選択されたオプションの値と置き換えられます。

Tree-AS ノードのモデル・オプション

「モデル・オプション」タブで、モデルの名前を指定するか、自動的に名前を生成するかを選択できます。モデルのスコアリング中に、確信度値を計算し、識別 ID を追加するように選択することもできます。

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

確信度の計算**:** モデルがスコアリングされるときに、確信度フィールドを追加するには、このチェック・ボックスを選択します。

ルール識別子**:** モデルがスコアリングされるときに、レコードが割り当てられた葉ノードの ID を含むフィールドを追加するには、このチェック・ボックスを選択します。

Tree-AS モデル・ナゲット

Tree-AS モデル・ナゲット出力

Tree-AS モデルを作成した後で、以下の情報が出力ビューアーで使用可能になります。

「モデル情報」テーブル

「モデル情報」テーブルは、モデルについての重要な情報を提供します。テーブルは次のようなハイレベルなモデル設定を特定します。

- 使用されるアルゴリズム・タイプ。CHAID または Exhaustive CHAID のいずれか。
- データ型ノードまたは Tree-AS ノードのいずれかの「フィールド」タブで選択された対象フィールドの名前。
- データ型ノードまたは Tree-AS ノードのいずれかの「フィールド」タブで予測フィールドとして選択されたフィールドの名前。
- データ内のレコード数。度数の重みを付けてモデルを構築する場合は、この値が、ツリーのベースとなるレコード数を表す重み付きの有効なカウントとなります。
- 生成されたツリー内の葉ノード の数。
- ツリー内のレベル数。つまり、ツリーの深さ。

予測変数の重要度

「予測変数の重要度」グラフは、モデル内の上位 10 個の入力 (予測値) の重要度を棒グラフとして表示します。

グラフ内に 10 個を超えるフィールドがある場合は、グラフの下のスライダーを使用して、グラフ内に含まれる予測値の選択を変更できます。スライダー上のインディケーター・マークは固定幅であり、スライダー上の各マークは 10 個のフィールドを表します。スライダーに沿ってインディケーター・マークを移動して、予測変数の重要度の順序で並べられた次の 10 個または前の 10 個のフィールドを表示できます。

グラフをダブルクリックして、グラフ設定を編集するための別個のダイアログ・ボックスを開くことができます。例えば、グラフのサイズ、使用されるフォントのサイズと色などの項目を修正できます。この別個の編集ダイアログ・ボックスを閉じると、「出力」タブに表示されるグラフに変更が適用されます。

「上位ディシジョン・ルール (Top Decision Rules)」テーブル

デフォルトでは、この対話式のテーブルは、葉ノード内に含まれる合計レコードの割合に基づいて、出力内の上位 5 個の葉ノードのルールの統計量を表示します。

テーブルをダブルクリックして、テーブル内に表示されるルール情報を編集するための別個のダイアログ・ボックスを開くことができます。ダイアログ・ボックス内に表示される情報および使用可能なオプションは、対象のデータ型 (カテゴリー型または連続型など) によって異なります。

テーブルには次のルール情報が表示されます。

- ルール ID
- ルールの適用方法および構成内容の詳細
- 各ルールのレコード件数。度数の重みを付けてモデルを構築する場合は、この値が、ツリーのベースとなるレコード数を表す重み付きの有効なカウントとなります。
- 各ルールのレコード割合

さらに、連続型対象の場合は、各ルールの「平均」値を示す追加の列がテーブル内に表示されます。

以下の「テーブル・コンテンツ **(Table contents)**」オプションを使用して、ルール・テーブル・レイアウトを変更できます。

- 上位ディシジョン・ルール **(Top decision rules):** 葉ノード内に含まれる合計レコードの割合によって、上位 5 個のディシジョン・ルールがソートされます。
- すべてのルール **(All rules):** テーブルには、モデルによって生成されたすべての葉ノードが含まれますが、1 ページあたり 20 個のルールのみが表示されます。このレイアウトを選択すると、「**ID** でルールを検索 **(Find rule by ID)**」および「ページ」の追加オプションを使用して、ルールを検索できます。

さらに、カテゴリー型対象の場合は、「カテゴリーごとの上位ルール **(Top rules by category)**」オプションを使用して、ルール・テーブル・レイアウトを変更できます。選択した「対象カテゴリー」の合計レコードの割合によって、上位 5 個のディシジョン・ルールがソートされます。

ルール・テーブルのレイアウトを変更する場合は、ダイアログ・ボックスの左上にある「ビューアーにコピー」ボタンをクリックして、変更されたルール・テーブルを元の出力ビューアーにコピーできます。

Tree-AS モデル・ナゲットの設定

Tree-AS モデル・ナゲットの「設定」タブで、モデル・スコアリング中の確信度のオプションと SQL 生成のオプションを指定できます。このタブは、モデル・ナゲットがストリームに追加された後にのみ使用できます。

確信度の計算**:** スコアリング操作に確信度を含めるには、このチェック・ボックスを選択します。データベースでモデルをスコアリングするときに、確信度を除外すると、より効率的な SQL を生成できます。回帰ツリーでは確信度は割り当てられないことに注意してください。

ルール識別子**:** 各レコードが割り当てられるターミナル・ノードの ID を示すフィールドをスコアリング出力に追加するには、このチェック・ボックスを選択します。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL の生成方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

Random Trees ノード

Random Trees ノードは、分散環境内のデータと共に使用できます。このノードでは、複数のディシジョン ツリーで構成されるアンサンブル モデルを構築します。

Random Trees ノードは、分類と回帰ツリー方法論に基づいて作成される、ツリー・ベースの分類および予測方法です。この予測方法では、C&R Tree と同様に、再帰的な区分化を使用して、学習レコードが同様の出力フィールド値を持つセグメントに分割されます。ノードでは、まず使用可能な入力フィールドが検証され、分割による不純度の減少を測定することで最適な分割が検出されます。分割によって 2 つのサブグループが定義されます。停止基準が起動されるまで、2 つのサブグループへの分割が繰り返されます。すべての分割は 2 分割 (2 つのサブグループのみ) です。

Random Trees ノードでは、置き換えを伴うブートストラップ サンプリングを使用してサンプル データを生成します。サンプル データは、ツリー モデルを成長させるために使用します。ツリーの成長中、Random Trees はデータを再度サンプリングしません。代わりに、予測値の一部をランダムに選択し、最良の予測値を使用してツリー ノードを分割します。このプロセスは、各ツリー ノードの分割時に繰り返されます。これがランダム フォレストにおけるツリーの成長の基本的な概念です。

Random Trees では、C&R Tree に似たツリーを使用します。そのようなツリーは 2 分岐であるため、分割用の各フィールドは、2 つの枝に分岐します。複数のカテゴリがあるカテゴリ型フィールドの場合、カテゴリは、内部分割基準に基づいて、2 つのグループにグループ化されます。各ツリーは、できる限り大きくなるように成長します (剪定は行われません)。スコアリングでは、Random Trees は、多数決 (分類の場合) または平均 (回帰の場合) によって個別のツリーのスコアを結合します。

Random Trees と C&R Tree には、以下のような違いがあります。

- Random Trees ノードは、指定した数の予測値をランダムに選択し、選択したもので最良の予測値を使用してノードを分割します。対照的に、C&R Tree は、すべての予測値で最良のものを探します。
- Random Trees の各ツリーは通常、各葉ノードに単一のレコードが含まれるまで十分に成長します。そのため、ツリーの深さは非常に大きくなる可能性があります。一方、標準的な C&R Tree では、ツリーの成長にさまざまな停止規則が使用され、通常、はるかに浅いツリーになります。

Random Treesでは、C&R Tree と比較して以下の 2 つの機能が追加されています。

- 最初の機能であるバギング では、元のデータセットから置換してサンプリングすることによって、学習データセットの複製が作成されます。このアクションによって、元のデータセットと同じサイズのブートストラップ サンプルが作成され、それを元に各複製の上にコンポーネント モデル が作成されます。同時にこれらのコンポーネント・モデルがアンサンブル・モデルを形成します。
- 2 番目の機能では、ツリーの各分割において、入力フィールドのサンプリングのみが不純度測定の対象となります。

要件**:** Random Trees モデルを学習するには、1 つ以上の入力 フィールドと 1 つの対象 フィールドが必要です。対象フィールドおよび入力フィールドは、連続型 (数値範囲) またはカテゴリーとなります。両方またはなし に設定されているフィールドは無視されます。モデルで使用されるフィールドは、その型を完全にインスタンス化している必要があり、モデルで使用されるすべての順序型 (順序セット) フィールドは、数値ストレージ (文字列不可) である必要があります。必要な場合、データ分類ノードを使用して変換できます。

利点**:** Random Trees モデルは、大規模なデータセットと大量のフィールドを扱う場合に堅固なモデルです。バギングとフィールド サンプリングを使用することにより、このモデルはオーバーフィットになる可能性が小さくなるため、検定で得られる結果は新規データを使用した場合にも繰り返される可能性が高くなります。

Random Trees ノードのフィールド・オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

定義済みの役割を使用**:** このオプションを選択すると、上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) が使用されます。

カスタム・フィールド割り当ての使用: 対象、予測、およびその他の役割を手動で割り当てるには、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

目標: 1 つのフィールドを予測の対象として選択します。

予測値**:** 1 つ以上のフィールドを予測の入力として選択します。

分析の重み付け**:** フィールドをケースの重みとして使用するには、ここでフィールドを指定します。ケースの重みを使用して、出力フィールドのレベル間の分散における相違を処理します。 詳しくは、 33 ページの『度数フィールドと重みフィールドの使用』を参照してください。

Random Trees ノードの作成オプション

「作成オプション」タブで、モデルを構築するすべてのオプションを設定します。もちろん、「実行」 ボタンをクリックすると、すべてデフォルト・オプションのモデルが構築されますが、通常は、それぞれの目的で作成をカスタマイズする必要があります。

このタブに含まれる数種類のペインを使用して、モデルに固有のカスタマイズを設定します。

Random Trees ノード - 基本

デシジョン ツリーを作成する方法に関する基本オプションを指定します。

構築するモデルの数: ノードが構築できるツリーの最大数を指定します。

サンプル サイズ**:** デフォルトでは、ブートストラップ サンプルのサイズは元の学習データと同じになります。大きなデータセットを扱う場合は、サンプル サイズを縮小することでパフォーマンスを高めることができます。0 から 1 の比率です。例えば、サンプル サイズを 0.6 に設定すると、サイズが元の学習データ サイズの 60% に削減されます。

バランス調整をしていないデータを処理する: モデルの対象がフラグの結果である場合 (例えば、購入するかしないかのどちらか) に、望ましくない結果に対する望ましい結果の比率が非常に小さいと、データが不均衡になり、モデルによって行われるブートストラップのサンプリングがモデルの精度に影響する可能性があります。精度を改善するには、このチェック ボックスを選択します。モデルで収集される望ましい結果の比率が上がり、より適切なモデルが生成されます。

変数選択に重み付きサンプルを使用する: デフォルトでは、各葉ノードの変数が同じ確率でランダムに選択されます。変数に重みを付けて選択プロセスを改善するには、このチェック ボックスを選択します。重みは、Random Trees ノード自体によって計算されます。重要度の高い (重みの大きい) フィールドは、予測変数として選択される可能性が高くなります。

ノードの最大数: 個々のツリーで許容される葉ノードの最大数を指定します。次の分割でこの数を超える場合は、ツリーの成長が停止して分割が行われません。

最大ツリー深度: ルート ノードの下の葉ノード の最大レベル数、つまりサンプルを (再帰的に) 分割できる最大回数を指定します。

子ノードの最小サイズ: 親ノードが分割された後の子ノードに最低限含まれていなければならないレコード数を指定します。子ノードに含まれるレコード数が指定した数より少なくなる場合は、親ノードは分割されません。

分割に使用する予測値の数を指定する: 分割モデルを構築する場合、各分割の構築に使用する予測値の最小個数を設定します。これにより、分割によって極端に小さいサブグループが作成されるのを防ぎます。このオプションを選択しなかった場合のデフォルト値は、■*sqrt*(M)■ (分類の場合) および ■M/3■ (回帰の場合) です (M は、予測変数の総数です)。このオプションを選択した場合、指定した数の予測値が使用されます。

注**:** 分割用の予測値の数をデータ内の予測値の総数より多くすることはできません。

精度の向上が見込めない場合は構築を中止する: Random Trees は、学習を停止するタイミングを決定するために特定の手順を使用します。具体的には、現在のアンサンブルの精度の向上が指定しきい値より小さい場合、新規ツリーの追加が停止されます。このため、「構築するモデルの数」オプションに指定した値よりもツリーが少ないモデルになることがあります。

Random Trees ノード - コスト

状況によっては、特定の誤りコストが他の誤りコストに比べて高いことがあります。例えば、信用リスクの高い申請者を低リスクに分類した場合 (ある種の誤分類) のコストは、リスクの低い申請者を高リスクに分類した場合 (別種の誤分類) よりも高くなります。誤分類コストでは、さまざまな予測の誤りに対し、相対的な重要度を指定できます。

誤分類コストは、基本的には、特定の結果に対して適用される重みです。これらの重みは、モデルに組み込まれ、(コストの高い誤りを防ぐための手段として) 実際に予測値に影響する場合があります。

コストを含むモデルは、コストを含まないモデルに比べてエラーが多く、全体の精度に関して低くランク付けされる可能性があります。ただし、コストが少ないエラーを優先するという組み込まれたバイアスがあるため、実際面ではパフォーマンスが優れている場合があります。

コスト行列には、可能な各予測カテゴリーや実際のカテゴリーの組み合わせのコストが表示されます。デフォルトでは、すべての誤分類コストが 1.0 に設定されています。コストの値を自分で入力するには、「誤分類コストを使用」 を選択して、コスト行列に独自の値を入力します。

誤分類コストを変更するには、目的の予測値と実際の値の組み合わせに対応するセルを選択して、セルの内容を削除してから、適切なコストを入力してください。コストは自動的に対称的にはなりません。例えば *A* を *B* として誤分類した場合のコストを 2.0 に設定しても、*B* を *A* として誤分類した場合のコストは、変更しない限りデフォルト値 (1.0) のまま変わりません。

順序型対象の場合にのみ、「順序型対象のデフォルトのコスト増加」を選択し、コスト行列内のデフォルト値を設定することができます。使用可能なオプションを次のリストに示します。

- 増加なし - 正しくないすべての予測にデフォルト値 1.0 を使用します。
- 線型 - 後続の不正確な予測が検出されるごとに、コストは 1 ずつ増加します。
- 平方 **(Square)** - 後続の不正確な予測が検出されるごとに、コストは線型値の平方になります。この場合、値は 1、4、9 などになります。
- カスタム - テーブル内のいずれかの値を手動で編集すると、ドロップダウン・オプションは、「カスタム」に自動的に変更されます。ドロップダウン選択を他のいずれかのオプションに変更すると、編集された値は、選択されたオプションの値と置き換えられます。

Random Trees ノード - 詳細設定

デシジョン ツリーを作成する方法に関する拡張オプションを指定します。

欠損値の最大パーセンテージ**:** 入力で許容される欠損値の最大パーセンテージを指定します。パーセンテージがこの数値を超えた場合は、その入力がモデルの構築から除外されます。

単一カテゴリの多数派が割合を超えるフィールドを除外する**:** 1 つのカテゴリがフィールド内に持つことのできるレコードの割合の最大値を指定します。この値を超えるレコードの割合を示しているカテゴリ値がある場合は、フィールド全体がモデルの構築から除外されます。

フィールド カテゴリの最大数**:** フィールドに入れることのできるカテゴリの最大数を指定します。カテゴリの数がこの数値を超えた場合は、そのフィールドがモデルの構築から除外されます。

最小値フィールド変動**:** 連続型フィールドの変動係数がここで指定した値より小さい (つまり、フィールドがほぼ一定である) 場合は、そのフィールドがモデルの構築から除外されます。

ビン数**:** 連続型入力に使用される等度数ビンの数を指定します。選択可能なオプションは、2、4、5、10、20、25、50、または 100 です。

報告する関心の高いルールの数 **(Number of interesting rules to report):** 報告するルールの数を指定します (最小値は 1、最大値は 1000、デフォルト値は 50 です)。

Random Trees ノードのモデル・オプション

「モデル・オプション」タブで、モデルの名前を指定するか、自動的に名前を生成するかを選択できます。モデルのスコアリング時に予測変数の重要度を計算するように指定することもできます。

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

Random Trees モデル ナゲット

Random Trees モデル ナゲットの出力

Random Trees モデルを作成すると、以下の情報が出力ビューアーに表示されます。

「モデル情報」テーブル

「モデル情報」テーブルは、モデルについての重要な情報を提供します。このテーブルには、常に以下の上位モデル設定が含まれています。

- データ型ノードまたは Random Trees ノードの「フィールド」タブで選択された対象フィールドの名前。
- モデル作成方法 - Random Trees。
- モデルへ入力される予測値の数。

テーブルに表示されるその他の詳細は、分類モデルと回帰モデルのいずれを作成するか、およびモデルが不均衡なデータを処理するように作成されるかによって異なります。

- 分類モデル (デフォルト設定)
  - モデル精度
  - 誤分類ルール
- 分類モデル (「バランス調整をしていないデータを処理する」を選択)
  - 全平均
  - 真陽性率 (クラスに分割されます)。
- 回帰モデル
  - 平均平方根誤差
  - 相対誤差
  - 説明された分散

レコード要約

要約には、モデルを適合させるために使用されたレコード数と除外されたレコード数が表示されます。レコードの数と全体の数のパーセンテージの両方が表示されます。度数の重みを含めるようにモデルが作成され

ている場合、重み付きのないレコードのうち含められた数と除外された数も表示されます。

予測変数の重要度

「予測変数の重要度」グラフは、モデル内の上位 10 個の入力 (予測値) の重要度を棒グラフとして表示します。

グラフ内に 10 個を超えるフィールドがある場合は、グラフの下のスライダーを使用して、グラフ内に含まれる予測値の選択を変更できます。スライダー上のインディケーター・マークは固定幅であり、スライダー上の各マークは 10 個のフィールドを表します。スライダーに沿ってインディケーター・マークを移動して、予測変数の重要度の順序で並べられた次の 10 個または前の 10 個のフィールドを表示できます。

グラフをダブルクリックして、グラフ・サイズを編集できる別個のダイアログ・ボックスを開くことができます。この別個の編集ダイアログ・ボックスを閉じると、「出力」タブに表示されるグラフに変更が適用されます。

「上位ディシジョン・ルール (Top Decision Rules)」テーブル

デフォルトでは、この対話式のテーブルには、上位ルールの統計が関心度でソートされて表示されます。

テーブルをダブルクリックして、テーブル内に表示されるルール情報を編集するための別個のダイアログ・ボックスを開くことができます。ダイアログ・ボックス内に表示される情報および使用可能なオプションは、対象のデータ型 (カテゴリー型または連続型など) によって異なります。

テーブルには次のルール情報が表示されます。

- ルールの適用方法および構成内容の詳細
- 結果が最も頻繁に出現するカテゴリに含まれるかどうか
- ルールの精度
- ツリーの精度
- 関心度インデックス

関心度インデックスは、以下の式を使用して計算されます。

$$I_{index}(t) = P(A(t)) * P(B(t)) * (P(B(t)|A(t)) + P(\bar{B}(t)|\bar{A}(t)))$$

この式で、各項目は次のとおりです。

- P(A(t)) はツリーの精度
- P(B(t)) はルールの精度
- P(B(t)|A(t)) はツリーとノードの両方による正しい予測を表す
- 式の残りの部分は、ツリーとノードの両方による誤った予測を表しています。

以下の「テーブル コンテンツ **(Table contents)**」オプションを使用して、ルール テーブル レイアウトを変更できます。

- 最上位の決定ルール**:** 関心度インデックスでソートされた上位 5 個の決定ルール。
- すべてのルール **(All rules):** このテーブルには、モデルにより生成されたすべてのルールが含まれていますが、1 ページあたり 20 個のルールのみが表示されます。このレイアウトを選択すると、「**ID** でルールを検索 **(Find rule by ID)**」および「ページ」の追加オプションを使用して、ルールを検索できます。

さらに、カテゴリー型対象の場合は、「カテゴリーごとの上位ルール **(Top rules by category)**」オプションを使用して、ルール テーブル レイアウトを変更できます。選択した「対象カテゴリー」の合計レコードの割合によって、上位 5 個のディシジョン・ルールがソートされます。

注**:** カテゴリー型対象の場合、このテーブルは「作成オプション」の「基本」タブで「バランス調整をしていないデータを処理する」を選択していない場合にのみ使用できます。

ルール テーブルのレイアウトを変更する場合は、ダイアログ ボックスの左上にある「ビューアーにコピー」ボタンをクリックして、変更されたルール テーブルを元の出力ビューアーにコピーできます。

混同マトリックス

分類モデルの場合、混同マトリックスに予測結果の数と実際の観測結果の数の比較、および正しい予測の割合が表示されます。

注**:** 混同マトリックスは回帰モデルでは使用できず、「作成オプション」の「基本」タブで「バランス調整をしていないデータを処理する」を選択している場合にも使用できません。

Random Trees モデル ナゲットの設定

Random Trees モデル ナゲットの「設定」タブでは、モデル スコアリング時の確信度のオプションおよびSQL 生成のオプションを指定します。このタブは、モデル・ナゲットがストリームに追加された後にのみ使用できます。

確信度の計算**:** スコアリング操作に確信度を含めるには、このチェック・ボックスを選択します。データベースでモデルをスコアリングするときに、確信度を除外すると、より効率的な SQL を生成できます。回帰ツリーでは確信度は割り当てられないことに注意してください。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL の生成方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

C&R Tree、CHAID、QUEST、および C5.0 ディシジョン・ツリー・モデル・ナゲット

ディシジョン・ツリー・モデル・ナゲットは、ディシジョン・ツリー・モデル作成ノード (C&R Tree、CHAID、QUEST、または C5.0) のいずれかによって発見された特定の出力フィールドを予測するためのツリー構造を表します。ツリー・モデルはツリー構築ノードまたはインタラクティブ・ツリー・ビルダーから間接的に生成できます。 詳しくは、トピック 87 ページの『インタラクティブ・ツリー・ビルダー』を参照してください。

スコアリング・ツリー・モデル

ツリー・モデル・ナゲットを持つストリームを実行する場合、その結果は具体的にはツリーの種類によって異なります。

- 分類ツリー（カテゴリー対象）で、レコードごとに予測フィールドと確信度を含む 2 つの新しいフィールドがデータに追加されます。予測は、レコードが割り当てられたターミナル・ノードの最も頻度の高いカテゴリーによって決まります。つまり、あるノードで過半数の回答者が *yes* と答えた場合、そのノードに割り当てられたレコードの予測はすべて「yes」です。
- 回帰ツリーでは、予測値のみが生成され、確信度は割り当てられません。
- オプションとして、CHAID、QUEST、および C&R Tree のモデルに、もう 1 つのフィールドを追加することができますが、これは各レコードを割り当てるノードに ID を示すためのものです。

新規フィールド名はモデル名から派生し、接頭辞が付けられます。C&R Tree、CHAID、および QUEST の接頭辞は、予測フィールドに *$R-* 、確信度フィールドに *$RC-*、また、識別子フィールドに *$RI-* です。C5.0 ツリーの場合、予測フィールドの接頭辞は *$C-* で、確信度フィールドの接頭辞は *$CC-* です。複数のツリー・モデル・ノードが存在する場合、新しいフィールド名には、必要に応じて、接頭辞 にノード識別用の数字が含まれます。例えば、*$R1-*、*$RC1-*、*$R2-* などです。

ツリー・モデル・ナゲットの処理

モデルに関する情報を、さまざまな方法で保存、またはエクスポートできます。

注**:** これらのオプションの多くは、ツリー・ビルダー・ウィンドウからも利用できます。

ツリー・ビルダーと生成されたモデルのどちらからでも、次を実行できます。

- 現在のツリーに基づいて、フィルター・ノードまたは条件抽出ノードを生成する。詳しくは、 98 ページの『フィルター・ノードおよび条件抽出ノードの生成』を参照してください。
- ツリー構造を、ツリーのターミナル・ブランチを定義するルールのセットとして表す、新しいルール・セット・ノードを作成する。詳しくは、 98 ページの『ディシジョン・ツリーからのルールセットの生成』を参照してください。
- さらに、ツリー・モデル・ナゲットについてのみ、モデルを PMML 形式でエクスポートできます。詳しくは、 41 ページの『モデル・パレット』を参照してください。モデルがユーザー定義の分割を含んでいる場合、その情報は、エクスポートされた PMML には保存されません。(分割は保存されますが、アルゴリズムによる選択ではなく、ユーザー定義であるという情報は保存されません。)
- 現在のツリーの選択した部分に基づいてグラフを生成する。なお、ストリーム内のその他のノードに接続している場合のナゲットにのみ生成できます。詳しくは、 130 ページの『グラフの生成』を参照してください。
- ブーストされた C5.0 モデルについてのみ、選択中のルールから新しいルール・セットを作成するために 「シングル ディシジョン ツリー **(**キャンバス**)**」 または「シングル ディシジョン ツリー **(GM** パレット**)**」 を選択できます。 詳しくは、トピック 130 ページの『ブーストされた C5.0 モデル』を参照してください。

注**:** ルール・ノードは、C&R Tree ノードに置き換えられていますが、元々ルール・ノードを使用して作成された既存のストリーム内のディシジョン・ツリー・ノードは依然正しく機能することに注意してください。

単一ツリー・モデル・ナゲット

「単一ツリーの構築」 をモデル作成ノードの主な目的に選択すると、作成されるモデル・ナゲットには次のタブが含まれます。

表 7. 単一ツリー・ナゲットのタブ

タブ	説明	詳細情報
モデル	モデルを定義するルールが表示されます。	詳しくは、トピック『ディシジョン・ツリー・モデル ルール』を参照してください。
ビューアー	モデルのツリー・ビューが表示されます。	詳しくは、トピック 128 ページの『ディシジョン・ツリー・モデル・ビューアー』を参照してください。
要約	フィールド、作成設定、およびモデル推定プロセスについての情報が表示されます。	詳しくは、トピック 44 ページの『モデル・ナゲットの要約/情報』を参照してください。
設定	モデル・スコアリング時の確信度および SQL 生成のオプションを指定できます。	詳しくは、トピック 128 ページの『ディシジョン・ツリー/ルール・セット・モデル・ナゲットの設定』を参照してください。
注釈	説明の注釈を追加し、カスタム名を指定、ツールヒントを追加し、モデルの検索キーワードを指定できます。	

ディシジョン・ツリー・モデル ルール

ディシジョン・ツリー・ナゲットの「モデル」タブには、モデルを定義するルールが表示されます。オプションで、予測変数の重要度のグラフおよび時系列、度数、代理変数に関する情報を含む 3 番目のパネルを表示することができます。

注: CHAID ノードの「作成オプション」タブ (「目的」パネル) で「非常に大きいデータセットのモデルを作成」オプションを選択すると、「モデル」タブにはツリー・ルールの詳細のみが表示されます。

ツリーのルール

左側の領域には、アルゴリズムが発見したデータの分岐を定義する条件が表示されています。これは、基本的には異なる予測フィールドの値に基づいて、子ノードに個別のレコードを割り当てるために使用できる一連のルールです。

ディシジョン・ツリーは、入力フィールド値に基づいて回帰的にデータを分岐させることによって機能します。データの分岐をブランチといいます。初期のブランチ (ルート) には、すべてのデータ・レコードが含まれます。ルートは、特定の入力フィールド値を基準にして、サブセットまたは子ブランチに分割されます。各子ブランチはさらに分割でき、それをさらに分割していくことができます。ツリーの最下位レベルは、それ以上分割されないブランチです。そのようなブランチを、ターミナル・ブランチまたは葉と呼びます。

ツリーのルールの詳細

ツリー ブラウザーには、各分岐 (ブランチ) を定義する入力値と、その分割内のレコードの出力フィールド値の要約が表示されます。モデル・ブラウザー使用法に関する一般情報については、 43 ページの『モデル・ナゲットの参照』を参照してください。

数値型フィールドに基づく分割の場合、ブランチは次のような 1 行の書式で表示されます。

```
fieldname relation value [summary]
```

ここで *relation* は数値の関係を表します。例えば、「*revenue*」 フィールドの 100 より大きい値によって定義されるブランチは、次のように表示されます。

```
revenue > 100 [summary]
```

シンボル値フィールドに基づく分割の場合、ブランチは次のような 1 行の書式で表示されます。

```
fieldname = value [summary] or fieldname in [values] [summary]
```

この *values* はブランチを定義するフィールド値を表しています。例えば、*region* の値が *North*、*West*、または *South* のいずれかになるレコードを含むブランチは、次のように表されます。

```
region in ["North" "West" "South"] [summary]
```

ターミナル・ブランチの場合は、ルール条件の最後に矢印と予測値を追加すると、予測も提供されます。例えば、*revenue* > 100 によって定義され、出力フィールドに対して *high* の値を予測するリーフは、次のように表示されます。

```
revenue > 100 [Mode: high] → high
```

ブランチの要約は、シンボル値出力フィールドと数値出力フィールドでは異なる方法で定義されます。数値出力フィールドを含むツリーの場合、要約はそのブランチの平均値であり、そのブランチの効果は、そのブランチの平均とその親ブランチの平均の差として定義されます。シンボル値出力フィールドを含むツリーの場合、要約はそのブランチ内にあるレコードのモード (最頻値) になります。

ブランチを完全に説明するには、そのブランチを定義する条件に加えて、ツリーの上位レベルの分割を定義する条件を含める必要があります。例えば、次のようなツリーがあるとします。

```
revenue > 100
  region = "North"
  region in ["South" "East" "West"]
    revenue <= 200
```

この場合、2 行目に記載されているブランチは、条件 *revenue > 100* および *region = "North"* で定義されています。

ツールバーの 「インスタンス/確信度の表示」 をクリックすると、各ルールによってさらに、ルールが適用されるレコード数 (インスタンス)、およびルールが真 (true) であるケースの比率 (確信度) の情報も表示されます。

予測変数の重要度

オプションで、モデルの推定時に各予測値の相対的重要度を示すグラフを「モデル」タブに表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。

注: このグラフは、モデル生成前に「分析」タブで「予測変数の重要度の計算」が選択されている場合にのみ使用できます。詳しくは、トピック 44 ページの『予測変数の重要度』を参照してください。

モデルの付加情報

ツールバーの 「付加情報パネルを表示」 をクリックすると、選択したルールの詳細情報を記載したパネルが、ウィンドウの下部に表示されます。情報パネルには、3 種類のタブがあります。

時系列 : ルート・ノードから選択されたノードまでの分岐条件をトレースしています。ここには、選択されているノードにレコードが割り当てられる時期を決める条件が記載されています。すべての条件が真のノードは、このノードに割り当てられます。

度数分布表 : 対象フィールドがシンボル値のモデルの場合、有効な各対象値に対して、その対象値を持つこのノード (学習データ中) に割り当てられているレコード数を表示します。パーセントで表された度数の図も表示されます (最大で小数点以下 3 桁までを表示)。数値の対象値を持つノードの場合、このタブには何も表示されません。

代理変数: 適用可能な場合、選択されているノードで、主分岐フィールド用の代理変数が表示されます。代理変数は、あるレコードで主予測値が欠損値の場合に、代わりに使用されるフィールドです。ツリー構築ノードでは、ある分割で使用できる代理変数の個数の最大値を指定します。ただし、実際の個数は、学習用データに依存します。一般に、欠損値データが多いほど、使用される代理変数も多くなります。他のディシジョン・ツリー・モデルの場合、このタブには何も表示されません。

注: 代理変数をモデルに含めるには、代理変数を学習フェーズ中に識別する必要があります。学習用サンプルに欠損値がない場合、代理変数は識別されません。また、テストまたはスコアリング中に出現した、欠損値を持つレコードは、自動的に最大のレコード数を持つ子ノードに分類されます。テストまたはスコアリング中に欠損値が予測される場合は、その値が学習用サンプルでも欠損値であることを確認してください。代理変数は、CHAID ツリーでは使用できません。

効果

ノードの効果は、平均値の増大または減少です (親ノードと比較した予測値)。例えば、ノードの平均値が 0.2 で、その親の平均値が 0.6 の場合、ノードの効果は 0.2-0.6=-0.4 です。この統計量は、連続型対象の場合にのみ適用されます。

ディシジョン・ツリー・モデル・ビューアー

生成されたディシジョン・ツリー・モデルの「ビューアー」タブは、ツリー・ビルダーでの表示に似ています。主な違いは、モデル・ナゲットを参照する場合、ツリーを大きくしたり修正したりできないことです。表示および表示をカスタマイズするためのその他のオプションは、2 つのコンポーネント間で似ています。詳しくは、トピック 90 ページの『ツリー・ビューのカスタマイズ』を参照してください。

注: 「ビューアー」タブは、「目的」パネルの「作成オプション」タブで「非常に大きいデータセットのモデルを作成」オプションを選択した場合に作成される CHAID モデル・ナゲットには表示されません。

「分割」タブに分割ルールが表示されると、大かっこは隣接する値が範囲内含まれ、かっこは隣接する値が範囲から除外されていることを示します。式 (23,37] は 23 を除き 37 を含む 23 ～ 37 の範囲を示します。「モデル」タブで同じ条件は次のように示されます。

```
Age > 23 and Age <= 37
```

ディシジョン・ツリー/ルール・セット・モデル・ナゲットの設定

ディシジョン・ツリーまたはルール・セット・モデル・ナゲットの「設定」タブで、確信度のオプションとモデル・スコアリング中の SQL 生成 を指定することができます。このタブは、モデル・ナゲットがストリームに追加された後にのみ使用されます。

確信度の計算: スコアリング操作に確信度を含める場合に選択します。データベースでモデルをスコアリングする場合、確信度を除外することで、より効率的な SQL を生成することができます。回帰ツリーでは確信度は割り当てられないことに注意してください。

注: CHAID モデルの「作成オプション」タブ (「方法」パネル) で「非常に大きいデータセットのモデルを作成」オプションを選択すると、このチェック・ボックスは、名義型またはフラグ型のカテゴリー対象のモデル・ナゲットでのみ使用できます。

未調整傾向スコアの計算: 対象がフラグ型 (yes または no の予測を返す) であるモデルの場合は、対象フィールドに true の結果が指定される尤度を示す傾向スコアを要求できます。また、スコアリング時に生成することができるその他の予測および確信度値があります。

注: CHAID モデルの「作成オプション」タブ (「方法」パネル) で「非常に大きいデータセットのモデルを作成」オプションを選択すると、このチェック・ボックスは、フラグ型のカテゴリー対象のモデル・ナゲットでのみ使用できます。

調整済み傾向スコアの計算: 未調整傾向スコアは、学習データのみに基づくものであり、多くのモデルがこのデータにオーバーフィットする傾向があるため、楽観的になり過ぎる場合があります。調整済み傾向は、テストまたは検証用データ区分に対してモデルのパフォーマンスを評価することによって補正しようとします。このオプションでは、モデルの生成前にデータ区分フィールドをストリーム内で定義し、調整済み傾向スコアがモデル作成ノードで有効化されている必要があります。

注: 調整済み傾向スコアは、ブースティング・ツリーおよびルール・セット・モデルには使用できません。詳しくは、トピック 130 ページの『ブーストされた C5.0 モデル』を参照してください。

ルール識別子: CHAID、QUEST、および C&R Tree モデルでは、このオプションによって、スコアリング出力にフィールドが追加されます。このフィールドは、各レコードを割り当てるターミナル・ノードの ID を示すものです。

注: このオプションを選択した場合、SQL 生成は使用できません。

このモデルの **SQL** を生成: データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト: **Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- 欠損値のサポートのないネイティブ **SQL** への変換によるスコア: これを選択すると、欠損値処理によるオーバーヘッドを発生させることなく、データベース内でモデルをスコアリングするネイティブ SQL が生成されます。このオプションでは、ケースのスコアリング時に欠損値が見つかると、予測にヌル (`$null$`) を設定します。

  注: このオプションは、CHAID モデルでは使用できません。他のモデルの場合、ディシジョン・ツリーの場合にのみ使用できます (ルール・セットでは使用できません)。
- 欠損値のサポートのあるネイティブ **SQL** への変換によるスコア: CHAID、QUEST、および C&R Tree モデルで、欠損値を全面的にサポートしながら、データベース内でモデルをスコアリングするネイティブ SQL を生成できます。この場合、モデル中で指定されているように欠損値が処理されて、SQL が生成されます。例えば、C&R Tree は代理変数ルールと Biggest Child Fallback (あるレコードの分割フィールド、およびその分割に対するすべての代理変数フィールドに欠損値がある場合、そのレコードは

重み付けされた最大の度数を持つ子ノードに割り当てられる (ケースまたは度数の重みが使用中の場合を除き、通常は大半のレコードが割り当てられた子ノードになる) を使用します。

注: C5.0 モデルの場合、このオプションはルール・セットの場合にのみ使用できます (ディシジョン・ツリーでは使用できません)。

- データベースの外部でスコアリング: このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

ブーストされた C5.0 モデル

この機能は SPSS Modeler Professional および SPSS Modeler Premium で使用できます。

ブーストされた C5.0 モデル (ルール・セットまたはディシジョン・ツリー) を作成する場合は、実際には関連する複数のモデルのセットを作成します。ブーストされた C5.0 モデル用のモデル ルール・ブラウザーでは、階層の最上位レベルのモデルのリストと、各モデルの推定精度、ブーストされたモデルの全体の精度が表示されます。特定のモデルに関するルールまたは分割を調べるには、単一モデル内のルールやブランチと同じように、そのモデルを選択して展開します。

また、ブーストされたモデルのセットから特定のモデルを抽出し、そのモデルだけを含む新しいルール・セット・モデル・ナゲットを生成できます。ブーストされた C5.0 モデルから新規ルール・セットを作成するには、対象のルール・セットまたはツリーを選択し、「ノードの生成」メニューから「シングル ディシジョン ツリー (GM パレット)」または「シングル ディシジョン ツリー (キャンバス)」を選択します。

グラフの生成

ツリー・ノードは多くの情報を提供します。ただし、その情報はビジネス・ユーザーが容易にアクセスできる形式であるとは限りません。ビジネス・レポート、プレゼンテーションなどに用意に組み込むことができる方法でデータを提供するために、選択したデータのグラフを作成できます。例えば、モデル・ナゲットの「モデル」タブまたは「ビューアー」タブから、またはインタラクティブ・ツリーの「ビューアー」タブから、ツリーの選択した部分のグラフを生成でき、そのため選択したツリーまたはブランチノードのケースのグラフのみを生成できます。

注 : ストリームのそのほかのノードに接続している場合のみ、ナゲットからグラフを生成できます。

グラフの生成

まず、次のように、グラフに表示する情報を選択します。

- ナゲットの「モデル」タブで、左側のウィンドウ枠の条件とルールのリストを展開し、関心のあるリストを条件またはルールを選択します。
- ナゲットの「ビューアー」タブで、ブランチのリストを展開して関心のあるブランチを選択します。
- インタラクティブ・ツリーの「ビューアー」タブで、ブランチのリストを展開して関心のあるブランチを選択します。

注 : 「ビューアー」タブの最上位ノードは選択できません。

表示するデータを選択する方法に関係なく、グラフを作成する方法は同じです。

1. 「生成」メニューの 「グラフ (選択項目から)」 を選択します。また、「ビューアー」タブの左下の「グラフ (選択項目から) ボタンを選択します。グラフボードの「基本」タブが表示されます。

   注 : この方法でグラフボードを表示した場合、「基本」タブと「詳細」タブのみを使用できます。
2. 「基本」タブまたは「詳細」タブいずれかの設定を使用し、グラフに表示する詳細を指定します。

3. 「OK」をクリックしてグラフを生成します。

グラフの見出しは選択されたノードまたはルールを識別します。

ブースティング、バギング、非常に大きいデータセットのモデル・ナゲット

「モデル精度の強化 **(**ブースティング**)**」、「モデルの安定性を拡張 **(**バグ**)**」、または 「非常に大きなデータセットのモデルを作成」 をモデル作成ノードの主な目的に選択する場合、IBM SPSS Modeler は複数モデルのアンサンブルを作成します。 詳しくは、トピック 46 ページの『アンサンブルのモデル』を参照してください。

生成されるモデル・ナゲットには次のタブが含まれます。「モデル」タブには、さまざまなモデルのビューが表示されます。

表 *8.* モデル・ナゲットで使用可能なタブ

タブ	ビュー	説明	詳細情報
モデル	モデル要約	アンサンブルの品質 (ブースティング・モデルおよび連続型対象を除く) および多様性の概要、異なるモデルで予測値がどのように異なるかについての測定が表示されます。	詳しくは、トピック 47 ページの『モデルの要約』を参照してください。
	予測変数の重要度	モデルを推定する際に各予測値 (入力フィールド) の相対重要度を示すグラフが表示されます。	詳しくは、トピック 47 ページの『予測変数の重要度』を参照してください。
	予測度数	各予測値がモデルのセットに使用する相対度数を示すグラフが表示されます。	詳しくは、トピック 47 ページの『予測値の頻度』を参照してください。
	コンポーネント モデルの精度	アンサンブル内のさまざまなモデルの予測制度に関するグラフを表示します。	
	コンポーネント モデルの詳細	アンサンブル内のさまざまなモデルの詳細が表示されます。	詳しくは、トピック 48 ページの『コンポーネント・モデルの詳細』を参照してください。
	情報	フィールド、作成設定、およびモデル推定プロセスについての情報が表示されます。	詳しくは、トピック 44 ページの『モデル・ナゲットの要約/情報』を参照してください。
設定		スコアリング操作に確信度を含めることができます。	詳しくは、トピック 128 ページの『ディシジョン・ツリー/ルール・セット・モデル・ナゲットの設定』を参照してください。
注釈		説明の注釈を追加し、カスタム名を指定、ツールヒントを追加し、モデルの検索キーワードを指定できます。	

C&R Tree、CHAID、QUEST、C5.0、および Apriori ルール・セットのモデル・ナゲット

ルール・セットモデル・ナゲットは、アソシエーション・ルール・モデル作成ノード (Apriori) によって、またはツリー作成ノード (C&R Tree、CHAID、QUEST、または C5.0) のいずれかによって検出された特定の出力フィールドを予測するルールを示します。アソシエーション・ルールの場合、ルールセットは未調整ルール ナゲットから生成する必要があります。ツリーの場合、ルール・セットは、インタラクティブ・ツリー・ビルダー、C5.0 モデル構築ノード、または任意のツリー・モデル・ナゲットから生成できます。未調整ルール ナゲットとは異なり、ルール・セット ナゲットはストリーム内に置いて予測を生成できます。

ルール・セット ナゲットを含むストリームを実行すると、データに対して各レコードごとに予測された値と確信度を含むストリームに、2 つの新規フィールドが追加されます。新規フィールド名はモデル名から派生し、接頭辞が付けられます。アソシエーション・ルール・セットの場合、予測フィールドの接頭辞は *$A-* で、確信度フィールドの接頭辞は *$AC-* です。C5.0 ルール・セットの場合、予測フィールドの接頭辞は *$C-* で、確信度フィールドの接頭辞は *$CC-* です。C&R Tree ルール・セットの場合、予測フィールドの接頭辞は *$R-* で、確信度フィールドの接頭辞は *$RC-* です。同じ出力ノードを連続して予測する複数のルール・セット・ナゲットを含むストリーム内では、新規フィールド名を区別するためにそれぞれの接頭辞に番号が追加されます。ストリーム内の最初のアソシエーション・ルール・セット・ナゲットでは通常の名前を使用します。2 番目のノードでは *$A1-* と *$AC1-* で始まる名前、3 番目のノードでは *$A2-* と *$AC2-* で始まる名前というように名前が付けられていきます。

ルールの適用方法 **:** アソシエーション・ルールから生成されたルール・セットは、他のモデル・ナゲットとは異なります。アソシエーション ルールから生成されたルール・セットが他のモデル・ナゲットと異なる理由は、特定のレコードについて複数の予測が生成される場合があり、それらの予測がすべて一致するとは限らないためです。ルール・セットから予測を生成するには、次の 2 つの方法があります

注**:** どちらの方法を採るかにかかわらず、ディシジョン・ツリーから生成されたルール・セットは同じ結果を返します。これは、1 つのディシジョン・ツリーから得られる複数のルールは相互排他的であるためです。

- 票決**:** この方法では、レコードに適用されるすべてのルールの予測の結合を試行します。各レコードのすべてのルールを調べ、レコードに適用される各ルールを使用して予測および関連付けられた確信度を生成します。各出力値の確信度値の合計を計算し、最も大きい確信度合計を持つ値を最終的な予測として選択します。最終的な予測の確信度は、その値の確信度合計をそのレコードに該当するルールの数で割ったものになります。
- 最初のヒット**:** この方法では、単純にルールを順番に検定し、レコードに最初に適用されるルールを使用して予測を生成します。

使用する方法は、「ストリーム・オプション」で制御できます。

ノードの生成**:** 「生成」メニューを使用し、ルール・セットに基づいて新しいノードを作成することができます。

- フィルター ノード: ルール・セット内のルールで使用されないフィールドにフィルターをかけるための新規フィルター・ノードを生成します。
- 条件抽出ノード: 選択したルールを適用するレコードを選択するための新規条件抽出ノードを生成します。生成されたノードは、ルールのすべての先行条件が真 (true) であるレコードを選択します。このオプションではルールを選択する必要があります。

- ルール・トレース・ノード: 各レコードの予測の作成に使用されたルールを示すフィールドを算出する、新規スーパーノードを作成します。ルール・セットが最初のヒット方法で評価される場合、これは該当する最初のルールを示す単なる記号になります。ルール・セットが票決方法で評価される場合、これは票決メカニズムへの入力を示すより複雑な文字列になります。
- シングル ディシジョン ツリー **(**キャンバス**)**/シングル ディシジョン ツリー **(GM** パレット**)**: 現在選択されているルールから派生する単一の新規ルール・セット・ナゲットを作成します。ブーストされた **C5.0** モデルの場合にのみ使用できます。 詳しくは、トピック 130 ページの『ブーストされた C5.0 モデル』を参照してください。
- モデルをパレットに: モデルをモデル・パレットに戻します。これは、同僚から、モデル自体ではなくモデルを含むストリームが送信されてきた場合に役立ちます。

注**:** ルール・セット・ナゲットの「設定」タブおよび「要約」タブは、ディシジョン・ツリー・モデルで使用されているものと同じです。

ルール・セットの「モデル」タブ

ルール・セット ナゲットの「モデル」タブで、アルゴリズムによってデータから抽出されたルールのリストが表示されます。

ルールは、結果 (予測されるカテゴリー) ごとに分類され、次の形式で表示されます。

```
if antecedent_1
and antecedent_2
...
and antecedent_n
then predicted value
```

consequent と antecedent_1 から antecedent_n まではすべて条件です。ルールは、「antecedent_1 から antecedent_n がすべて true であるレコードの場合、consequent も true である可能性が高い」ものとして解釈されます。ツールバーの 「インスタンス**/**確信度の表示」 ボタンをクリックすると、さらに各ルールが適用されるレコード数、つまり前提条件が真 (true) (インスタンス)、およびルール全体が真 (true) であるレコードの比率 (確信度) に関する情報も表示されます。

C5.0 ルール・セットの場合は確信度がやや異なる方法で計算されることに注意してください。C5.0 では、次の式を使用してルールの確信度を計算します。

```
(1 + number of records where rule is correct)
/
(2 + number of records for which the rule's antecedents are true)
```

この確信度推定の計算によって、ディシジョン・ツリーからルールを生成するプロセス (C5.0 がルール・セットを作成するときに実行する処理) が調整されます。

AnswerTree 3.0 からのプロジェクトのインポート

IBM SPSS Modeler は、次の手順で示すように、標準の「ファイルを開く」ダイアログ・ボックスを使用すると、AnswerTree 3.0 または 3.1 で保存したプロジェクトをインポートできます。

1. IBM SPSS Modeler のメニューから次の項目を選択します。

   「ファイル」 > 「ストリームを開く」
2. 「ファイルの種類」ドロップダウン・リストから、「**AT Project** ファイル **(*.atp, *.ats)**」を選択します。

インポートされた各プロジェクトは、次のノードを使用して IBM SPSS Modeler ストリームに変換されます。

- データ・ソースを定義する入力ノードが 1 個使用されます (例えば、IBM SPSS Statistics データ・ファイルまたはデータベース ソース)。
- プロジェクトにある各ツリー (複数ある場合があります) について、タイプ、役割 (入力つまり予測値フィールドまたは、出力各フィールド)、欠損値および他のオプションを含む、各フィールド (変数) のプロパティーを定義するデータ型ノードが 1 つ作成されます。
- プロジェクトにある各ツリー (複数ある場合があります) について、データを学習用とテスト用のサンプルに分割するデータ区分ノードが 1 つ作成され、さらに、ツリー構築ノードが 1 つ作成され、ツリーを生成するためのパラメーターを定義します (C&R Tree、QUEST、または CHAID ノードのいずれか)。

3. 生成されたツリーを表示するには、ストリームを実行します。

コメント

- IBM SPSS Modeler で生成されたディシジョン・ツリーは、AnswerTree にエクスポートできません。AnswerTree から IBM SPSS Modeler へのインポートは、一方通行です。
- AnswerTree で定義されたプロフィットは、プロジェクトが IBM SPSS Modeler にインポートされるときに保存されません。

第 7 章 Bayesian network (ベイズ) モデル

Bayesian network (ベイズ) ノード

Bayesian network (ベイズ**)** ノードを使用すると、観測された情報および記録された情報を「常識」という実際の知識を組み合わせることによって確率モデルを作成し、表面的にはリンクしていない属性を使用して発生の尤度を確立できます。ノードは主に分類に使用される Tree Augmented Naïve Bayes (TAN) および Markov Blanket ネットワークに焦点を当てています。

Bayesian network (ベイズ) は、あらゆる状況で予測を行うために使用されます。以下に例を示します。

- デフォルトのリスクが低い、ローンの機会を選ぶ。
- センサーの入力および既存のレコードに基づき、機器にサービス、部品、置換が必要な時期を推定する。
- オンラインのトラブルシューティング・ツールを使用して顧客の問題を解決する。
- 携帯電話ネットワークをリアルタイムで診断およびトラブルシューティングする。
- 発生しうるリスクおよび研究開発プロジェクトの報酬を評価し、リソースを最も良い機会に集める。

Bayesian network (ベイズ) は、データセットに変数 (多くの場合、ノードとして参照) を表示、および変数間の確率的または条件的独立性を表示するグラフィカルなモデルです。ノード間の因果関係は、Bayesian network (ベイズ) によって表されますが、(**arcs** と呼ばれる) ネットワークのリンクは直接的な原因と結果を必ずしも表すわけではありません。例えば、グラフに表示された症状と病気の間の確率的独立性が真である場合、Bayesian network (ベイズ) を使用して、特定の症状およびその他の関連データが存在または非存在を考慮し、Bayesian network (ベイズ) を使用して、患者が特定の病気を持つ確率を計算できます。情報がない場合、ネットワークは非常に強力で、存在するすべての情報を使用して、最善の予測を行います。

Bayesian network (ベイズ) の一般的で基本的な例は、Lauritzen および Spiegelhalter によって作成されていました (1988 年)。この例は、「アジア」モデルとして参照され、医師の新しい患者、因果関係にほとんど対応するリンクの方向を診断するために使用されるネットワークを単純化したものです。それぞれのノードは、患者の状況に関連するファセットを表します。例えば、「Smoking」は常習喫煙者を表し、「VisitAsia」は最近アジアをに行ったことを表します。確率の関係はノード間のリンクによって表されます。例えば、喫煙すると患者が気管支炎および肺ガンを患う可能性が上昇し、年齢は肺ガンを発症する可能性にのみ関連するように考えられます。同様に、肺の X 線での異常は結核または肺ガンによるものであることが考えられますが、気管支炎または肺ガンも患っている場合、患者が呼吸困難に陥っている可能性が大きくなります。

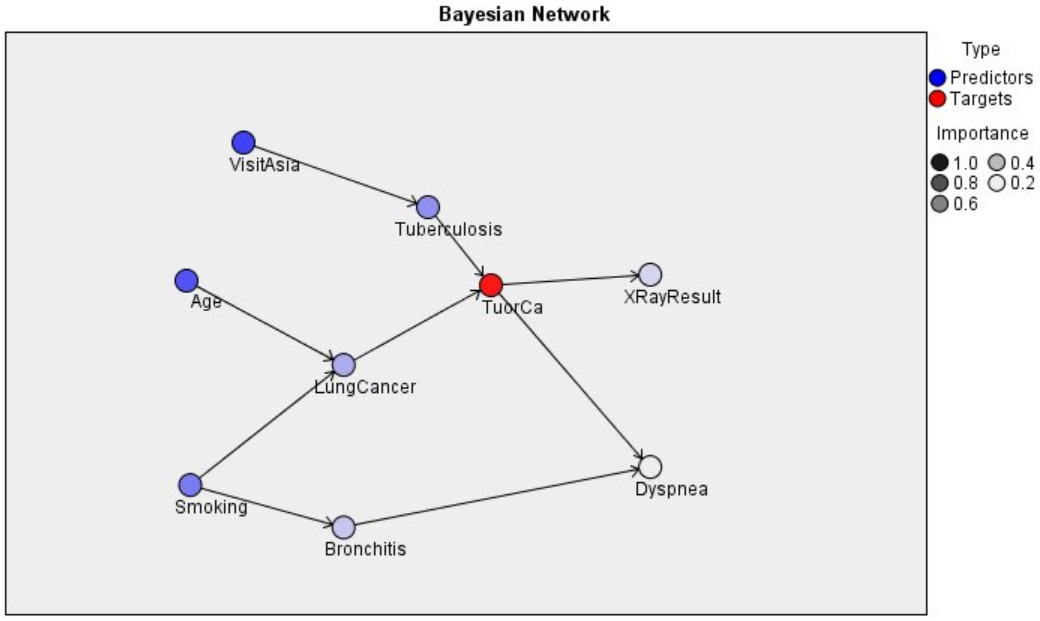


図 *29. Lauritzen* および *Spegelhalter* のアジア・ネットワークの例

Bayesian network (ベイズ) を使用する理由は、下記のとおりです。

- 因果関係について学習することができます。これにより、問題の領域を理解し、干渉の結果を予測できます。
- ネットワークは、データのオーバーフィットを防止するための効果的な手法を提供します。
- 関係の明確な視覚化が、容易に観測されます。

要件 :対象フィールドはカテゴリーでなければならず、測定の尺度は、名義型、順序型、またはフラグ型のどれでもかまいません。入力フィールドは、いかなるタイプのフィールドでもかまいません。連続した入力フィールド (数値範囲型) は自動的に分割されます。ただし、分散が歪んでいる場合、Bayesian network (ベイズ) ノードの前にデータ分割ノードを使用して手動でフィールドを分割し、より良い結果を取得できます。例えば、スーパバイザ フィールドが、Bayesian network (ベイズ) ノードの対象フィールドと同じ場合、最適データ分割を使用します。

例: 銀行のアナリストは、ローンの返済を履行しない顧客または潜在的顧客を予測できる必要があります。Bayesian network (ベイズ) モデルを使用して、滞納すると考えられる顧客の特性と特定し、複数のタイプのモデルを構築して潜在的な滞納者を予測するために最良のモデルを確定します。

例: 通信会社のオペレータは、解約する顧客 (「顧客離れ」) の数を減らし、前月のデータを使用して毎月ベースでモデルを更新したいと考えています。Bayesian network (ベイズ) モデルを使用し、離れると考えられる顧客の特性を特定し、新規データで毎月モデルの学習を継続します。

Bayesian network (ベイズ) ノードの「モデル」オプション

モデル名: ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

各分割のモデルの構築 : 分割フィールドとして指定される入力フィールドの各値の個別モデルを作成します。 詳しくは、トピック 28 ページの『分割モデルの作成』を参照してください。

データ区分: このフィールドでは、モデル構築の学習ステージ、テスト・ステージ、検証ステージ用に、データを個別のサンプルに区分するためのフィールドを指定することができます。1 組のサンプルをモデルの生成に使用し、別のサンプルで生成したモデルをテストすることにより、そのモデルが、このデータに似た性質を持つより大きなデータセットにどの程度適用できるかについての良い目安を得ることができます。データ型ノードまたはデータ区分ノードを使用して複数のデータ区分フィールドが定義されている場合、データ区分を使用する各モデル作成ノードの「フィールド」タブで単一のデータ区分フィールドを選択する必要があります。(1 つのデータ区分だけが存在している場合、データ分割を有効にすると、そのデータ区分が必ず自動的に使用されます)。また、選択したデータ区分を分析に適用するには、そのノードの「モデル・オプション」タブでデータ区分が有効になっている必要があります (このオプションの選択を解除すると、フィールド設定を変更することなくデータ区分を無効にすることができます)。

分割。分割モデルについて、分割フィールドを選択します。これは、データ型ノードのフィールドの役割を「分割」 に設定するのと似ています。測定の尺度が「フラグ型」、「名義型」、「順序型」または「連続型」のフィールドのみ、分割フィールドとして指定できます。分割フィールドとして選択されたフィールドは、対象フィールド、入力フィールド、データ区分フィールド、度数フィールドまたは重みフィールドとして使用できません。 詳しくは、トピック 28 ページの『分割モデルの作成』を参照してください。

既存モデルの学習を継続: このオプションを選択すると、モデル・ナゲットの「モデル」タブに表示される結果はモデルが実行されるたびに再生成されて更新されます。例えば、新規または更新済みのデータ・ソースを既存のモデルに追加する場合にこの処理を実行します。

注 : 既存のネットワークのみ更新されます。ノードまたは接続を追加または削除できません。モデルを再学習するごとに、ネットワークは同じ形状となり、条件付き確率および予測変数の重要度のみが変更されます。新しいデータが古いデータに類似している場合も、同じ事柄が重要であると考えるため、大きな問題ではありません。ただし、重要である事柄 を確認または更新する場合 (どれくらい重要であるかではなく)、新しいモデル、つまり新しいネットワークを構築する必要があります。

構造タイプ : Bayesian Network (ベイズ) を構築時に使用する構造を選択します。

- **TAN:** Tree Augmented Naïve Bayes モデル (TAN) は、標準の Naïve Bayes モデルの改良型である単純な Bayesian Network (ベイズ) モデルを作成します。これは各予測値が目標変数のほかに、別の予測値に依存することが可能となるため、その結果、分類精度を向上させることができます。
- **Markov Blanket:** 目標変数の親、その子、その子の親を含むデータ・セットのノード群を選択します。基本的に、Markov Blanket は目標変数を予測するために必要なネットワークのすべての変数を識別します。このネットワーク構築方法はより正確なものと考えられていますが、大きなデータセットの場合には、関連する多くの変数によって処理時間のペナルティーが生じることがあります。処理の量を削減するには、「エキスパート」タブの 「特徴量選択」 オプションを使用して、目標変数に特に関連する変数を選択できます。

特徴量選択の前処理ステップを含む: このボックスをオンにすると、「エキスパート」タブの 「特徴量選択」 オプションを使用できます。

パラメーター学習方法 : Bayesian Network (ベイズ) パラメーターは、親の値が与えられた各ノードの条件付き確率を参照します。親の値が認識されるノード間の条件付き確率テーブルを推定するタスクを制御するために、次の 2 つを選択できます。

- 最尤法 : 大きなデータセットを使用する場合は、このボックスを選択します。これがデフォルトの設定です。

- 小さいセルの度数の **Bayes** 調整 **:** 小さいデータセットの場合、ゼロ度数の上限の可能性とともにモデルがオーバーフィットする危険性があります。このオプションを選択すると、ゼロ度数の効果および信頼できない推定効果を減らす平滑法を適用してこれらの問題を緩和します。

Bayesian network (ベイズ) ノードの「エキスパート」オプション

ノードのエキスパート・オプションを使用すると、モデル構築プロセスを微調整できます。エキスパート・オプションを利用するには、「エキスパート」タブで「モード」に 「エキスパート」 を設定してください。

欠損値。デフォルトで、IBM SPSS Modeler ではモデルで使用されるすべてのフィールドに有効な値を持つレコードだけが使用されます。(これは、欠損値のリストごとの削除とも呼ばれます。)欠損値が大量にある場合は、この方法では多くのレコードが除外され、データ不足で適切なモデルを作成できなくなることがあります。このような場合、「完全なレコードのみ使用」 オプションを選択解除できます。IBM SPSS Modeler は、フィールドの一部に欠損値のあるレコードなど、モデルを推定するためにできる限り多くの情報を使用します(これは、欠損値のペアごとの削除とも呼ばれます)。ただし、状況によっては、このようにして不完全なレコードを使用すると、モデルの推定に計算上の問題が発生することがあります。

すべての確率を追加**:** 出力フィールドの各カテゴリーの確率を、ノードで処理される各レコードに追加するかどうかを指定します。このオプションを選択しないと、予測されたカテゴリーの確率だけが追加されます。

独立性検定**:** 独立性検定によって、2 つの変数のペアの観測がお互いに独立しているかどうかを評価します。使用される検定の種類を選択します。使用できるオプションは次のとおりです。

- 尤度比**:** 2 つの異なる仮説に基づく結果の最大確率間の比率を計算して、対象予測値の独立性を検定します。
- **Pearson** のカイ **2** 乗 **:** 観測されたイベントが発生する相対度数が指定された度数分布に従うという帰無仮説を使用して、対象予測値の独立性を検定します。

検定されたペアを超えて追加変数が使用される場合、Bayesian network (ベイズ) モデルは、独立性の条件検定を行います。さらに、モデルは対象値および予測値間の関係だけでなく、予測値自体の間の関係を探索します。

注**:** 「独立性検定」オプションは、「モデル」タブで Markov Blanket の「特徴量選択の前処理ステップを含む」または「構造タイプ」のいずれかを選択した場合にのみ使用できます。

有意水準**:** 独立性検定設定と組み合わせて使用し、検定実行時に使用されるカットオフ値を設定できます。値が低いと、ネットワーク内のリンクが少なくなります。デフォルトの水準は 0.01 です。

注**:** このオプションは、「モデル」タブで Markov Blanket の「特徴量選択の前処理ステップを含む」または「構造タイプ」のいずれかを選択した場合にのみ使用できます。

最大条件セットサイズ**:** Markov Blanket 構造を作成するためのアルゴリズムでは、サイズが増加する条件セットを使用して、独立性検定を実行し、ネットワークの不要なリンクを削除します。上限の条件変数を含む検定には処理するための時間およびメモリーが必要であるため、含まれる変数の数を制限できます。これは、多くの変数間で強い依存関係があるデータの処理をする場合に特に役に立ちます。ただし、結果として生じるネットワークには、不要なリンクが含まれている場合があります。

独立性検定に使用する条件変数の最大数を指定します。デフォルトは 5 です。

注: このオプションは、「モデル」タブで Markov Blanket の「特徴量選択の前処理ステップを含む」または「構造タイプ」のいずれかを選択した場合にのみ使用できます。

特徴量選択 : これらのオプションを使用すると、モデルを処理する場合に使用する入力数を制限し、モデル構築プロセスの時間を短縮できます。これは、多くの潜在入力数により Markov Blanket 構造を作成する場合に特に役に立ちます。目標変数に大きく関連する入力を選択できます。

注: 特徴量選択オプションは、「モデル」タブの「特徴量選択の前処理ステップを含む」を選択する場合にのみ使用できます。

- 常に選択された入力: フィールド ピッカー (テキスト フィールドの右側にあるボタン) を使用して、Bayesian network (ベイズ) モデルを構築する場合に常に使用するデータ セットのフィールドを選択します。この対象フィールドは常に選択されます。他の検定で有意でないと見なされた項目は、Bayesian network (ベイズ) でもモデル構築プロセス中にリストから除去される可能性があることに注意してください。そのため、このオプションは単にリスト内の項目がモデル構築プロセスで使用されることを指定するものであり、生成される Bayesian モデルにこれらの項目が必ず使用されることを保証するものではありません。
- 入力フィールドの最大数: Bayesian network (ベイズ) モデル構築時に使用するデータセットの入力フィールドの合計数を指定します。入力できる上限値は、データセットの入力フィールド数の合計です。

  注: 「常に選択された入力」で選択されたフィールド数が「入力フィールドの最大数」を超える場合、エラー・メッセージが表示されます。

Bayesian network (ベイズ) モデル・ナゲット

注: モデル作成ノードの「モデル」タブで「既存パラメータの学習を継続」を選択すると、このモデル・ナゲットの「モデル」タブに表示される情報が、モデルを再生成するたびに更新されます。

モデル・ナゲットの「モデル」タブは、2 つのペインに分かれています。

左側ペイン

基本: このビューには、対象とその最重要予測値の関係、および予測値同士の関係を表示する、ノードのネットワーク・グラフが表示されます。濃い色は重要な予測値を表し、薄い色は重要度の低い予測値を表します。

範囲を示すノードのビン値は、マウス・ポインターをノード上に移動すると、ツールヒントに表示されます。

IBM SPSS Modeler のグラフ ツールを使用して、グラフを対話的に操作することや、グラフを編集および保存することができます。例えば、MS Word などの他のアプリケーションで使用できます。

ヒント: ネットワークに多くのノードが含まれている場合は、ノードをクリックして選択し、ドラッグすることで、グラフをより見やすくすることができます。

分布: このビューでは、ネットワーク内の各ノードの条件付き確率が小さいグラフに表示されます。マウス・ポインターをグラフ上に移動すると、その値がツールヒントに表示されます。

右側ペイン

予測変数の重要度: モデルを推定する際の各予測値の相対重要度を示すグラフが表示されます。詳しくは、44 ページの『予測変数の重要度』を参照してください。

条件付き確率: 左側ペインでノードまたは小さい分布グラフを選択すると、関連する条件付き確率の表が右側ペインに表示されます。この表には、各ノード値および親ノードの値の各組み合わせの条件付き確率が含まれます。また、各レコード値および親ノードの値の各組み合わせの観察されたレコード数も含まれます。

Bayesian network (ベイズ) モデル設定

Bayesian network (ベイズ) モデル・ナゲットの「設定」タブは、構築したモデルを修正するオプションを指定します。例えば、同じデータと設定を用いていくつかの異なるモデルを構築するために Bayesian network (ベイズ) ノードを使用し、設定を少しだけ修正して結果に及ぼす影響を確認するにはそれぞれのモデルの同じタブを使用します。

注**:** このタブは、モデル・ナゲットがストリームに追加された後にのみ使用されます。

未調整傾向スコアを計算**:** フラグ型対象 (yes または no の予測を返す) を持つモデルの場合、対象フィールドに指定された真 (true) の結果の尤度を示す傾向スコアが必要な場合があります。また、スコアリング時に生成することができるその他の予測および確信度値があります。

調整済み傾向スコアを計算 **:** 未調整傾向スコアは、学習データにのみ基づき、このデータがオーバーフィットする多くのモデルの傾向によって過度に楽観的な場合があります。調整済み傾向は、テストまたは検証用データ区分に対してモデルのパフォーマンスを評価することによって補正しようとします。このオプションでは、モデルの生成前にデータ区分フィールドをストリーム内で定義し、調整済み傾向スコアがモデル作成ノードで有効化されている必要があります。

すべての確率を追加**:** 出力フィールドの各カテゴリーの確率を、ノードで処理される各レコードに追加するかどうかを指定します。このオプションを選択しないと、予測されたカテゴリーの確率だけが追加されます。

このチェック・ボックスのデフォルト設定は、モデル作成ノードの「エキスパート」タブの対応するチェック・ボックスによって決まります。 詳しくは、トピック 138 ページの『Bayesian network (ベイズ) ノードの「エキスパート」オプション』を参照してください。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

Bayesian network (ベイズ) モデル要約

モデル・ナゲットの「要約」タブで、モデルそのもの (精度分析)、モデルで使用するフィールド (フィールド)、モデルの構築時に使用する設定 (構築の設定)、およびモデルの学習 (学習の要約) についての情報を表示します。

ノードを初めて参照する場合、「要約」タブの結果は閉じられています。目的の結果を表示するには、項目の左側にある展開コントロールを使用して項目を展開するか、または 「すべて展開」 ボタンをクリックしてすべての結果を表示します。見終わった結果を隠すには、展開コントロールを使用して目的の結果を省略するか、または 「すべて閉じる」 ボタンをクリックしてすべての結果を非表示にします。

精度分析**:** 特定のモデルについての情報を表示します。

フィールド**:** 対象フィールドおよびモデル構築時の入力として使われるフィールドが表示されます。

構築の設定**:** モデル構築時に使われる設定情報が表示されます。

学習の要約 **:** モデルの種類、モデルの作成に使われたストリーム、モデルの作成者、モデルの作成日時、およびモデルの構築時間などの情報が表示されます。

第 8 章 ニューラル・ネットワーク

ニューラル・ネットワークは、モデルの構造および推定について最小限の要件で幅広い予測モデルの見積もりができます。関係の形式は、学習プロセスで決定します。対象フィールドと予測フィールドの線型の関係が適切である場合、ニューラル・ネットワークの結果から従来の線型モデルの結果を見積もります。非線型の関係がより適切である場合、ニューラル・ネットワークは自動的に「適切な」モデル構造を見積もります。

この柔軟性における矛盾点は、ニューラル・ネットワークが容易に解釈できないという点です。対象フィールドと予測フィールドの関係を構築する基底プロセスを説明しようとする場合、従来の統計モデルを使用することが適しています。ただし、モデルの解釈が重要でない場合、ニューラル・ネットワークを使用して適切な予測を取得できます。

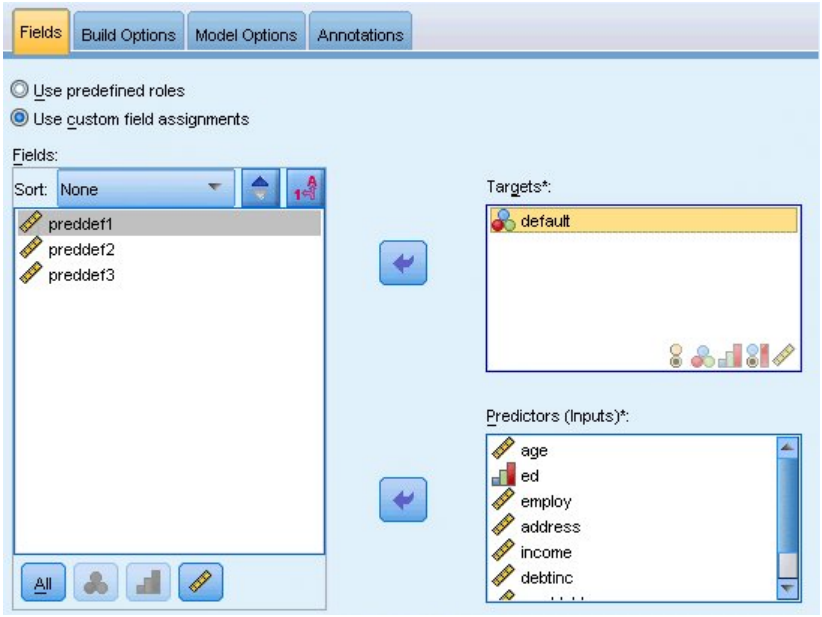


図 *30.* 「フィールド」タブ

フィールドの要件**:** 少なくとも 1 つの対象フィールドと、1 つの入力フィールドが必要です。「両方」または「なし」が設定されているフィールドは無視されます。対象フィールドまたは予測フィールド (入力) に測定の尺度の制限はありません。 詳しくは、 31 ページの『モデル作成ノードのフィールド・オプション』を参照してください。

モデルの構築中にニューラル・ネットワークに割り当てられる初期の重み (そのため、生成される最終モデル) は、データ内のフィールドの順序に依存します。SPSS Modeler は、学習のためにニューラル・ネットワークにデータを提示する前に、フィールド名によってデータを自動的にソートします。これは、ランダム・シードがモデル・ビルダーで設定されている場合、データ上流のフィールドの順序を明示的に変更しても、生成されるニューラル・ネット・モデルは影響を受けないことを意味します。ただし、ソート順を変更するように入力フィールド名を変更した場合、ランダム・シードがモデル・ビルダーで設定されていても、異なるニューラル・ネットワーク・モデルが生成されます。フィールド名のソート順が異なっていても、モデル品質は大きな影響は受けません。

ニューラル・ネットワーク・モデル

ニューラル・ネットワークは、神経系の動作を模倣した単純なモデルです。基本ユニットはニューロンと呼ばれ、次の図に示すように、層で編成されています。

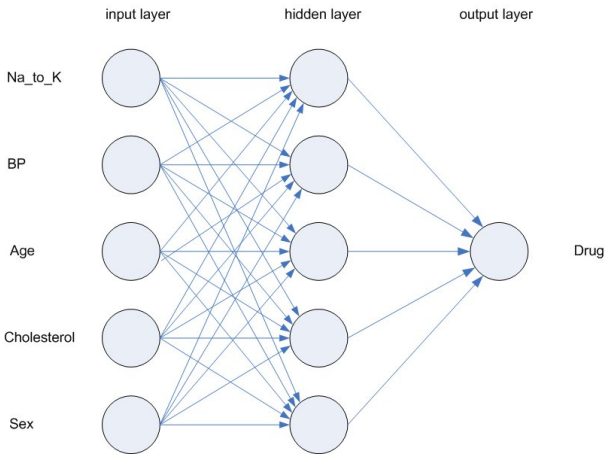


図 *31.* ニューラル・ネットワークの構造

ニューラル・ネットワークは、人間の脳が情報を処理する方法を単純化したモデルです。ニューラル・ネットワーク・ノードは、連係する多数の単純な処理単位をシミュレートします。処理ユニットは、ニューロンを抽象化したものと表現できます。

処理ユニットは、複数の層で編成されています。通常ニューラル・ネットワークは 3 つの部分から構成されています。入力フィールドを表すユニットから構成される入力層、隠れ層、および対象フィールドを表すユニットから構成される出力層。ユニットは、さまざまな接続強度 (重み) で接続されています。入力データが最初の層に送られ、その層の各ニューロンから次の層の全ニューロンに、値が伝達されます。最後に、結果が出力層から供給されます。

ネットワークは、各レコードを検証してレコードごとに予測を生成し、不正確な予測が行われた場合は重みを調整することで、学習していきます。この過程を何度も繰り返し、1 つ以上の停止基準が満たされるまで予測の改善を継続します。

当初は、重みはすべて無作為なので、ネットワークからの応答はあまり意味がありません。しかし、ネットワークは学習していきます。既知の結果の例が繰り返しネットワークに送られ、ネットワークからの応答と既知の結果が比較されます。この比較情報がネットワークに戻され、次第に重みが変更されていきます。学習が進むに従って、ネットワークの応答は精度を増し、既知の結果に近づいてきます。学習が終了すると、結果のわからない今後のケースに、ネットワークを適用できるようになります。

古いストリームでのニューラル・ネットワークの使用

バージョン 14 の IBM SPSS Modeler では、ブースティングおよびバギングの手法や非常に大きいデータセットの最適化をサポートする、新しいニューラル・ネット・ノードを導入しました。古いノードが含まれている既存のストリームでも、後のリリースでモデルを構築およびスコアリングできます。ただし、このサポートは今後のリリースで廃止されるため、新しいバージョンを使用することをお勧めします。

バージョン 13 以降では、値が不明 (学習データに値が存在しいない) のフィールドは自動的に欠損値として処理されず、値 `$null$` としてスコアリングされます。そのため、バージョン 13 以降で値が不明のフィ

ールドを以前 (13 より前) のニューラル・ネットワーク・モデルを使用して Null 以外の値としてスコアリングしたい場合、不明の値を欠損値としてマークする必要があります (例: データ型ノードを使用)。

互換性を維持するために、古いノードを依然として含んでいる古いストリームは、「ツール」>「ストリーム・プロパティー」>「オプション」の「設定サイズの制限 *(Limit set size)*」 を使用している場合があります。このオプションは、バージョン 14 以降は、Kohonen ネットおよび *K*-Means ノードにのみ適用されます。

目的

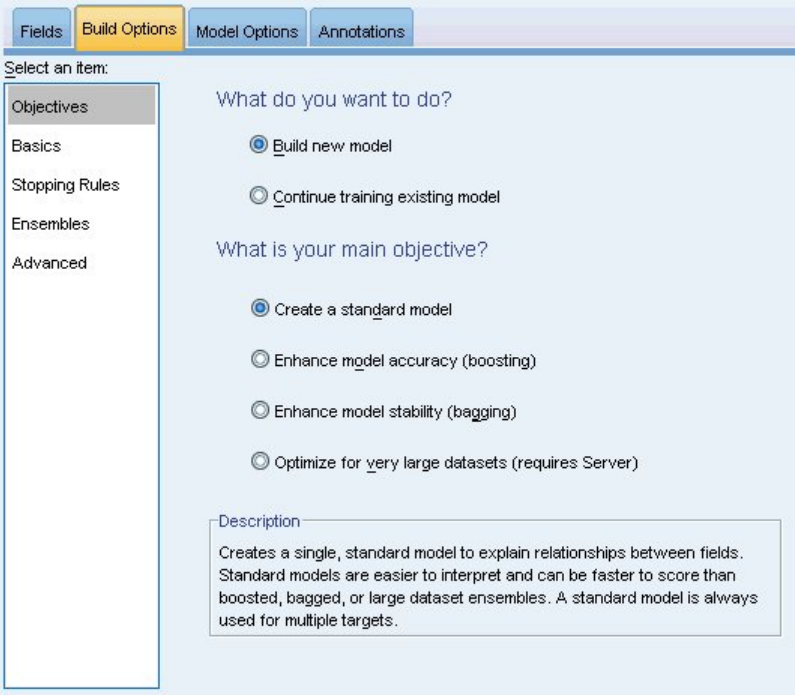


図 *32.* 「目的」の設定

実行する作業

- 新しいモデルを作成: 完全に新しいモデルを作成します。これはノードの役立つ操作です。
- 既存モデルの学習を継続: ノードによって正常に作成された最後のモデルで学習が継続します。これにより、元のデータにアクセスすることなく既存のモデルを更新またはリフレッシュできます。また、新規レコードまたは更新されたレコードのみがストリームに適用されるため、パフォーマンスが大幅に向上する場合があります。以前のモデルの詳細はモデル作成ノードで保存されるので、以前のモデル・ナゲットがストリームまたは「モデル」パレットでもう使用できない場合でも、このオプションを使用することができます。

  注**:** このオプションが有効な場合、「フィールド」タブと「作成オプション」タブにある他のすべてのコントロールが無効になります。

主な目的は**?**: 該当する目的を選択します。

- 標準モデルを作成: この方法では、予測変数を使用して対象を予測する単一モデルが作成されます。一般的に、ブースティング、バギング、または大規模なデータ・セット・アンサンブルと比べ、標準モデルは解釈が容易であり、素早くスコアリングできます。

注**:** 分割モデルの場合、「既存モデルの学習を継続」とともにこのオプションを使用するには、Analytic Server に接続されている必要があります。

- モデル精度の強化 **(**ブースティング**):** ブースティングを使用して、アンサンブルモデルを構築します。この方法では、ブースティングを使用してアンサンブル・モデルが作成されます。これによって、より正確な予測を得るための一連のモデルが生成されます。アンサンブルは、標準モデルと比べて作成とスコアリングに時間がかかる場合があります。

  ブースティングによって一連の「コンポーネント・モデル」が生成されます。各コンポーネント・モデルはデータ・セット全体に作成されます。連続する各コンポーネント・モデルを作成する前に、以前のコンポーネント・モデルの残差に基づきレコードに重みが付けられます。残差が大きいケースには比較的大きな分析の重みが与えられるため、次のコンポーネント・モデルは、これらのレコードの予測に重点を置きます。これらのコンポーネント・モデルがまとまってアンサンブル・モデルを形成します。アンサンブル・モデルは結合ルールを使用して新しいレコードをスコアリングします。使用できる規則は、対象の測定レベルによって異なります。
- モデルの安定性を拡張 **(**バグ**):** バギング (ブートストラップ集計) を使用して、アンサンブルモデルを構築します。この方法では、バギング (ブートストラップ集計) を使用してアンサンブル・モデルが作成されます。これによって、より信頼性の高い予測を得るための複数のモデルが生成されます。アンサンブルは、標準モデルと比べて作成とスコアリングに時間がかかる場合があります。

  ブートストラップ集計 (バギング) は、元のデータ・セットから置換を行うサンプリングによって、学習データ・セットの複製を作成します。これにより、元のデータ・セットと同じサイズのブートストラップ・サンプルが作成されます。その後、「コンポーネント・モデル」が複製ごとに作成されます。これらのコンポーネント・モデルがまとまってアンサンブル・モデルを形成します。アンサンブル・モデルは結合ルールを使用して新しいレコードをスコアリングします。使用できる規則は、対象の測定レベルによって異なります。
- 非常に大きなデータ セットのモデルを作成**:** この方法では、データ・セットを別々のデータ・ブロックに分割することにより、アンサンブル・モデルが作成されます。上記のモデルのいずれかを作成するにはデータ・セットが大きすぎる場合、または増分モデル作成の場合、このオプションを選択します。このオプションは、作成には時間がかからないものの、標準モデルと比べてスコアリングに時間がかかる場合があります。

複数の対象がある場合、選択した目的に関係なく、この方法では標準モデルを作成するだけです。

基本

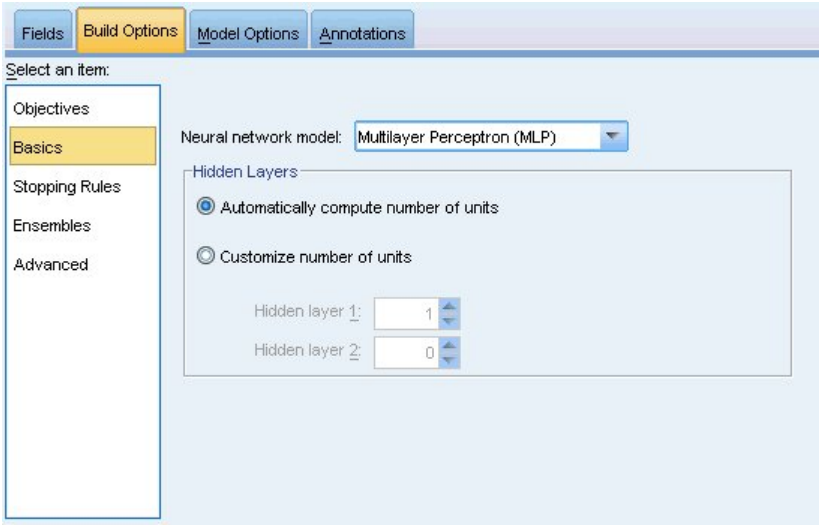


図 *33.* 「基本」の設定

ニューラル・ネットワーク・モル**:** このモデルを使用して、ネットワークが隠れ層を介して予測フィールドを対象フィールドにどのように接続するかを決定します。多層パーセプトロン **(MLP)** は、学習およびスコアリングに時間のかかる、より複雑なリレーションシップに使用できます。放射基底関数 **(RBF)** は、学習およびスコアリングに時間はかかりませんが、MLP と比較して予測の精度が低くなります。

隠れ層。ニューラル・ネットワークの隠れ層には、観測不可能な単位が含まれています。各隠れ層の単位は予測フィールドの関数です。関数の正確な形式は、ネットワークの種類によって一部異なります。多層パーセプトロンには 1 つまたは 2 つの隠れ層があり、放射基底関数ネットワークには 1 つの隠れ層があります。

- 単位数を自動的に計算**:** 隠れ層が 1 つのネットワークを構築し、隠れ層に最適な数の単位を計算します。
- 単位数をカスタマイズ**:** 隠れ層ごとに単位数を指定できます。最初の隠れ層には少なくとも 1 つの単位を指定する必要があります。2 番目の隠れ層の単位数を 0 と指定すると、隠れ層が 1 つの多層パーセプトロンが構築されます。

  注**:** ノード数が連続型予測フィールドの数とすべてのカテゴリー型 (フラグ型、名義型、順序型) 予測フィールドのカテゴリー数の合計を合わせた数を超えないように値を選択する必要があります。

停止規則

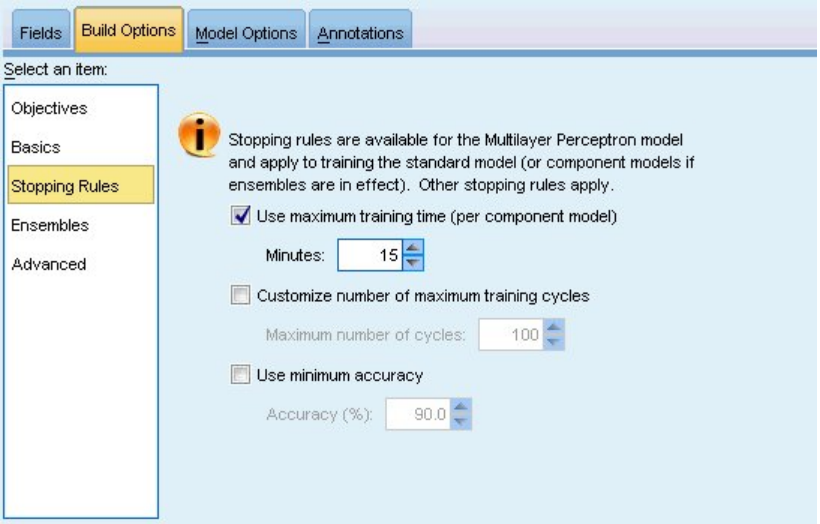


図 *34.* 「停止規則」の設定

これらは、多層パーセプトロン ネットワークの学習を停止する時期を決定する規則です。これらの設定は、放射基底関数アルゴリズムが使用される場合は無視されます。学習は少なくとも 1 回のサイクル (データ・パス) で続行し、次の基準に従って停止できます。

(コンポーネント・モデルあたりの**)** 最大学習時間を使用**:** アルゴリズムを実行する最大時間 (分単位) を指定するかどうかを選択します。0 より大きい数値を指定してください。アンサンブル・モデルを構築する場合、この値が、アンサンブルの各コンポーネント・モデルで許可される学習時間になります。最後のサイクルを完了するために指定の制限時間を多少超えることがあります。

最大学習サイクル数をカスタマイズ**:** 可能な最大学習サイクル数。最大学習サイクル数を超えると、学習が停止します。サイクルの最大数を超えた場合、学習が停止します。0 より大きい整数を指定します。

最小精度を使用**:** このオプションを選択すると、指定の精度に達するまで学習が続行されます。指定の精度に達しない可能性もありますが、任意の時点で学習を中断し、それまでに達成された最高精度のネットワークを保存することができます。

各サイクルの後オーバーフィット防止セットのエラーが減らない場合、学習エラーの変化が比較的小さい場合、または現在の学習エラーが最初のエラーと比較して小さい場合も、学習アルゴリズムが停止します。

アンサンブル

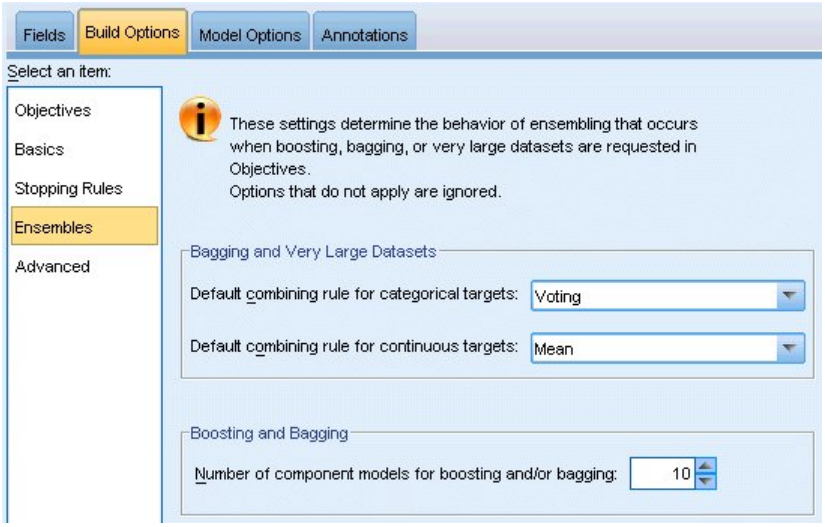


図 *35.* 「アンサンブル」の設定

これらの設定は、「目的」で、ブースティング、バギング、または非常に大きなデータ・セットが要求される場合に起きるアンサンブルの動作を決定します。選択された目的に適用されないオプションは無視されます。

バギングおよび非常に大きなデータ・セット: アンサンブルをスコアリングする場合、基本モデルの予測値を結合するために使用するルールで、アンサンブル・スコア値を計算します。

- カテゴリー型対象のデフォルト結合ルール。 カテゴリー型対象のアンサンブル予測値は、票決、最も高い確率、または最も高い平均確率を使用して結合できます。「票決」は、基本モデルで最も頻繁であり、最も確率が高いカテゴリーを選択します。「高確率」は、すべての基本モデルで最も高い単独の確率に達したカテゴリーを選択します。「最高平均確率」は、基本モデルでカテゴリーの確率が平均化された場合の、最も値の高いカテゴリーを選択します。
- 連続型対象のデフォルトの結合規則: 連続型対象のアンサンブル予測値は、基本モデルの予測値の平均または中央地を使用して結合できます。

モデルの精度を上げることが目的である場合、結合規則の選択が無視されることに注意してください。ブースティングは常に、カテゴリー対象のスコアリングには重み付き多数決を使用し、連続型対象のスコアリングには重み付き中央値を使用します。

ブースティングおよびバギング: バギングの場合は、ブートストラップ数となります。正の整数でなければなりません。

拡張

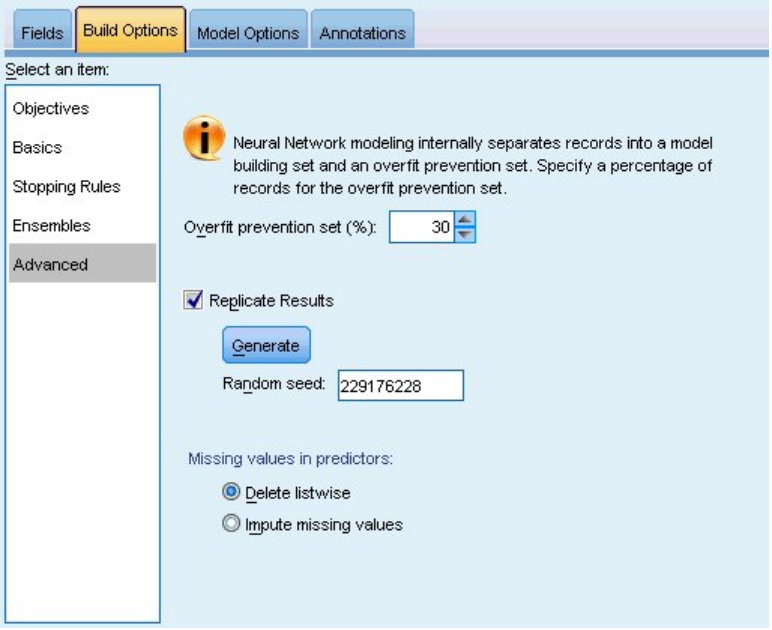


図 *36.* 詳細設定

詳細設定では、設定のその他のグループにあまり適合しないオプションの内容を変更することができます。

オーバーフィット防止セット。 ニューラル・ネットワーク メソッドは、レコードをモデル作成セットとオーバーフィット防止セットに内部的に分割します。オーバーフィット防止セットは学習時のエラーの追跡に使用されるデータ・レコードの独立したセットで、メソッドがデータ内の偶然変動のモデル作成を行わないようにします。レコードの割合を指定します。デフォルトは 30 です。

結果の再現: ランダム・シードを設定すると、分析を再現することができます。整数を指定するか、「生成」をクリックします。「生成」をクリックすると、1 から 2147483647 までの整数の疑似乱数が作成されます。デフォルトでは、分析は、シード 229176228 で複製されます。

予測フィールドの欠損値**:** 欠損値の処理方法を指定します。リストごとに削除すると、予測フィールドに欠損値のあるレコードがモデル構築から削除されます。欠損値を代入すると、予測フィールドの欠損値が置き換えられ、これらのレコードが分析に使用されます。連続型フィールドは、観測値の最小値および最大値の平均を代入します。カテゴリー型フィールドでは、最も頻度の高いカテゴリーを代入します。「フィールド」タブで指定されたその他のフィールドに欠損値があるレコードは、必ずモデルの作成から除外されます。

モデル・オプション

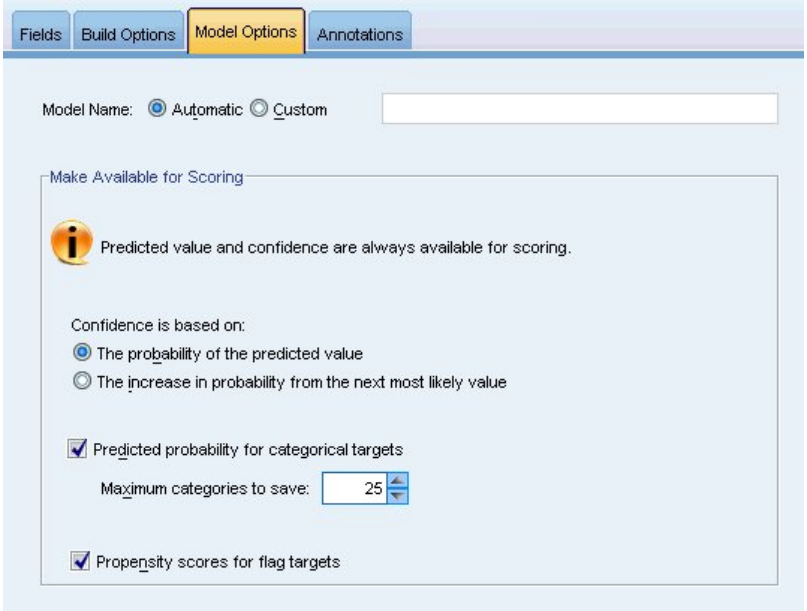


図 *37.* 「モデル・オプション」タブ

モデル名: 対象フィールドに基づいて自動的にモデル名を生成するか、またはカスタム名を指定できます。自動的に生成される名前は、対象フィールド名です。複数の対象がある場合、モデル名はそれらのフィールド名が順番にアンパサンドで区切られた形式となります。例えば、対象フィールドが *field1 field2 field3* の場合、モデル名は *field1 & field2 & field3* となります。

スコアリングで使用可能にする：モデルをスコアリングする場合、このグループで選択された項目を作成する必要があります。すべての対象フィールドの予測された値とカテゴリー型対象の確信度は、モデルをスコアリングする場合必ず計算されます。計算される確信度は、予測値の確率 (最も高い予測確率) または最も高い予測確率と 2 番目に高い予測確率との差を基準とする場合があります。

- カテゴリー型対象の予測確率： カテゴリー型対象の予測確率を生成します。カテゴリーごとにフィールドが作成されます。
- フラグ型対象の傾向スコア: フラグ型対象フィールド (yes または no の予測を返す) を持つモデルの場合、対象フィールドに指定された真 (true) の結果の尤度を示す傾向スコアが必要な場合があります。モデルは、傾向スコア (調整なし) を生成します。データ区分が有効な場合、モデルは検定データ区分に基づいて、調整済み傾向スコアも生成します。

モデルの要約

Model Summary

Target	Previously defaulted
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	4

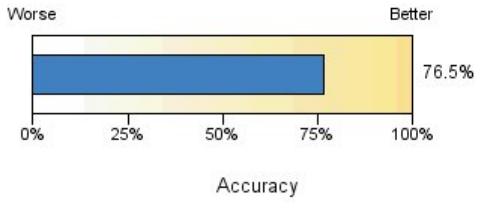


図 *38.* 「ニューラル・ネットワーク・モデルの要約」ビュー

「モデルの要約」ビューはスナップショットで、ニューラル・ネットワークの予測または分類の精度についての要約が一目でわかります。

モデルの要約。テーブルには、対象、学習したニューラル・ネットワークの種類、学習を停止した停止規則(多層パーセプトロン・ネットワークを学習した場合に表示)、ネットワークの隠れ層ごとのニューロン数が表示されます。

ニューラル ネットワークの品質**:** このグラフには、最終モデルの精度が表示されます (値が大きいほど精度が高いことを示します)。カテゴリー型対象の場合は、予測値が観測値に一致するレコードの割合となります。連続型対象の場合は、精度が R^2 値で示されます。

複数の対象 **(Multiple targets):** 対象が複数ある場合は、各対象がテーブルの「対象」行に表示されます。グラフに表示される精度は各対象の精度の平均です。

予測変数の重要度

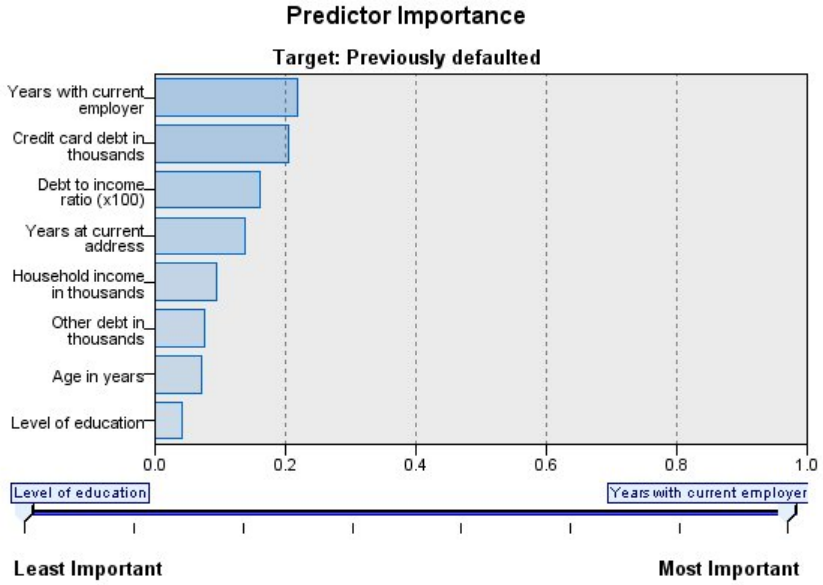


図 *39.* 「予測変数の重要度」ビュー

一般にモデリングの作業では、最も重要な予測フィールドの編集に集中して取り組み、最も重要でない予測フィールドは削除するか無視してしまいたくなるものです。予測値の重要度グラフを使用すると、モデル推定時に各予測値の相対重要度が示されるので便利です。値が相対的であるため、表示されるすべての予測変数の値の合計は 1.0 となります。予測値の重要度は、モデルの精度に関連しません。予測が正確かどうかに関係なく、予測時における各予測の重要度にのみ関連します。

複数の対象**:** 複数の対象がある場合、各対象は個別のグラフに表示され、表示する対象を制御する 「対象」ドロップダウンがあります。

予測と観測

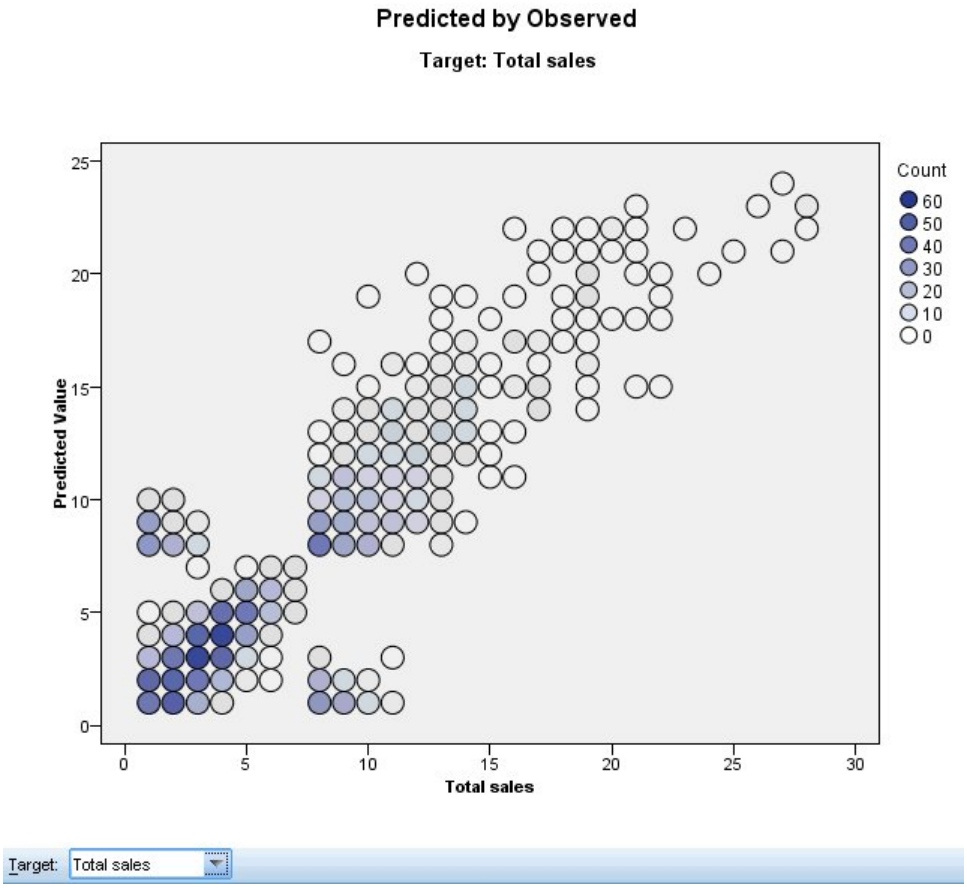


図 *40.* 「予測対観測」ビュー

連続型対象の場合、縦軸に予測値を、横軸に観測値を示した分割散布図を表示します。

複数の対象**:** 複数の連続型対象がある場合、各対象は個別のグラフに表示され、表示する対象を制御する「対象」ドロップダウンがあります。

分類

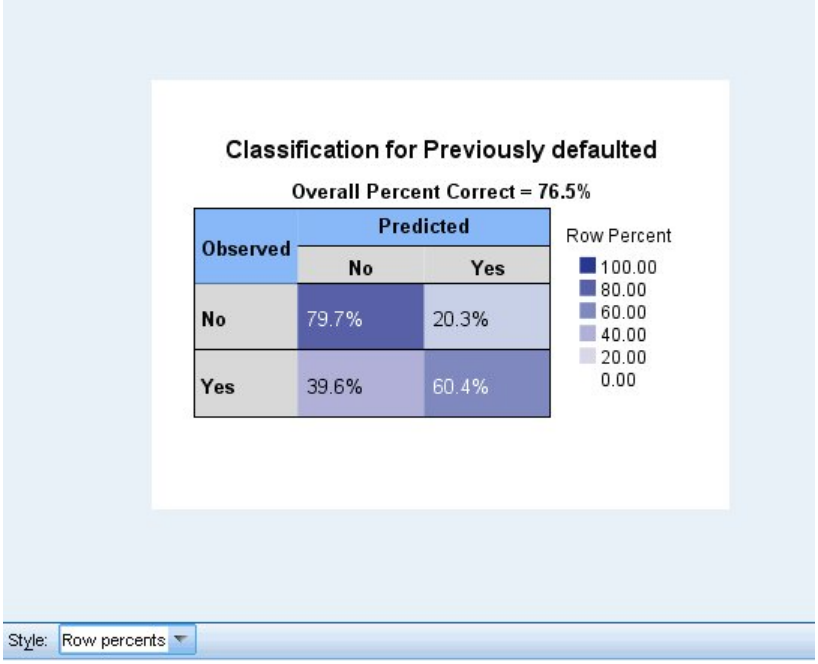


図 *41.* 「分類」ビュー、行パーセントのスタイル

カテゴリー型対象の場合、観測値と予測値のクロス分類と、すべての正分類パーセントをヒート・マップに表示します。

テーブルのスタイル。さまざまな表示スタイルがあり、「スタイル」ドロップダウン・リストからアクセスできます。

- 行パーセント。セルの行パーセント (行合計のパーセントで表現されるセルの度数) が表示されます。これがデフォルトです。
- セルの度数。セルのセル度数が表示されます。ヒート・マップの色の濃さは、行パーセントに基づいています。
- ヒート・マップ。セルの値は表示しません。陰影付けのみ表示します。
- 圧縮。セルの行または列のヘッダー、セルの値を表示しません。この方法は、対象にカテゴリー数が多い場合に役立ちます。

欠損値。対象に欠損値があるレコードがある場合、レコードはすべての有効な行の下の 「**(欠損値)**」行に表示されます。欠損値のあるレコードは、すべての正分類パーセントには貢献しません。

複数の対象。複数のカテゴリー対象がある場合、各対象は別々のテーブルに表示され、「対象」ドロップダウン・リストを使用して表示する対象を制御します。

大型テーブル。表示する対象に 100 を超えるカテゴリーがある場合、テーブルは表示されません。

ネットワーク

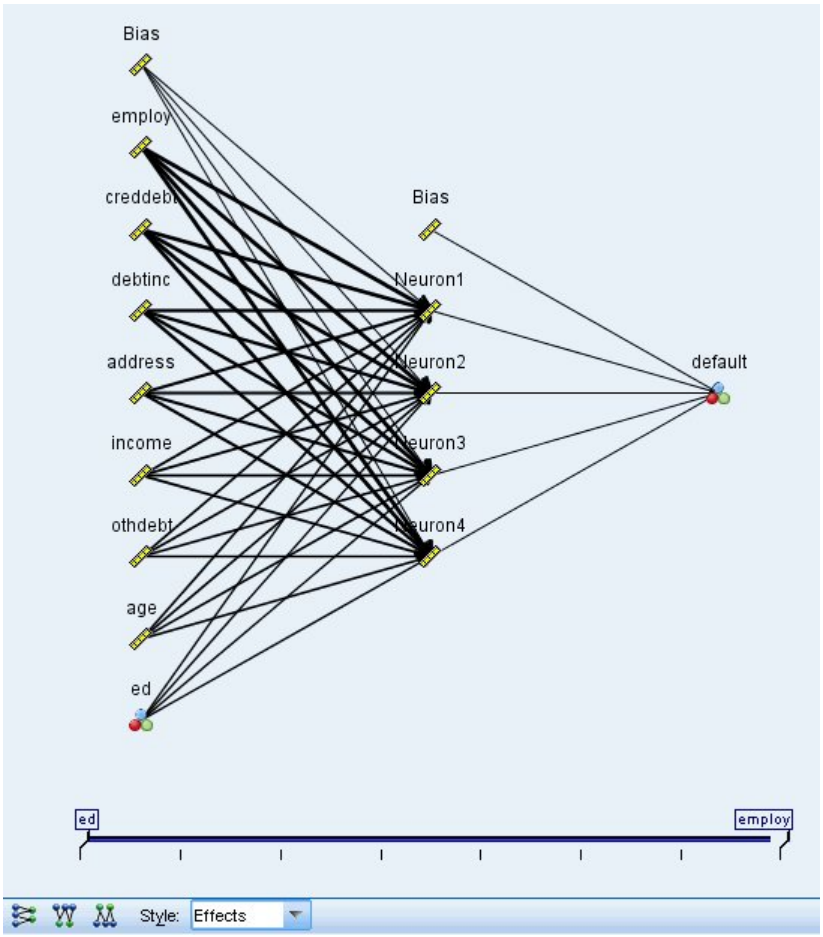


図 *42.* 「ネットワーク」ビュー、左側に入力、効果のスタイル

ニューラル・ネットワークのグラフィカルな表示が行われます。

グラフ・スタイル: 2 つの異なる表示スタイルがあり、「スタイル」ドロップダウンから選択できます。

- 効果**:** 各予測と対象が、測定値の尺度が連続型かカテゴリー型かに関係なく、1 つのノードとしてダイアグラムに表示されます。これはデフォルトです。
- 係数**:** カテゴリー型予測および対象に複数のインジケーター・ノードが表示されます。係数スタイルのダイアグラムでつながった線は、シナプスの重みの推定値に基づいて色分けされます。

ダイアグラムの方向**:** デフォルトでは、ネットワーク・ダイアグラムは、左側に入力、右側に対象が配置されます。ツールバーのコントロールを使用して、入力を上、対象を下に表示、あるいは入力を下、対象を上に表示するよう、方向を変更できます。

予測値の重要度ダイアグラム内の接続線には予測変数の重要度に基づいて重みが付けられ、線が太いほど重要度が高いことを示します。ツールバーの「予測変数の重要度」スライダーで、ネットワーク・ダイアログ内に表示された予測値を制御します。このスライダーを使用してもモデルは変更されませんが、最も重要な予測値に焦点を当てることができます。

複数の対象**:** 複数の対象がある場合、すべての対象がグラフに表示されます。

設定

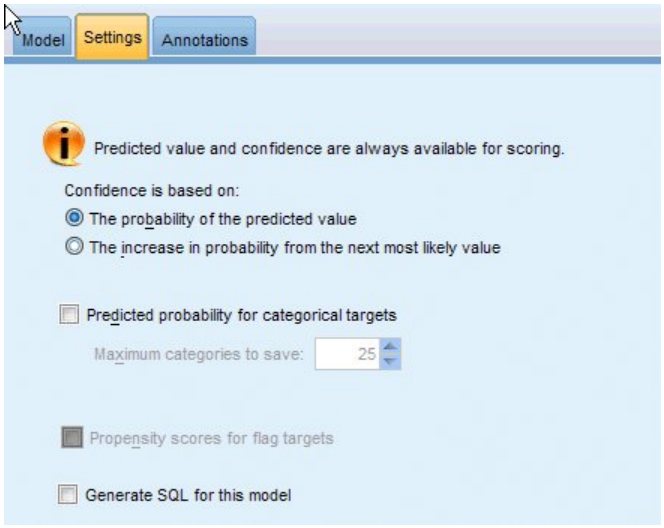


図 *43.* 「設定」タブ

モデルをスコアリングする場合、このタブで選択された項目を作成する必要があります。すべての対象フィールドの予測された値とカテゴリー型対象の確信度は、モデルをスコアリングする場合必ず計算されます。計算される確信度は、予測値の確率 (最も高い予測確率) または最も高い予測確率と 2 番目に高い予測確率との差を基準とする場合があります。

- カテゴリー型対象の予測確率： カテゴリー型対象の予測確率を生成します。カテゴリーごとにフィールドが作成されます。
- フラグ型対象の傾向スコア**:** フラグ型対象フィールド (yes または no の予測を返す) を持つモデルの場合、対象フィールドに指定された真 (true) の結果の尤度を示す傾向スコアが必要な場合があります。モデルは、傾向スコア (調整なし) を生成します。データ区分が有効な場合、モデルは検定データ区分に基づいて、調整済み傾向スコアも生成します。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。

ネイティブの **SQL** に変更してスコアリング**:** これを選択すると、データベース内でモデルをスコアリングするためのネイティブ SQL が生成されます。

注**:** このオプションの方が短時間で結果を得ることができますが、モデルの複雑度が増大すると、ネイティブ SQL のサイズと複雑度も、それに応じて増大します。

データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

第 9 章 ディシジョン・リスト

ディシジョン・リスト モデルは、母集団に関連する与えられた 2 値 (yes/no) の結果の高いもしくは低い尤度を示すサブグループまたはセグメントを識別します。例えば、離れる可能性の最も少ない、もしくはキャンペーンに好意的に参加する可能性のある顧客を探すことができます。ディシジョン・リスト・ビューアー を使用すると、モデルを完全に制御でき、さらにセグメントを編集し独自のビジネス・ルールを追加、各セグメントのスコアリング方法を指定し、さまざまな方法でモデルをカスタマイズしてすべてのセグメントのヒット数の割合を最適化します。それは、メーリング・リストの生成またはどのレコードを特定のキャンペーンのターゲットとするかの識別に適しています。複数のマイニング・タスクを使用し、例えば同一モデル内の高度または低度の実行セグメントを識別し、必要に応じてスコアリングの段階でそれぞれを選択または除外することで、モデル作成方法を結合することができます。

セグメント、ルールおよび条件

モデルはセグメントのリストで構成され、それぞれは一致するレコードを選択するルールによって定義されます。指定されたルールには、複数の条件があります。例えば、次のとおりです。

```
RFM_SCORE > 10 and
MONTHS_CURRENT <= 9
```

ルールは与えられたレコードに対して順番に適用され、最初に一致するルールによって結果が判別されます。それぞれを見ると、ルールまたは条件が重複していることがありますが、ルールの順番によって曖昧さを解決します。一致するルールがない場合、レコードは残りのルールに割り当てられます。

スコアリングの完全な制御

ディシジョン・リスト・ビューアー を使用すると、セグメントを参照、修正、再構築し、スコアリングのためにどのセグメントを選択または除外するかを選ぶことができます。例えば、今後のオファーから顧客グループを 1 つ除外して他の顧客グループを追加し、全体のヒット率にどのように影響を与えるかをすぐに確認できます。ディシジョン・リスト モデルの含まれるセグメントに対しては *Yes* を返し、残りも含めたその他に対しては *$null$* を返します。このスコアリングに対する直接的な制御によって、ディシジョン・リスト モデルはメーリング リストの生成には理想的なモデルとなり、コール・センターまたはマーケティング・アプリケーションを含む顧客関係管理システムで幅広く使用されています。

マイニング・タスク、測定および選択

モデル作成プロセスは、マイニング・タスクによって決まります。各マイニング・タスクは新規モデル作成の実行を効果的に開始し、代替モデルの新しいセットを返してそこから選択します。デフォルト・タスクはディシジョン・リスト ノードの初期の指定に基づいていますが、カスタム タスクのすべての数値を定義することができます。タスクをインタラクティブに適用することもできます。例えば、 学習セット全体で高い確率の検索を実行し、その後残りで低い確率の検索を行い、低度の実行セグメントを除外します。

データ選択

モデルの構築および評価のために、データの選択とカスタムのモデル測定を定義することができます。例えば、マイニング・タスクでデータの選択を指定してモデルを特定の地域に合わせて調整し、ユーザー定義の測定を作成して国全体でモデルがどのように実行されているのか評価することができます。マイニング・タスクと違い、測定は基になっているモデルを変更しませんが、別のレンズを提供して実行状態を評価します。

ビジネスに関する知識の追加

アルゴリズムに識別されたセグメントを調整または拡張し、ディシジョン・リスト・ビューアー を使用すると、ビジネスに関する知識をモデルに導入することができます。モデルによって生成されたセグメントを編集し、指定のルールに基づいてセグメントを追加することができます。その後、変更を適用し結果をプレビューすることができます。

さらに詳しい調査については、Excel とのダイナミック・リンクを使用すると、データを Excel にエクスポートし、そこでプレゼンテーション用のグラフを作成したり、複合利益や ROI 指標などのカスタム指標を計算する、あるいはモデルを構築しながらそれらを ディシジョン・リスト・ビューアー で表示するなどの作業ができます。

例: 金融機関のマーケティング部門では、それぞれの顧客に合った適切な提案を行うことで、今後さらに収益を上げることを望んでいます。ディシジョン・リスト・モデルを使用すると、以前の販売促進を基に顧客が最も好意的な反応を示す特徴を識別し、その結果に基づいてメーリング リストを生成できます。

要件: 予測する 2 値の結果 (yes/no) を示すフラグ型 または名義型 の測定の尺度を持つ単一カテゴリー対象フィールドと、1 つ以上の入力フィールド。対象フィールドが名義型の場合は、ヒントまたは回答として処理される単一の値を手動で選択する必要があります。その他の値はすべて、ヒントでないとして一まとめにされます。任意でフリクエンシ フィールドも指定することができます。連続する日付/時刻型フィールドは無視されます。連続する数値範囲型の入力フィールドは、モデル作成ノードの「エキスパート」タブで指定されたアルゴリズムにより自動的に分割されます。分割を細かく制御するために、アップストリームにデータ分割ノードを追加し、測定の尺度が順序型の入力として分割フィールドを使用します。

ディシジョン・リストのモデル関連のオプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

分割モデルを作成 **:** 分割フィールドとして指定される入力フィールドの各値の個別モデルを作成します。詳しくは、 28 ページの『分割モデルの作成』を参照してください。

モード**:** モデルを構築するために使用する方法を指定します。

- モデルの生成**:**ノードが実行されるときにモデル・パレット上に自動的にモデルを生成します。結果のモデルをスコアリング目的でストリームに追加することはできますが、以後の編集はできません。
- インタラクティブ セッションの起動**:** 対話式の ディシジョン・リスト・ビューアー モデル作成 (出力) ウィンドウが開きます。このウィンドウで、複数の代替案から選択して異なる設定値で繰り返しアルゴリズムを適用することにより、段階的にモデルを大きくしたり変更したりすることができます。詳しくは、トピック 163 ページの『ディシジョン・リスト・ビューアー』を参照してください。
- 保存済みインタラクティブ セッション情報の使用**:** 前回保存した設定を使用してインタラクティブ セッションを起動します。以前に保存した設定を使用してインタラクティブ・セッションを起動します。インタラクティブ・セッションの設定は、「生成」メニュー (モデルまたはモデル作成ノードを作成するため) または「ファイル」メニュー (セッションが起動されたノードを更新するため) を使用して、ディシジョン・リスト・ビューアー から保存できます。

対象値**:** モデル化する結果となる、対象フィールドの値を指定します。例えば、対象フィールドで顧客離れを「0 = no」および「1 = yes」とコーディングしている場合、どのレコードが他社に乗り換えそうかを表すルールを識別するには 1 と指定します。

次を含むセグメントを検索**:** 対象の変数の検索で 「高確率」 または 「低確率」 のどちらの発生を検索するかを示します。それらを検出して除外すると効率的にモデルを改善できます。これは剰余の度数が低いときは特に効果的です。

最大セグメント数**:** 返される最大のセグメント数を指定します。上位 *N* 個のセグメントが作成されます。確率が最も高いセグメント、または複数のモデルで確率が同じ場合は、範囲が最も広いセグメントが最善のセグメントとなります。設定可能な最小値は 1 です。最大値はありません。

最小セグメント サイズ**:** 下に示される 2 つの設定で、最小セグメント・サイズを指定します。2 つの値の大きい方が優先されます。例えば、パーセンテージ値が絶対値よりも大きい場合は、パーセンテージの設定が優先されます。

- 前のセグメントのパーセント **(%)** として**:** グループの最小サイズをレコードのパーセンテージとして指定します。設定可能な最小値は 0 です。設定可能な最大値は 99.9 です。
- 絶対値として **(N):** グループの最小サイズをレコードの絶対数として指定します。設定可能な最小値は 1 です。最大値はありません。

セグメント ルール**:**

最大属性数**:** セグメント・ルールあたりの条件の最大数を指定します。設定可能な最小値は 1 です。最大値はありません。

- 属性の再使用の許可**:** 有効にすると、前のサイクルで使用されたものも含めて、すべての属性が各サイクルで考慮されます。各サイクルで新しい条件が追加されるため、セグメントの条件はサイクルごとに累積していきます。サイクルの数は 「最大属性数」 設定を使用して定義します。

新しい条件の信頼区間 **(%):** セグメントの有意性をテストするための信頼水準を指定します。この設定は、返されるセグメントがある場合はその数と、「セグメントあたりの条件数」のルールに大きな影響を与えます。この値を大きくするほど、返される結果セットが少なくなります。設定可能な最小値は 50 です。設定可能な最大値は 99.9 です。

ディシジョン・リスト・ノードのエキスパート関連のオプション

エキスパート・オプションを使用すると、モデル構築プロセスを微調整できます。

データ分割手段**:** 連続フィールドを分割するために使用する方法です (等カウントまたは等幅)。

ビン数**:** 連続フィールドに対して作成するビンの数です。設定可能な最小値は 2 です。最大値はありません。

モデル検索幅。次のサイクルに使用できる、サイクルあたりのモデル結果の最大数です。設定可能な最小値は 1 です。最大値はありません。

ルール検索幅。次のサイクルに使用できる、サイクルあたりのルール結果の最大数です。設定可能な最小値は 1 です。最大値はありません。

ビンのマージ因子。セグメントを隣のセグメントと結合するときに、セグメントを拡張すべき最小量です。設定可能な最小値は 1.01 です。最大値はありません。

- 条件内の欠損値を許可: True の場合は、規則内で IS MISSING テストを使用することができます。
- 中間結果を破棄: True の場合は、検索プロセスの最終結果だけが返されます。最終結果は、検索プロセスでそれ以上調整されない結果です。False の場合は、中間結果も返されます。

  代替の最大数 : マイニング・タスクを実行して返される代替の最大数を指定します。設定可能な最小値は 1 です。最大値はありません。

  マイニング・タスクは、指定されている最大数まで、実際の代替数のみを返します。例えば、最大数が 100 に設定され、3 つの代替のみが検出された場合、その 3 つのみが表示されます。

ディシジョン・リスト・モデル・ナゲット

モデルはセグメントのリストで構成され、それぞれは一致するレコードを選択するルールによって定義されます。モデルを生成する前にセグメントを簡単に表示または変更し、含めたり除外したりするセグメントを選択できます。スコアリングに使用される場合、ディシジョン・リスト・モデルは含まれるセグメントに対しては *Yes* を返し、残りも含めたその他に対しては *$null$* を返します。この直接的な制御によって、ディシジョン・リスト・モデルはメーリング リストの生成には理想的なモデルとなり、コール・センターまたはマーケティング・アプリケーションを含む顧客関係管理で幅広く使用されています。

ディシジョン・リスト・モデルを含むストリームを実行するときに、ノードにより 3 つの新しいフィールドが追加されます。含まれたフィールドに対しては *1* (*Yes* の意味)、除外されたフィールドに対しては *$null* のどちらかのスコア、レコードがセグメントに収まる確立 (ヒット率)、セグメントの ID 番号の 3 フィールドです。新規フィールドの名前は予測された出力フィールドの名前から派生し、スコアには接頭辞の *$D-*、確率には *$DP-*、セグメント ID には *$DI-* が付けられます。

モデルは、構築時に指定された対象値に基づいてスコアリングされます。*$null* とスコアリングされるように、セグメントを手動で除外することができます。例えば、平均よりも低いヒット率のセグメントを検出するために低い確率の検索を行う場合、そのような平均より「低い」セグメントは、手動で除外しないかぎり *Yes* とスコアリングされます。必要に応じて、フィールド作成ノードまたは置換ノードを使用して、ヌルを *No* と記録することができます。

PMML

ディシジョン・リスト・モデルは、「最初にヒットした」選択基準の PMML RuleSetModel として、格納できます。ただし、ルールのすべてが同じスコアを持つように要求されます。対象フィールドまたは対象値を変更できるようにするには、最初のモデルに適合しないケースは 2 番目のモデルへ渡されるというようにして、複数のルールセット・モデルを適用される順序に従って 1 ファイル内へ格納できます。アルゴリズム名の *DecisionList* がこの標準以外の動作を示すために使用され、この名前の付いたルールセット・モデルのみがディシジョン・リスト・モデルとして認識され、そのようにスコアリングされます。

ディシジョン・リスト・モデル・ナゲットの設定

ディシジョン・リスト・モデル・ナゲットの「設定」タブを使用すると、傾向スコアの取得や、SQL 最適化の有効化または無効化が可能です。このタブは、モデル・ナゲットをストリームに追加した後に使用できます。

未調整傾向スコアを計算: フラグ型対象 (yes または no の予測を返す) を持つモデルの場合、対象フィールドに指定された真 (true) の結果の尤度を示す傾向スコアが必要な場合があります。また、スコアリング時に生成することができるその他の予測および確信度値があります。

調整済み傾向スコアを計算 **:** 未調整傾向スコアは、学習データにのみ基づき、このデータがオーバーフィットする多くのモデルの傾向によって過度に楽観的な場合があります。調整済み傾向は、テストまたは検証用データ区分に対してモデルのパフォーマンスを評価することによって補正しようとします。このオプションでは、モデルの生成前にデータ区分フィールドをストリーム内で定義し、調整済み傾向スコアがモデル作成ノードで有効化されている必要があります。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- ネイティブの **SQL** に変更してスコアリング**:** これを選択すると、データベース内でモデルをスコアリングするためのネイティブ SQL が生成されます。

  注**:** このオプションの方が短時間で結果を得ることができますが、モデルの複雑度が増大すると、ネイティブ SQL のサイズと複雑度も、それに応じて増大します。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

ディシジョン・リスト・ビューアー

使いやすい、タスク・ベースの ディシジョン・リスト・ビューアー のグラフィカル・インターフェースによって、モデル構築プロセスの複雑さが軽減され、ユーザーはデータ・マイニング手法に含まれる低レベルの細かい作業から解放されるとともに、分析の中でも、目標の設定、対象グループの選択、分析の結果、最適なモデルの選択といったユーザーの判断が必要な部分に注意を集中できます。

作業モデル領域

作業モデル領域には、現在のモデルと、その作業モデルに適用されるマイニング・タスクなどのアクションが表示されます。

ID: セグメントの順序を連番で識別します。モデル・セグメントは ID 番号に従って順番に計算されます。

セグメント・ルール **:**セグメントの名前と、セグメントに定義されている条件です。セグメント名はデフォルトではフィールド名か、条件で使用されるフィールド名をコンマで区切って連結したものです。

スコア **:**値を予測するフィールドです。この値は、他のフィールド (予測フィールド) の値と関連すると考えられます。

注 : 以下のオプションは、「 174 ページの『モデル指標の編成』」ダイアログで表示を切り替えることができます。

カバー **:**円グラフで、カバー全体に対する各セグメントの範囲を視覚的に分類します。

カバー **(n)** **:**カバー全体に対する各セグメントの範囲を一覧表示します。

度数: 各カバーに対して得られたヒット数を一覧表示します。例えば、カバーが 79 でフリクエンシが 50 であれば、選択したセグメントでは 79 のうち 50 が該当したことになります。

確率 **:**セグメントの確率を示します。例えば、カバーが 79 でフリクエンシが 50 であれば、そのセグメントの確率は 63.29% (50 を 79 で除算) になります。

エラー **:**セグメントのエラーを示します。

領域下部に表示される情報は、モデル全体でのカバー、フリクエンシ、確率です。

作業モデル ツールバー

作業モデル領域のツールバーでは、以下の機能を利用できます。

注 : 一部の機能は、モデル・セグメントを右クリックしても利用できます。

表 *9.* 作業モデル ツールバー・ボタン

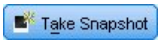
ツールバー・ボタン	説明
	新しいモデル・ナゲットを作成するためのオプションが用意された 「新規モデルを生成」ダイアログを起動します。
	インタラクティブ セッションの現在の状態を保存します。マイニング・タスク、モデルのスナップショット、データ選択、およびカスタム指標を含めて、ディシジョン・リスト・モデル作成ノードが現在の設定で更新されます。セッションをこの状態に復元するには、モデル作成ノードの「モデル」タブの 「保存済みセッション情報の使用」 ボックスをオンにし、「実行」 をクリックします。
	「モデル指標の編成」ダイアログ・ボックスが表示されます。 詳しくは、トピック 174 ページの『モデル指標の編成』を参照してください。
	「データ選択肢の編成」ダイアログ・ボックスが表示されます。 詳しくは、トピック 169 ページの『データ選択肢の編成』を参照してください。
	「スナップショット」タブを表示します。 詳しくは、トピック 165 ページの『「スナップショット」タブ』を参照してください。
	「代替」タブを表示します。 詳しくは、トピック 165 ページの『「代替」タブ』を参照してください。
	現在のモデル構造のスナップショットを取得します。スナップショットは「スナップショット」タブに表示され、モデルの比較のために共通で使用されます。
	新しいモデル・セグメントを作成するためのオプションが用意された 「セグメントの挿入」ダイアログを起動します。
	モデル・セグメントに条件を追加したり、モデル・セグメントで以前に定義した条件を変更するオプションが用意された 「セグメント・ルールの編集」ダイアログを起動します。

表 9. 作業モデル ツールバー・ボタン (続き)

	選択したセグメントをモデル階層で上方向に移動します。
	選択したセグメントをモデル階層で下方向に移動します。
	選択したセグメントを削除します。
	選択したセグメントをモデルに含めるかどうかを切り替えます。セグメントを除外すると、その結果は剰余に追加されます。これは、セグメントを再びアクティブできるという点で、セグメントの削除とは異なります。

「代替」タブ

「代替」タブには、「セグメントの検索」 をクリックすると生成される、作業モデル領域で選択したモデルまたはセグメントに対する代替のマイニング結果が、すべて一覧表示されます。

代替を作業モデルとするために、該当する代替を強調表示し、「ロード」 をクリックします。代替モデルが作業モデル領域に表示されます。

注 : ディシジョン・リスト・モデル作成ノードの「エキスパート」タブで「最大代替数」を設定して複数の代替を作成した場合にのみ、「代替」タブが表示されます。

生成された代替モデルでは、具体的なモデル情報が表示されます。

名前: 各代替は、順番にナンバリングされます。通常は、最初の代替モデルが持つ結果が最善の結果です。

目標: 次に例を示します。例えば、1 は真 (true) になります。

セグメント数 : 代替モデルで使用するセグメント・ルール数。

カバー : 代替モデルのカバー。

度数 :各カバーに関連するヒット数。

確率 :代替モデルの確率をパーセントで示します。

注 : 代替の結果はモデルと一緒には保存されません。これらの結果は、アクティブ・セッションの間のみ有効です。

「スナップショット」タブ

スナップショットは、特定の時点におけるモデルの姿です。例えば、作業モデル領域に別の代替モデルを読み込みたいが、現在のモデルは失いたくない場合に、スナップショットを作成できます。「スナップショット」タブには、作業モデルの状態を残すために手作業で作成したスナップショットがいくつでも、すべて一覧表示されます。

注 : スナップショットはモデルとともに保存されます。最初のモデルをロードする場合、スナップショットを採取することをお勧めします。このスナップショットにはオリジナルのモデル構造が保存されているため、モデルはいつでも元の状態に戻すことができます。生成されたスナップショットの名前は、いつ生成されたかを示すタイムスタンプとして表示されます。

モデルのスナップショットの作成

1. 作業モデル領域に表示する適切なモデル/代替モデルを選択します。
2. 作業モデルに必要な変更を加えます。
3. 「スナップショットの採取」をクリックします。「スナップショット」タブに新しいスナップショットが表示されます。

   「名前」。スナップショット名。スナップショット名をダブルクリックして、スナップショット名を変更することができます。

   目標: 次に例を示します。例えば、1 は真 (true) になります。

   セグメント数: モデルで使用するセグメント・ルール数。

   カバー: モデルのカバー。

   度数:各カバーに関連するヒット数。

   確率:モデルの確率をパーセントで示します。
4. スナップショットを作業モデルとするために、該当するスナップショットを強調表示し、「ロード」をクリックします。スナップショットが作業モデル領域に表示されます。
5. 「削除」 をクリックするか、スナップショットを右クリックしてメニューから 「削除」 を選択すると、スナップショットを削除できます。

ディシジョン・リスト・ビューアー の作業

顧客の応答や行動を最もよく予測できるモデルは、さまざまな段階を経て構築されます。ディシジョン・リスト・ビューアー が起動すると、定義されたモデル・セグメントおよび指標が作業モデルに読み込まれます。ユーザーはここからマイニング・タスクを開始し、必要に応じてセグメント/指標を修正して、新しいモデルやモデル作成ノードを生成できます。

ユーザーは、セグメント・ルールを 1 つまたは複数追加していくことで、満足できるモデルを構築します。セグメント・ルールは、マイニング・タスクを実行したり、「セグメント ルールの編集」 機能を使用してモデルに追加します。

モデル構築プロセスでは、指標データに対するモデルの検証、グラフによるモデルの視覚化、カスタムの Excel 指標の生成などを行って、モデルのパフォーマンスを評価できます。

モデルの品質が確認されたら、新しいモデルを生成し、IBM SPSS Modeler のキャンバスやモデル・パレットに配置できます。

マイニング・タスク

マイニング・タスクは、新しいルールをどのように生成するかを規定するパラメーターの集合です。これらのパラメーターの中には選択が可能なものもあり、ユーザーは新しい状況にモデルを柔軟に適合させることができます。タスクはタスク・テンプレート (種類)、対象、ビルド選択肢 (マイニング用のデータ・セット) で構成されています。

以下のセクションでは、さまざまなマイニング・タスク操作を説明します。

- 『マイニング・タスクの実行』
- 『マイニング・タスクの作成および編集』
- 169 ページの『データ選択肢の編成』

マイニング・タスクの実行**:** ディシジョン・リスト・ビューアー では、マイニング・タスクを実行したり、モデル間でセグメント・ルールをコピー/ペーストすることで、手作業でモデルにルールを追加できます。マイニング・タスクには、新しいセグメント・ルールの生成方法 (検索の戦略、ソースの属性、検索の幅、信頼水準などのデータ・マイニング・パラメーター設定)、予測する顧客行動、調査するデータに関する情報が含まれています。最善のセグメント・ルールを探し出すことがマイニング・タスクの目的です。

モデル・セグメント・ルールを生成するには、次の手順に従ってください。

1. 「剰余」行をクリックします。作業モデル領域にすでにセグメントが表示されている場合は、いずれかのセグメントを選択し、そのセグメントに基づく追加のルールを検索することもできます。剰余またはセグメントを選択した後、以下の方法のいずれかを使用してモデル、または代替モデルを生成します。

- 「ツール」メニューの 「セグメントの検索」 を選択します。
- 「剰余」 の行/セグメントを右クリックし、「セグメントの検索」 を選択します。
- 作業モデル領域の 「セグメントの検索」 ボタンをクリックします。

タスクの処理中は作業領域の下部に進捗状況が表示され、処理が終了するとユーザーにそれが通知されます。タスクの完了に要する正確な時間は、マイニング・タスクの複雑さやデータセットのサイズによって変わります。結果内にモデルが 1 つだけある場合、タスクが完了するとすぐにモデルが作業モデル領域に表示されます。ただし、結果に複数のモデル場ある場合、「代替」タブに表示されます。

注 : タスクの結果は、モデル付き、モデルなし、失敗のいずれかです。

新しいセグメント・ルールを探す操作は、モデルに追加するルールがなくなるまで繰り返します。これによって、有意な顧客グループがすべて見つかったことになります。

マイニング・タスクは、既存の任意のモデル・セグメントについて実行することもできます。タスクから思ったような結果が得られない場合は、同じセグメントに対して別のマイニング・タスクを実行することも可能です。こうすることで、選択したセグメントに基づいた、新規のルールが見つかります。各セグメントは先行するセグメントに依存するため、選択したセグメントよりも「下」にあるセグメント (選択したセグメントよりも後にモデルに追加されたセグメント) は、新しいセグメントに置き換えられます。

マイニング・タスクの作成および編集**:** マイニング・タスクは、データ・モデルを構成する一連のルールを検出するためのメカニズムです。選択したテンプレートで定義されている検索基準に加えて、タスクでは対象 (「ダイレクトメールに反応する顧客はどのくらいか」など、分析の動機となる実際の質問) も定義するほか、使用するデータ・セットも指定します。最善のモデルを探し出すことがマイニング・タスクの目的です。

マイニング・タスクの作成

マイニング・タスクを作成するには、次の手順に従ってください。

1. 新規のセグメント条件をマイニングするためのセグメントを選択します。
2. 「設定」をクリックします。「マイニング タスクの作成/編集」ダイアログが開きます。このダイアログには、マイニング・タスクを定義するオプションがあります。

3. 必要な変更を行い、「**OK**」 をクリックして作業モデル・ウィンドウに戻ります。ディシジョン・リスト・ビューアー は設定をデフォルトとして使用し、代替タスクまたは設定が選択されるまで各タスクに実行します。
4. 「セグメントの検索」 をクリックし、選択したセグメントでマイニング・タスクを開始します。

マイニング・タスクの編集

「マイニング タスクの作成/編集」ダイアログには、新しいマイニング・タスクの定義や、既存のマイニング・タスクの編集のためのオプションが用意されています。

マイニング・タスクで利用できるパラメーターの大半は、ディシジョン・リスト・ノードで提供されるものと同じです。例外が、以下に示されます。 詳しくは、トピック 160 ページの『ディシジョン・リストのモデル関連のオプション』を参照してください。

ロード設定 **:** 複数のマイニング・タスクを作成した場合、必要なタスクを選択します。

新規**...** 現在表示されている設定に基づいて、新しいマイニング・タスクを作成します。

対象

対象フィールド **:**値を予測するフィールドです。この値は、他のフィールド (予測フィールド) の値と関連すると考えられます。

対象値**:** モデル化する結果となる、対象フィールドの値を指定します。例えば、対象フィールドで顧客離れを「0 = no」および「1 = yes」とコーディングしている場合、どのレコードが他社に乗り換えそうかを表すルールを識別するには 1 と指定します。

シンプル設定

代替の最大数 **:** マイニング・タスクを実行して表示される代替の数を指定します。設定可能な最小値は 1 です。最大値はありません。

エキスパート設定

編集**... :** 「詳細パラメーターの編集」ダイアログが開き、詳細設定を定義できるようになります。詳しくは、トピック『詳細パラメーターの編集』を参照してください。

データ

条件抽出の構築**:** ディシジョン・リスト・ビューアー が新しいルールを検索するために分析する必要がある評価測定を指定するためのオプションが表示されます。一覧表示されている評価測定は、「データ選択肢の編成」ダイアログで作成/編集されます。

利用可能なフィールド**:** すべてのフィールドを表示するか、表示するフィールドを手作業で選択するためのオプションが用意されています。

編集**... :** 「ユーザー設定」オプションを選択した場合は、「利用可能フィールドのカスタマイズ」ダイアログが開き、マイニング・タスクがセグメント属性として利用できるフィールドを選択できます。詳しくは、トピック 169 ページの『利用可能フィールドのカスタマイズ』を参照してください。

詳細パラメーターの編集**:** 「詳細パラメーターの編集」ダイアログには、以下の構成オプションが用意されています。

データ分割手段: 連続フィールドを分割するために使用する方法です (等カウントまたは等幅)。

ビン数: 連続フィールドに対して作成するビンの数です。設定可能な最小値は 2 です。最大値はありません。

モデル検索幅。次のサイクルに使用できる、サイクルあたりのモデル結果の最大数です。設定可能な最小値は 1 です。最大値はありません。

ルール検索幅。次のサイクルに使用できる、サイクルあたりのルール結果の最大数です。設定可能な最小値は 1 です。最大値はありません。

ビンのマージ因子。セグメントを隣のセグメントと結合するときに、セグメントを拡張すべき最小量です。設定可能な最小値は 1.01 です。最大値はありません。

- 条件内の欠損値を許可: True の場合は、規則内で IS MISSING テストを使用することができます。
- 中間結果を破棄: True の場合は、検索プロセスの最終結果だけが返されます。最終結果は、検索プロセスでそれ以上調整されない結果です。False の場合は、中間結果も返されます。

利用可能フィールドのカスタマイズ: 「利用可能フィールドのカスタマイズ」ダイアログでは、マイニング・タスクがセグメント属性として利用できるフィールドを選択できます。

使用可能： 現在セグメント属性として利用できるフィールドが一覧表示されます。一覧からフィールドを削除するには、該当するフィールドを選択し、「削除 >>」をクリックします。選択されたフィールドは、「使用可能」リストから「使用不能」リストに移動します。

使用不能： セグメント属性として利用できないフィールドが一覧表示されます。フィールドを「使用可能」リストに追加するには、該当するフィールドを選択し、「<< 追加」をクリックします。選択されたフィールドは、「使用不能」リストから「使用可能」リストに移動します。

データ選択肢の編成: データ選択肢 (マイニング・データセット) を編成することで、ディシジョン・リスト・ビューアー が新しいルールを探すために分析すべき評価測定を指定し、測定の基準として使用するデータ選択肢を選択できます。

データ選択肢を編成するには、次の手順に従ってください。

1. 「ツール」メニューから 「データ選択肢の編成」 を選択するか、セグメントを右クリックして、このオプションを選択します。「データ選択肢の編成」ダイアログが開きます。

   注 :「データ選択肢の編成」ダイアログでは、既存のデータ選択肢を編集または削除することもできます。
2. 「新規データ選択の追加」 ボタンをクリックします。既存のテーブルに新しいデータ選択肢のエントリーが追加されます。
3. 「名前」 をクリックし、適切な選択肢の名前を入力します。
4. 「データ区分」 をクリックし、適切なデータ区分の種類を選択します。
5. 「条件」 をクリックし、適切な条件のオプションを選択します。「指定」 を選択すると、「選択条件を指定」ダイアログが開き、特定のフィールドの条件を定義するためのオプションが表示されます。
6. 適切な条件を定義し、「**OK**」 をクリックします。

データ選択肢は、「マイニング・タスクの作成/編集」ダイアログ・ボックスの「ビルド選択肢」ドロップダウン・リストから選択できます。このリストで、特定のマイニング・タスクに対して使用する評価測定を選択できます。

セグメント規則

タスク・テンプレートに基づいてマイニング・タスクを実行し、モデル・セグメント・ルールを検索します。セグメント挿入またはセグメント・ルール編集の機能を使用して、セグメント・ルールをモデルに手動で追加できます。

新しいセグメント・ルールを探してマイニングを行うと、「インタラクティブ・リスト」ダイアログの「ビューアー」タブに表示されます。「モデル・アルバム」から結果を 1 つ選択して 「読み込み」 をクリックして、モデルをすばやく調整できます。このようにさまざまな結果を試すことで、最適な対象グループを正確に表現しうるモデルを構築していきます。

セグメントの挿入**:** セグメント挿入の機能を使用して、セグメント・ルールをモデルに手動で追加できます。

セグメント・ルール条件を追加するには、次の手順に従ってください。

1. 「インタラクティブ・リスト」ダイアログで、新しいセグメントを追加するモデルの場所を選択します。新しいセグメントは、選択したセグメントのすぐ上に挿入されます。
2. 「編集」メニューから 「セグメントの挿入」 を選択するか、セグメントを右クリックしてこの項目を選択します。

   「セグメントの挿入」ダイアログが開き、セグメント・ルール条件を挿入できます。
3. 「挿入」をクリックします。「条件の挿入」ダイアログが開き、新しいルール条件に対する属性を定義できます。
4. ドロップダウン・リストからフィールドと演算子を選択します。

   注 :「**Not in**」演算子を選択すると、選択された条件は除外条件となり、「ルールの挿入」ダイアログで赤で表示されます。例えば、条件「`region = 'TOWN'`」が赤で表示されている場合、それは「`TOWN`」が結果セットから除外されることを意味します。
5. 1 つまたは複数の値を入力するか、あるいは 「値の挿入」 アイコンをクリックして「値の挿入」ダイアログを表示します。ダイアログで、選択したフィールドに定義する値を選択します。例えば、フィールドが「**married**」である場合は、「**YES**」と「**NO**」のオプションが提供されます。
6. 「**OK**」 をクリックして「セグメントの挿入」ダイアログに戻ります。もう一度 「**OK**」 をクリックすると、作成したセグメントがモデルに追加されます。

新しいセグメントは、指定されたモデルの場所に表示されます。

セグメント・ルールの編集**:** 「セグメント ルールの編集」機能では、セグメント・ルール条件を追加、変更、削除できます。

セグメント・ルール条件を変更するには、次の手順に従ってください。

1. 編集するモデル・セグメントを選択します。
2. 「編集」メニューから 「セグメント ルールの編集」 を選択するか、ルールを右クリックして、この項目を選択します。

   「セグメント ルールの編集」ダイアログが表示されます。
3. 適切な条件を選択し、「編集」 をクリックします。

   「条件の編集」ダイアログが開き、選択したルール条件に対する属性を定義できます。
4. ドロップダウン・リストからフィールドと演算子を選択します。

注：「**Not in**」演算子を選択すると、選択された条件は除外条件となり、「セグメント ルールの編集」ダイアログで赤で表示されます。例えば、条件「region = 'TOWN'」が赤で表示されている場合、それは「TOWN」が結果セットから除外されることを意味します。

5. 1 つまたは複数の値を入力するか、あるいは 「値の挿入」 ボタンをクリックして「値の挿入」ダイアログを表示します。ダイアログで、選択したフィールドに定義する値を選択します。例えば、フィールドが「**married**」である場合は、「**YES**」と「**NO**」のオプションが提供されます。
6. 「**OK**」 をクリックして「セグメント ルールの編集」ダイアログに戻ります。もう一度 「**OK**」 をクリックし、作業モデルに戻ります。

選択したセグメントは、更新されたルール条件で表示されます。

セグメント・ルール条件の削除*:* セグメント・ルール条件を削除するには、次の手順に従ってください。

1. 削除するルール条件が含まれているモデル・セグメントを選択します。
2. 「編集」メニューから 「セグメント ルールの編集」 を選択するか、セグメントを右クリックして、この項目を選択します。

   「セグメント ルールの編集」ダイアログが開き、1 つ以上のセグメント規則条件を削除することができます。
3. 適切なルール条件を選択し、「削除」 をクリックします。
4. 「**OK**」をクリックします。

セグメント・ルール条件を 1 つ以上削除すると、作業モデル領域の指標メトリックがリフレッシュされます。

セグメントのコピー**:** ディシジョン・リスト・ビューアー には、モデル・セグメントを手軽にコピーする方法が用意されています。セグメントをあるモデルから別のモデルに適用する場合は、そのセグメントを一方のモデルからコピー (または切り取り) し、別のモデルに貼り付けるだけで適用されます。代替プレビュー領域に表示されているモデルからセグメントをコピーし、作業モデル領域に表示されているモデルに貼り付けることもできます。これらの切り取り、コピー、貼り付け機能では、システムのクリップボードを使用して一次データを格納または取得しています。つまり、クリップボードに条件や対象がコピーされるということです。クリップボードの内容は、ディシジョン・リスト・ビューアー でしか使用できないわけではなく、他のアプリケーションにも貼り付けることができます。例えば、クリップ ボードの内容をテキスト・エディターに貼り付けると、条件と対象が XML 形式で貼り付けられます。

モデル・セグメントをコピーまたは切り取るには、次の手順に従ってください。

1. 他のモデルで使用したいモデル・セグメントを選択します。
2. 「編集」メニューから 「コピー」 (または 「切り取り」) を選択するか、モデル・セグメントを右クリックして 「コピー」 または 「切り取り」 を選択します。
3. 適切なモデル (モデル・セグメントの貼り付け先) を開きます。
4. いずれかのモデル・セグメントを選択し、「貼り付け」 をクリックします。

注：切り取り、コピー、貼り付けコマンドの代わりに、**Ctrl+X**、**Ctrl+C**、**Ctrl+V** のキー・コンビネーションを使用することもできます。

コピー (切り取り) されたセグメントは、選択されていたモデル・セグメントの上に挿入されます。貼り付けられたセグメント以下の指標は再計算されます。

注：この手順のモデルはどちらも同じモデル・テンプレートをベースにし、含まれている対象も同じでなければなりません。それ以外の場合はエラー・メッセージが表示されます。

代替モデル: 複数の結果がある場合、「代替」タブには各マイニング・タスクの結果が表示されます。各結果は、選択したデータにおいて対象と最も一致率が高い条件と、「十分なレベル」の代替の結果で構成されています。表示される代替モデルの総数は、分析プロセス中に使用される検索基準によって異なります。

代替モデルを表示するには、次の手順に従ってください。

1. 「代替」タブで、代替モデルをクリックします。代替プレビュー領域に代替モデルのセグメントが表示されるか、現在のモデル・セグメントと入れ替わります。
2. 作業モデル領域で代替モデル使用するには、代替プレビュー領域でモデルを選択し、「ロード」 をクリックするか、「代替」タブで代替モデル名を右クリックして 「ロード」 を右クリックします。

注 : 新しいモデルを生成するとき、代替モデルは保存されません。

モデルのカスタマイズ

データは静的なものではありません。顧客は転居、結婚、転職します。製品は市場フォーカスを失い、意味をなくします。

ディシジョン・リスト・ビューアー によって、ビジネス・ユーザーは新しい状況に、モデルを手早く、柔軟に適合させることができます。モデルの変更は、特定のモデル・セグメントの編集、優先順位付け、削除、非アクティブ化によって行います。

セグメントの優先順位付け: モデル ルールは任意の順序でランク付けできます。デフォルトではモデル・セグメントは優先度の順に表示され、最初のセグメントに最高の優先度が与えられています。1 つまたは複数のセグメントに異なる優先順位を与えると、モデルはそれに従って変化します。セグメントの優先順位を必要に応じて上下することで、モデルを変えることができます。

モデル・セグメントの優先順位付けを行うには、次の手順に従ってください。

1. 別の優先順位を割り当てるモデル・セグメントを選択します。
2. 作業モデル領域のツールバーで、2 つの矢印ボタンのいずれかをクリックし、選択したモデル・セグメントを、一覧内で上下に移動します。

優先順位付けを行うと、それまでの評価結果は再計算され、新しい値が表示されます。

セグメントの削除: 1 つまたは複数のセグメントを削除するには、次の手順に従ってください。

1. モデル・セグメントを選択します。
2. 「編集」メニューから 「セグメントの削除」 を選択するか、作業モデル領域のツールバーから、削除ボタンをクリックします。

修正するモデル用の指標は再計算され、それに従ってモデルが変化します。

セグメントの除外: 特定のグループについて検索を行うとき、ビジネス・アクションの基にするのは、一連の選択したモデル・セグメントであることがほとんどです。モデルの展開時には、モデル内のセグメントを除外することができます。除外されたセグメントのスコアはヌル値になります。セグメントを除外するということは、そのセグメントを使用しないということではありません。セグメントを除外すると、このルールに一致するすべてのレコードがメーリング リストから除外されます。ルールは適用されていますが、適用方法が違うということです。

特定のモデル・セグメントを除外するには、次の手順に従ってください。

1. 作業モデル領域からセグメントを選択します。

2. 作業モデル領域のツールバーから、「トグル・セグメントの除外」 ボタンをクリックします。選択したセグメントの「目標」列には、「除外」 と表示されます。

注 : セグメントの削除とは異なり、除外されたセグメントは最終的なモデルで再利用できます。除外されたセグメントは、グラフの結果に影響します。

対象値の変更**:** 「対象値の変更」ダイアログ・ボックスでは、現在の対象フィールドについて対象値を変更できます。

作業モデルと対象値が異なるスナップショット/セッションの結果があるテーブルの行は、背景色が黄色に変わります。これは、そのスナップショット/セッションの結果が古いことを示しています。

「マイニング・タスクの編集」 ダイアログには、現在の作業モデルの対象値が表示されます。対象値はマイニング・タスクと一緒には保存されません。代わりに、作業モデルの値から取得されます。

保存したモデルを、代替の結果やスナップショットのコピーを編集するなどして、現在の作業モデルとは対象値が異なる作業モデルにレベル上げすると、保存されているモデルの対象値が作業モデルと同じ値に変わります (作業モデル領域に表示される対象値は変わりません)。モデルのメトリックは新しい対象を使用して再評価されます。

新規モデルを生成

「新規モデルを生成」ダイアログには、モデルを命名し、新しいノードの作成場所を選択するための.オプションが用意されています。

モデル名**:** ストリーム・キャンバスに表示される、自動的に生成されたノード名を変更したり固有の名前を作成したりする場合は、「カスタム」を選択します。

ノードの生成先**:** 「キャンバス」を選択すると、作業キャンバスに新しいモデルが配置されます。「**GM** パレット」を選択すると、モデル・パレットに新しいモデルが配置されます。「両方」を選択すると、作業キャンバスとモデル・パレットの両方に新しいモデルが配置されます。

インタラクティブ セッション・ステートを含める **:**有効にすると、生成されるモデルにインタラクティブ セッション・ステートが組み込まれます。後からそのモデルを使用してモデル作成ノードを生成すると、そのステートが引き継がれ、インタラクティブ セッションの初期化に使用されます。このオプションが選択されているかどうかに関わらず、モデル自体は新しいデータを同じくスコアリングします。このオプションが選択されていなくてもモデルは構築ノードを作成できますが、それはより汎用的な構築ノードになり、前のセッションが停止した場所から開始するのではなく、インタラクティブ セッションを新しく開始するノードになります。ノードの設定を変更したもの、保存されたステートで実行すると、変更した設定は無視されて、保存されたステートが優先されます。

注 : モデルに付随するメトリックは標準メトリックのみです。その他のメトリックはインタラクティブ ステートに保存されます。保存されているマイニング・タスクのインタラクティブ ステートは、生成されるモデルには現れません。ディシジョン・リスト・ビューアー を起動すると、ビューアーを使用して指定された元の設定が表示されます。

詳しくは、トピック 50 ページの『モデル作成ノードの再生成』を参照してください。

モデルの評価

正常にモデル作成を行うには、運用環境で実装する前にモデルを慎重に評価する必要があります。ディシジョン・リスト・ビューアー では、実際にモデルに対する影響の評価に使用できる、多くの統計モデルおよ

びビジネス用のモデルが用意されています。これらにはゲイン・グラフや Excel との完全互換性も含まれており、展開の効果を評価する、コスト/利益に関わるシナリオのシミュレーションが可能になっています。

モデルは以下の方法で評価できます。

- ディシジョン・リスト・ビューアー にあらかじめ定義されている統計およびビジネス上のモデル指標を使用する (確率、度数)
- Microsoft Excel からインポートした指標を評価する
- ゲイン・グラフを使用してモデルを視覚化する

モデル指標の編成**:** ディシジョン・リスト・ビューアー には指標を定義するためのオプションが用意されており、それらは列として計算および表示されます。各セグメントには、デフォルトのカバー、フリクエンシ、確率、エラーの各指標を列の形で含めることができます。また、新しい指標を列として作成することもできます。

モデル指標の定義

モデルに指標を追加したり、既存の指標を定義したりするには、次の手順に従ってください。

1. 「ツール」メニューから 「モデル指標の編成」 を選択するか、モデルを右クリックして、この項目を選択します。「モデル指標の編成」ダイアログが開きます。
2. 「モデル指標の新規追加」 ボタン (「表示」列の右側) をクリックします。テーブルに新しい指標が表示されます。
3. 指標に名前を付け、適切な種類、表示オプション、選択肢を選択します。「表示」列は、作業モデルで指標を表示するかを示します。既存の指標を定義するときは、適切なメトリックと選択肢を選択し、その指標を作業モデルで表示するかを指定します。
4. 「**OK**」 をクリックし、ディシジョン・リスト・ビューアー の作業領域に戻ります。新しい指標の「表示」列がチェックされている場合は、作業モデルで新しい指標が表示されます。

Excel のカスタム メトリック

詳しくは、トピック『Excel での評価』を参照してください。

指標のリフレッシュ*:* 既存のモデルを新しい顧客に適用する場合など、状況によっては、モデル指標を再計算しなければなりません。

モデル指標を再計算するには、次の手順に従ってください。

「編集」メニューから、「すべての指標のリフレッシュ」 を選択します。

または

F5 キーを押します。

すべての指標が再計算され、作業モデルの新しい値が表示されます。

Excel での評価**:** ディシジョン・リスト・ビューアー は Microsoft Excel との統合が可能で、これによってユーザー独自の値の計算や利益の式をモデル構築プロセスの内部で直接使用し、コスト/利益のシナリオをシミュレートできます。Excel とのリンクによって、データを Excel にエクスポートし、そこでプレゼンテーション用のグラフを作成したり、複合利益や ROI 指標などのカスタム指標を計算する、あるいはモデルを構築しながらそれらを ディシジョン・リスト・ビューアー で表示するなどの作業ができます。

注 : Excel スプレッドシートで作業するには、CRM 分析を熟知したユーザーが、ディシジョン・リスト・ビューアー と Microsoft Excel を同期させるための構成情報を定義しなければなりません。この構成はディシジョン・リスト・ビューアー と Excel の間でどの情報を転送するかを表し Excel スプレッドシート・ファイルに格納されます。

以下の手順は MS Excel がインストールされている場合にのみ有効です。Excel がインストールされていない場合は、Excel とモデルを同期させるためのオプションは表示されません。

モデルを MS Excel と同期させるには、次の手順に従ってください。

1. モデルを開き、インタラクティブ セッションを実行し、ツール・メニューから 「モデル指標の編成」を選択します。
2. 「**Excel** 内のカスタム指標を計算」 オプションで 「はい」 を選択します。「ワークブック」フィールドがアクティブになり、あらかじめ構成されている Excel ワークブックのテンプレートを選択できます。
3. 「**Excel** への接続」 ボタンをクリックします。「開く」ダイアログが開き、ローカルまたはネットワークのファイル・システム上にある、構成済みのテンプレートの場所に移動できます。
4. 適切な Excel テンプレートを選択し、「開く」 をクリックします。選択された Excel テンプレートを起動するには、Windows タスクバーを使用して (または Alt + Tab キー)「カスタム指標のための入力の選択」ダイアログに再度移動します。
5. Excel テンプレートで定義されているメトリック名と、モデルのメトリック名との間で適切なマッピングを選択し、「**OK**」 をクリックします。

リンクが確立されると、構成済みの Excel テンプレートで Excel が起動し、モデルのルールがスプレッドシートに表示されます。Excel で計算された結果は、ディシジョン・リスト・ビューアー の新しい列に表示されます。

注 : モデルを保存しても Excel のメトリックは保存されません。それらのメトリックが有効になるのはアクティブ・セッション中のみです。ただし、Excel メトリックを含むスナップショットは作成できます。スナップショット・ビューに保存された Excel メトリックは履歴を比較する目的でのみ有効であり、開き直してもリフレッシュされません。詳しくは、トピック 165 ページの『「スナップショット」タブ』を参照してください。Excel メトリックは、Excel テンプレートとの接続を再確立してからでなければスナップショットに表示されません。

MS Excel の統合のセットアップ**:** ディシジョン・リスト・ビューアー と Microsoft Excel の間の統合は、あらかじめ構成されている Excel スプレッドシートのテンプレートを使用して行われます。このテンプレートは、以下の 3 つのワークシートで構成されています。

モデル指標 **:**インポートされた ディシジョン・リスト・ビューアー 指標、カスタム Excel 指標、計算の合計値 (「設定」 ワークシートで定義) を表示します。

設定 **:**インポートした ディシジョン・リスト・ビューアー 指標とカスタム Excel 指標に基づいた計算を生成する変数を提供します。

構成**:** ディシジョン・リスト・ビューアー からインポートする指標を指定したり、カスタム Excel 指標を定義するためのオプションを提供します。

警告 : 「構成」ワークシートの構造は厳格に定義されます。緑の領域のセルは編集してはいけません。

- モデルからのメトリック **:**計算で使用する ディシジョン・リスト・ビューアー メトリックを示します。

- モデルへのメトリック **:** Excel で生成されたメトリックのうち、ディシジョン・リスト・ビューアー に返すメトリックを示します。Excel 生成メトリックは ディシジョン・リスト・ビューアー の新しい指標の列に表示されます。

注 : 新しいモデルを生成すると Excel のメトリックはモデルから削除されます。それらのメトリックが有効になるのはアクティブ・セッション中のみです。

モデル指標の変更*:* 次の例でモデル指標の変更の方法を説明します。

- 既存の指標を変更。
- モデルから追加の標準指標をインポート。
- 追加のカスタム指標をモデルにエクスポート。

既存の指標を変更

1. テンプレートを開いて「構成」ワークシートを選択します。
2. 任意の 「名前」 または 「説明」 を、強調表示して上書きすることによって編集します。

   指標を変更する場合 (度数ではなく確率を指定するようユーザーに求める場合など) は、「モデルからのメトリック」で名前と説明を変更するだけです。そうすると、これがモデルに表示され、ユーザーは適切なマッピング先の指標を選択できます。

モデルから追加の標準指標をインポート

1. テンプレートを開いて「構成」ワークシートを選択します。
2. メニューから次の項目を選択します。

   「ツール」 > 「保護」 > 「シート保護の解除」
3. セル A5 を選択します。これは、黄色で塗りつぶされていて、「**End**」という文字が入力されています。
4. メニューから次の項目を選択します。

   「挿入」 > 「行」
5. 新しい指標の 「名前」 と 「説明」 を入力します。例えば、「**Error**」、「**Error associated with segment**」と入力します。
6. セル C5 に、式「**=COLUMN('Model Measures'!N3)**」を入力します。
7. セル D5 に、式「**=ROW('Model Measures'!N3)+1**」を入力します。

   これらの式を入力すると、現在空になっている「モデル指標」ワークシートの N 列に、新しい指標が表示されます。
8. メニューから次の項目を選択します。

   「ツール」 > 「保護」 > 「シートの保護」
9. 「**OK**」をクリックします。
10. 「モデル指標」ワークシートで、セル N3 に、新しい列のタイトルとして 「**Error**」が表示されていることを確認します。
11. N 列全体を選択します。
12. メニューから次の項目を選択します。

    「形式」 > 「セル」

13. デフォルトでは、数値カテゴリーはすべてのセルで 「標準」 になっています。「パーセンテージ」をクリックして数値の表示方法を変更します。こうすると、Excel で数値を確認しやすくなると同時に、グラフへの出力など、他の方法でデータを利用することができます。
14. 「**OK**」をクリックします。
15. 一意な名前にファイル拡張子 *.xlt* を付けて、スプレッドシートを Excel 2003 のテンプレートとして保存します。新しいテンプレートを見つけやすくするために、ローカルまたはネットワークのファイル・システム上の、事前に構成済みのテンプレート格納先に保存することをお勧めします。

追加のカスタム指標をモデルにエクスポート

1. 上記の例で「Error」列を追加したテンプレートを開き、「構成」ワークシートを選択します。
2. メニューから次の項目を選択します。

   「ツール」 > 「保護」 > 「シート保護の解除」
3. セル A14 を選択します。これは、黄色で塗りつぶされていて、「**End**」という文字が入力されています。
4. メニューから次の項目を選択します。

   「挿入」 > 「行」
5. 新しい指標の 「名前」 と 「説明」 を入力します。例えば、「**Scaled Error**」、「**Scaling applied to error from Excel**」と入力します。
6. セル C14 に、式「**=COLUMN('Model Measures'!O3)**」を入力します。
7. セル D14 に、式「**=ROW('Model Measures'!O3)+1**」を入力します。

   これらの式は、O 列がモデルに対して新しい指標を提供することを指定します。
8. 「設定」 ワークシートを選択します。
9. セル A17 に、説明として「**'- Scaled Error**」と入力します。
10. セル B17 に、調整値として「**10**」を入力します。
11. 「モデル指標」ワークシートで、「**Scaled Error**」を、セル O3 に新しい列のタイトルとして入力します。
12. セル O4 に、式「**=N4*Settings!B17**」を入力します。
13. セル O4 の角を選択し、セル O22 までドラッグして、式を各セルにコピーします。
14. メニューから次の項目を選択します。

   「ツール」 > 「保護」 > 「シートの保護」
15. 「**OK**」をクリックします。
16. 一意な名前にファイル拡張子 *.xlt* を付けて、スプレッドシートを Excel 2003 のテンプレートとして保存します。新しいテンプレートを見つけやすくするために、ローカルまたはネットワークのファイル・システム上の、事前に構成済みのテンプレート格納先に保存することをお勧めします。

このテンプレートを使用して Excel に接続すると、Error 値を新しいカスタム指標として使用できます。

モデルのビジュアル化

モデルの効果を理解する最善の方法は、それを視覚化することです。ゲイン・グラフを使用すると、業務に対する貴重な洞察を日単位で得られるうえ、複数の代替モデルをリアルタイムで検討することで、モデルに

とっても技術的なメリットがあります。『ゲイン・グラフ』セクションには、ランダム化意思決定におけるモデルのメリットが表示され、代替モデルがある場合は、複数のグラフを直接比較できます。

ゲイン・グラフ**:** ゲイン・グラフは、テーブルのゲイン % 列にある値を作図します。ゲインは、次の式を使用して、各増分のツリー中の全ヒット数に対する相対的な割合として定義されています。

(増加中のヒット数 / 全ヒット数) x 100%

ゲイン・グラフは、指定された割合のすべてのヒットをツリーから獲得するために、網をどれだけ広げたかをわかりやすく図示しています。対角線は、モデルを使用しない場合に、すべてのサンプルで期待される回答を作図したものです。この場合、1 人が別の人と全く同じように応答するため、回答割合は定数です。売り上げを 2 倍にするには、2 倍の人に質問する必要があります。曲線は、ゲインに基づいてより高位割合にランクされている人だけを含めることで、回答をどの程度、改善できるのかを示しています。例えば、上位の 50 % を含めると、70% を上回る肯定的な応答を網羅できます。カーブが急になるほど、ゲインも高くなります。

ゲイン・グラフを表示するには、次の手順に従ってください。

1. ディシジョン・リスト・ノードが含まれたストリームを開き、そのノードからインタラクティブ セッションを開始します。
2. 「ゲイン」タブをクリックします。指定するデータ区分により、1 つまたは 2 つ (例えばモデル指標に学習用とテスト用の両方のデータ区分が定義されている場合) のグラフ表示されます。

デフォルトでは、グラフはセグメントとして表示されます。「分位」 を選択し、ドロップダウン・メニューから適切な分位方法を選択すると、グラフの表示を分位ごとに切り替えることができます。

グラフ・オプション*:* 「グラフ・オプション」機能には、グラフ化するモデルとスナップショット、作図するデータ区分、セグメントのラベルの有無を選択するためのオプションが用意されています。

作図するモデル

現在のモデル**:** グラフ化するモデルを選択することができます。作業モデルのほか、作成されたスナップショット・モデルを選択できます。

作図するデータ区分

左側のグラフのデータ区分 **:**ドロップダウン・リストから、定義されているすべてのデータ区分を表示するか、またはすべてのデータを表示するかを選択できます。

右側のグラフのデータ区分 **:**ドロップダウン・リストから、定義されているすべてのデータ区分を表示するか、すべてのデータを表示するか、左側のグラフのみを表示するかを選択できます。「左側のみのグラフ」を選択している場合は、左側のグラフだけが表示されます。

セグメント・ラベルの表示 **:**有効にすると、各セグメントのラベルがグラフに表示されます。

第 10 章 統計モデル

統計モデルでは、数学の方程式を使用して、データから抽出した情報を符号化します。統計モデリング手法により、適切なモデルを非常に早く提供できます。柔軟性のあるマシン学習手法 (ニューラル・ネットワークなど) を使用すれば、より良い結果を出すことのできる問題でも、高度な手法の性能を判定するために統計モデルを基本予測モデルとして使用することができます。

以下の統計モデル作成ノードが利用できます。

線型モデルは、対象と 1 つまたは複数の予測値との線型の関係に基づいて連続型対象を予測します。

ロジスティック回帰は、入力フィールドの値に基づいてレコードを分類する統計手法です。線型と似ていますが、数値範囲ではなくカテゴリー対象フィールドを使用します。

因子分析モデル・ナゲットには、データの複雑性を整理する強力なデータ分解手法が 2 種類あります。主成分分析 (PCA) : 入力フィールドの線型結合が検出されます。成分が互いに直交する (直角に交わる) 場合に、フィールドのセット全体の分散を把握するのに役立ちます。因子分析 : 一連の観測フィールド内の相関パターンを説明する基本因子が識別されます。どちらの手法でも、元のフィールド・セットの情報を効果的に要約する少数の派生フィールドの検出が目標です。

判別分析によって、ロジスティック回帰より厳密な仮説を立てることができますが、これらの仮説が一致した場合、ロジスティック回帰分析に対する様々な代替あるいは補足になります。

一般化線型モデルは、指定したリンク関数によって従属変数が因子および共変量と線型関係になるよう、一般線型モデルを拡張したものです。さらにこのモデルでは、非正規分布の従属変数を使用することができます。 線型、ロジスティック回帰、カウント・データに関するログ線型モデル、そして区間打切り生存モデルなど、統計モデルの機能性が数多く含まれています。

一般化線型混合モデル (GLMN) は線型モデルを拡張したため、対象が非正規分布となる場合があり、指定されたリンク関数を介して因子および共変量に線形に関連し、観測が相関できるようになりました。一般化線型混合モデルには、単純な線型から、非正規分布の縦断的データを取り扱う複雑なマルチレベル・モデルまで、さまざまなモデルがあります。

Cox 線型回帰ノードを使用すると、打ち切りレコードの存在下でイベントまでの時間のデータの生存モデルを構築します。モデルは、対象のイベントが入力変数の指定の値で指定の時間 (*t*) に発生する確率を予測する生存関数を作成します。

線型ノード

線型は、数値型入力フィールドの値に基づいてレコードを分類する一般的な統計手法です。線型は、予測された出力値と実際の出力値の違いを最小限にする直線または面に適合します。

要件: 線型モデルでは、数値型フィールドだけを使用できます。正確に、1 つの対象フィールド (役割を出力に設定) と 1 つ以上の予測フィールド (役割を入力に設定) を指定する必要があります。役割が「両方」または「なし」のフィールドは、非数値型フィールドのため、無視されます。(必要な場合、非数値型フィールドはフィールド作成ノードを使用して再コード化できます。)

利点: 線型モデルは比較的単純で、予測の生成のために解釈しやすい数式が取得できます。線型は、古くから確立されている統計手法なので、モデルのさまざまな特徴が確認されています。また、一般に線型モデルの学習速度は非常に高速です。線型ノードでは、自動フィールド選択を利用して、式から重要 (有意) でない入力フィールドを削除することができます。

注: 対象フィールドが連続した範囲でなく、「はい/いいえ」または「チャーンする/チャーンしない」のようなカテゴリー型の場合、ロジスティック回帰を代わりに使用できます。ロジスティック回帰でも、これらのフィールドを再コード化する必要性を排除して、文字列入力がサポートされます。 詳しくは、トピック 191 ページの『ロジスティック回帰ノード』を参照してください。

線型モデル

線型モデルは、対象と 1 つ以上の予測変数との線型の関係に基づいて連続型対象を予測します。

線型モデルは比較的単純で、スコアリング用に、解釈が容易な数式を提供します。これらのモデルのプロパティーについてはよく理解されていて、同じデータ・セットの他のモデル・タイプ (ニューラル・ネットワークまたはディシジョン・ツリーなど) に比べて、通常、非常に素早く作成できます。

例: 住宅所有者の保険金請求の調査を行うにはリソースが限られている保険会社が、請求のコストを推定するためのモデルを作成したいと考えます。このモデルをサービス・センターに提供することによって、担当者は顧客との電話中に請求情報を入力し、過去のデータに基づいて「予測される」請求のコストをすぐに計算することができます。

フィールド要件: 対象と 1 つ以上の入力が必要です。デフォルトでは、事前定義された役割が「両方」または「なし」のフィールドは使用されません。対象は連続型 (スケール) でなければなりません。予測変数 (入力) には測定レベルの制限はありません。カテゴリー・フィールド (フラグ型、名義型、および順序型) はこのモデルで因子として使用され、連続型フィールドは共変量として使用されます。

目的

実行する作業

- 新しいモデルを作成: 完全に新しいモデルを作成します。これはノードの通常の操作です。
- 既存モデルの学習を継続: ノードによって正常に作成された最後のモデルで学習が継続します。これにより、元のデータにアクセスすることなく既存のモデルを更新またはリフレッシュできます。また、新

規レコードまたは更新されたレコードのみがストリームに適用されるため、パフォーマンスが大幅に向上する場合があります。以前のモデルの詳細はモデル作成ノードで保存されるので、以前のモデル・ナゲットがストリームまたは「モデル」パレットでもう使用できない場合でも、このオプションを使用することができます。

注**:** このオプションが有効な場合、「フィールド」タブと「作成オプション」タブにある他のすべてのコントロールが無効になります。

主な目的は**?**: 該当する目的を選択します。

- 標準モデルを作成: この方法では、予測変数を使用して対象を予測する単一モデルが作成されます。一般的に、ブースティング、バギング、または大規模なデータ・セット・アンサンブルと比べ、標準モデルは解釈が容易であり、素早くスコアリングできます。

  注**:** 分割モデルの場合、「既存モデルの学習を継続」とともにこのオプションを使用するには、Analytic Server に接続されている必要があります。
- モデル精度の強化 **(**ブースティング**):** ブースティングを使用して、アンサンブルモデルを構築します。この方法では、ブースティングを使用してアンサンブル・モデルが作成されます。これによって、より正確な予測を得るための一連のモデルが生成されます。アンサンブルは、標準モデルと比べて作成とスコアリングに時間がかかる場合があります。

  ブースティングによって一連の「コンポーネント・モデル」が生成されます。各コンポーネント・モデルはデータ・セット全体に作成されます。連続する各コンポーネント・モデルを作成する前に、以前のコンポーネント・モデルの残差に基づきレコードに重みが付けられます。残差が大きいケースには比較的大きな分析の重みが与えられるため、次のコンポーネント・モデルは、これらのレコードの予測に重点を置きます。これらのコンポーネント・モデルがまとまってアンサンブル・モデルを形成します。アンサンブル・モデルは結合ルールを使用して新しいレコードをスコアリングします。使用できる規則は、対象の測定レベルによって異なります。
- モデルの安定性を拡張 **(**バグ**):** バギング (ブートストラップ集計) を使用して、アンサンブルモデルを構築します。この方法では、バギング (ブートストラップ集計) を使用してアンサンブル・モデルが作成されます。これによって、より信頼性の高い予測を得るための複数のモデルが生成されます。アンサンブルは、標準モデルと比べて作成とスコアリングに時間がかかる場合があります。

  ブートストラップ集計 (バギング) は、元のデータ・セットから置換を行うサンプリングによって、学習データ・セットの複製を作成します。これにより、元のデータ・セットと同じサイズのブートストラップ・サンプルが作成されます。その後、「コンポーネント・モデル」が複製ごとに作成されます。これらのコンポーネント・モデルがまとまってアンサンブル・モデルを形成します。アンサンブル・モデルは結合ルールを使用して新しいレコードをスコアリングします。使用できる規則は、対象の測定レベルによって異なります。
- 非常に大きなデータ セットのモデルを作成**:** この方法では、データ・セットを別々のデータ・ブロックに分割することにより、アンサンブル・モデルが作成されます。上記のモデルのいずれかを作成するにはデータ・セットが大きすぎる場合、または増分モデル作成の場合、このオプションを選択します。このオプションは、作成には時間がかからないものの、標準モデルと比べてスコアリングに時間がかかる場合があります。

ブースティング、バギング、および非常に大きいデータ・セットに関する設定については、 183 ページの『アンサンブル』を参照してください。

基本

自動的にデータを準備する: このオプションでは、対象と予測変数を内部で変換してモデルの予測精度を最大化する手続きが可能になります。いずれの変換もモデルに保存され、スコアリングする新しいデータに適用されます。変換フィールドの元のバージョンはモデルから除外されます。デフォルトでは、次のデータの自動準備が実行されます。

- 日付および時刻の処理: 日付の各予測変数は、基準日 (1970-01-01) 以降の経過時間を含む新たな連続型予測値に変換されます。時刻の各予測変数は、基準時刻 (00:00:00) 以降の経過時間を含む連続型予測値に変換されます。
- 測定レベルの調整: 異なる値が 5 個より少ない連続型予測値は、順序型予測値に変更されます。異なる値が 10 個より多い順序型予測値は、連続型予測値に変更されます。
- 外れ値の処理: カットオフ値 (平均値からの標準偏差が 3) を超える連続型予測値の値がカットオフ値に設定されます。
- 欠損値の処理: 名義型予測値の欠損値は、学習データ区分の最頻値と置き換えられます。順序型予測値の欠損値は、学習データ区分の中央値と置き換えられます。連続型予測値の欠損値は、学習データ区分の平均値と置き換えられます。
- 監視結合: 対象と関連して処理するフィールドの数を減らすことにより、より節約的なモデルを作成します。類似するカテゴリーが、入力と対象との関係に基づいて特定されます。有意差のないカテゴリー (p 値が 0.1 より大きいカテゴリー) は結合されます。すべてのカテゴリーが 1 つに結合される場合、元のバージョンおよび派生したバージョンのフィールドは、予測変数としての値がないため、モデルから除外されます。

信頼度レベル: 係数ビューでモデル係数の区間の推定値を計算するために使用する信頼度のレベルです。 0 より大きく 100 より小さい値を指定してください。デフォルトは 95 です。

モデルの選択

モデル選択方法: モデル選択方法 (下記参照) のいずれかを選択するか、または主効果のモデル項として使用可能なすべての予測値を単に入力する「すべての予測値を含む」を選択します。デフォルトでは、「変数増加ステップワイズ法」が使用されます。

変数増加ステップワイズ法の選択: モデルの効果がない状態から、これ以上追加または削除できなくなるまで、ステップワイズ法の基準に従って徐々に効果を追加および削除します。

- 投入または除去の基準: これは、モデルに効果を加えるかどうか、またはモデルから効果を削除するかどうかを決定するときに使用する統計です。「情報量基準 **(AICC)**」はモデルを指定された学習セットの尤度に基づき、過度に複雑なモデルにペナルティーを科すよう調整されます。「**F** 統計量」はモデルのエラーの改善に対する統計検定に基づいています。「調整済み **R2** 乗」は学習セットの適合度に基づき、過度に複雑なモデルにペナルティーを科すよう調整されます。「オーバーフィット防止基準 **(ASE)**」は、オーバーフィット防止セットの適合度 (平均平方誤差 (ASE)) に基づきます。オーバーフィット防止セットは、モデルの学習に使用されない元のデータ・セットのおよそ 30% の無作為サブサンプルです。

  「**F** 統計量」以外の基準を選択した場合、各ステップでその基準での最も大きい正の増分に対応する効果がモデルに追加されます。その基準での減少に対応するモデルの効果はいずれも削除されます。

  「**F** 統計量」を基準として選択すると、「次の値より小さい **p** 値の効果を含む」で指定したしきい値よりも小さい、最小の *p* 値を持つ効果が、各ステップでモデルに追加されます。デフォルトは 0.05 です。「次の値より大きい **p** 値の効果を削除する」で指定したしきい値より大きい *p* 値を持つモデルの効果はいずれも削除されます。デフォルトは 0.10 です。

- 最終モデルの最大効果数をカスタマイズする: デフォルトでは、すべての使用可能な効果をモデルに投入できます。あるいは、ステップワイズ・アルゴリズムが指定した効果の最大数でステップを終了する場合は、アルゴリズムは効果の現在のセットで停止します。
- ステップの最大数をカスタマイズする: 特定のステップ数の後、ステップワイズ・アルゴリズムが停止します。デフォルトでは、これは使用可能な効果の数の 3 倍です。あるいは、ステップの最大数を正整数で指定します。

最適サブセットの選択: 「可能なすべての」モデル、または少なくとも変数増加ステップワイズ法より大きい、可能なモデルのサブセットをチェックし、最適サブセットの基準に従って最適サブセットを選択します。「情報量基準 **(AICC)**」はモデルを指定された学習セットの尤度に基づき、過度に複雑なモデルにペナルティーを科すよう調整されます。「調整済み **R2** 乗」は学習セットの適合度に基づき、過度に複雑なモデルにペナルティーを科すよう調整されます。「オーバーフィット防止基準 **(ASE)**」は、オーバーフィット防止セットの適合度 (平均平方誤差 (ASE)) に基づきます。オーバーフィット防止セットは、モデルの学習に使用されない元のデータ・セットのおよそ 30% の無作為サブサンプルです。

最大の基準値を持つモデルが最良のモデルとして選択されます。

注**:** 最適サブセットによる選択は、変数増加ステップワイズ法による選択に比べてより多くの計算リソースを使用します。最適サブセットが、ブースティング、バギング、または非常に大きいデータ・セットと組み合わせて実行されると、変数増加ステップワイズ法の選択を使用して作成された標準モデルよりも大幅に時間がかかる場合があります。

アンサンブル

これらの設定は、「目的」で、ブースティング、バギング、または非常に大きなデータ・セットが要求される場合に起きるアンサンブルの動作を決定します。選択された目的に適用されないオプションは無視されます。

バギングおよび非常に大きなデータ・セット: アンサンブルをスコアリングする際、この規則を使い、基本モデルの予測値を結合して、アンサンブル・スコア値を計算します。

- 連続型対象のデフォルトの結合規則: 連続型対象に対するアンサンブル予測値は、基本モデルの予測値の平均値または中央値を使用して結合できます。

モデルの精度を上げることが目的である場合、結合規則の選択が無視されることに注意してください。ブースティングは常に、カテゴリー対象のスコアリングには重み付き多数決を使用し、連続型対象のスコアリングには重み付き中央値を使用します。

ブースティングおよびバギング: モデルの精度または安定性の向上が目的である場合、作成する基本モデル数を指定します。バギングの場合、これはブートストラップ・サンプルの数になります。これは正整数である必要があります。

詳細設定

結果の再現: ランダム・シードを設定すると、分析を再現することができます。どのレコードをオーバーフィット防止セットに含むかを選択するには、乱数ジェネレーターを使用します。整数を指定するか、「生成」をクリックします。「生成」をクリックすると、1 から 2147483647 までの整数の疑似乱数が作成されます。デフォルトは 54752075 です。

モデル・オプション

モデル名: 対象フィールドに基づいて自動的にモデル名を生成するか、またはカスタム名を指定できます。自動的に生成される名前は、対象フィールド名です。

モデルがスコアリングされると、必ず予測値が計算されることに注意してください。新規フィールド名は、接頭辞として *$L-* が付いた対象フィールド名です。例えば、対象フィールド名が *sales* の場合、新規フィールド名は *$L-sales* になります。

モデルの要約

「モデルの要約」ビューでは、モデルとその適合度についてスナップショット的に一目でわかる要約が表示されます。

表　テーブルは次のようなハイレベルなモデル設定を特定します。

- 「フィールド」タブで指定した対象名。
- 「基本」設定で指定したとおりに、データの自動準備が実行されたか。
- 「モデル選択」設定で指定したモデルの選択方法および選択基準。最終モデルの選択基準の値も表示され、値が小さいほど適切である形式で提示されます。

グラフ

このグラフには、最終モデルの精度が表示されます (値が大きいほど精度が高いことを示します)。値は、100 × 最終モデルの調整済み R^2 乗です。

データの自動準備

このビューには、データの自動準備 (ADP) ステップで、どのフィールドが除外されたか、また変換されたフィールドがどのように派生したかについての情報が表示されます。テーブルには変換または除外されたフィールドごとに、フィールド名、分析での役割、および ADP ステップで実行されたアクションがリストされます。フィールドは、フィールド名のアルファベット順 (昇順) にソートされます。各フィールドに対して行うことのできるアクションには、以下のようなものがあります。

- 「期間の取得**:** 月」では、日付を含むフィールドの値から現在のシステムの日付までの経過時間 (月) が計算されます。
- 「期間の取得**:** 時間」では、時間を含むフィールドの値から現在のシステムの時間までの経過時間 (時間) が計算されます。
- 「測定の尺度を連続型から順序型に変更」では、固有値が 5 個未満の連続型フィールドが順序型フィールドに変更されます。
- 「測定の尺度を順序型から連続型に変更」では、固有値が 10 個を超える順序型フィールドが連続型フィールドに変更されます。
- 「外れ値を削除」では、カットオフ値 (平均値からの標準偏差が 3) を超える連続型予測フィールドの値がカットオフ値に設定されます。
- 「欠損値の置換」では、名義型の欠損値が最頻値に、順序型フィールドの欠損値が中央値に、連続型フィールドの欠損値が平均値に置き換えられます。
- 「カテゴリを結合して目標との関連性を最大化」では、入力と目標との関係に基づいて「類似する」予測カテゴリーが特定されます。有意差のないカテゴリー (*p* 値が 0.05 より大きいカテゴリー) が結合されます。
- 「一定の予測値を**/**外れ値の処理後**/**カテゴリーの結合後除外する」では、1 つの値を持つ予測変数が除外されます。これらは、他の ADP アクションが実行された後に行われる可能性があります。

予測値の重要度

一般にモデリングの作業では、最も重要な予測フィールドの編集に集中して取り組み、最も重要でない予測フィールドは削除するか無視してしまいたくなるものです。予測値の重要度グラフを使用すると、モデル推定時に各予測値の相対重要度が示されるので便利です。値が相対的であるため、表示されるすべての予測変

数の値の合計は 1.0 となります。予測値の重要度は、モデルの精度に関連しません。予測が正確かどうかに関係なく、予測時における各予測の重要度にのみ関連します。

予測対観測

縦軸の予測値に対し横軸に観測値を示した分割散布図が表示されます。点は 45 度の線上にあるのが理想です。このビューで、モデルによるレコードの予測にとりわけ問題があるかがわかります。

残差

モデル残差の診断グラフが表示されます。

グラフ・スタイル: 各種表示スタイルがあり、「スタイル」ドロップダウンから選択できます。

- ヒストグラム: これは、正規分布のオーバーレイが適用された、スチューデント化残差の分割ヒストグラムです。線型モデルでは残差は正規分布になる想定するため、ヒストグラムは理想的にはほぼ滑らかな線になるはずです。
- 正規 **P-P** プロット: これは、スチューデント化残差を正規分布と比較する、分割された確率 - 確率プロットです。作図された点の傾きが通常の線に比べて小さい場合、残差は正規分布より大きい変動を示し、傾きがより大きい場合、残差は正規分布より小さい変動を示します。作図された点が S 字曲線を示す場合、残差の分布は歪んでいます。

外れ値

このテーブルではモデルに悪影響を与えるレコードがリストされ、レコード ID (「フィールド」タブで指定している場合)、対象値、および Cook の距離が表示されます。Cook の距離は、特定のレコードがモデル係数の計算から除外された場合に、すべてのレコードの残差がどのくらい変化するかを示す測度です。Cook の距離が大きい場合、レコードを除外すると係数が大幅に変わることを示しているため、影響があると考えられます。

影響を及ぼすレコードは慎重に検証し、モデルの推定時に重み付けを小さくするか、外れ値を許容可能なしきい値に切り詰めるか、あるいは影響のあるレコードを完全に削除するかを判断する必要があります。

効果

このビューには、モデルの各効果のサイズが表示されます。

スタイル: さまざまな表示スタイルがあり、「スタイル」ドロップダウン・リストから選択できます。

- ダイアグラム: この図表では、効果が予測値の重要度の大きい順にソートされます。ダイアグラム内の接続線は、効果の有意性に基づいて重みが付けられます。効果の有意性が大きいほど (*p* 値が小さいほど) 線が太くなります。接続線の上にマウス・ポインターを置くと、*p* 値および効果の重要度を示すツールチップが表示されます。これはデフォルトです。
- テーブル: これは、モデル全体の効果および個別のモデル効果を示す分散分析テーブルです。個別の効果は、予測値の重要度が大きい順にソートされます。ただしデフォルトでは、テーブルが省略表示され、モデル全体の結果だけが表示されます。個別のモデル効果の結果を表示するには、テーブル内の「修正モデル」セルをクリックします。

予測値の重要度:「予測値の重要度」スライダーは、どの予測値がビュー内に表示されるかを制御します。このスライダーを使用してもモデルは変更されませんが、最も重要な予測値にフォーカスすることができます。デフォルトでは、上位 10 件の効果が表示されます。

有意確率:「有意確率」スライダーは、予測値の重要度に基づく表示のほか、さらにどの効果がビュー内に表示されるかを制御します。有意確率の値がスライダーの値より大きい効果は表示されません。このスライ

ダーを使用してもモデルは変更されませんが、最も重要な効果にフォーカスすることができます。デフォルトでは値が 1.00 になるため、有意確率に基づいてフィルタリングされる効果はありません。

係数

このビューには、モデルの各係数の値が表示されます。因子 (カテゴリー予測値) はモデル内で指標コード化されるため、因子を含む効果には通常、複数の関連する係数があることに注意してください。冗長 (参照) パラメーターに対応するカテゴリー以外は、カテゴリーごとに 1 つの係数があります。

スタイル: さまざまな表示スタイルがあり、「スタイル」ドロップダウン・リストから選択できます。

- ダイアグラム: この図表では、まず切片項が表示されてから、すべての効果が予測値の重要度が大きい順にソートされます。因子を含む効果内で、係数はデータ値の昇順でソートされます。ダイアグラム内の接続線は、係数の符号 (図のキーを参照) に基づいて色分けされ、係数の有意性に基づいて重みが付けられます。係数の有意性が大きいほど (*p* 値が小さいほど) 線が太くなります。接続線上にマウス・ポインターを置くと、係数の値、その *p* 値、およびパラメーターが関連する効果の重要度を示すツールチップが表示されます。これはデフォルトのスタイルです。
- テーブル: 各モデル係数の値、有意差検定、および信頼区間が表示されます。切片項の後、効果は予測値の重要度の降順でソートされます。因子を含む効果内で、係数はデータ値の昇順でソートされます。ただしデフォルトではテーブルが省略表示され、各モデル・パラメーターの係数、有意性、および重要度だけが表示されます。標準誤差、*T* 統計量、および信頼区間を表示するには、テーブルの「係数」セルをクリックします。テーブル内のモデル・パラメーターの名前上にマウス・ポインターを置くと、パラメーターの名前、パラメーターが関連する効果、および (カテゴリー予測値の場合は) モデル・パラメーターに関連する値のラベルを示すツールチップがテーブルに表示されます。これは特に、データの自動準備がカテゴリー予測値の類似カテゴリーを結合する時に作成された新しいカテゴリーを確認する場合に役立ちます。

予測値の重要度:「予測値の重要度」スライダーは、どの予測値がビュー内に表示されるかを制御します。このスライダーを使用してもモデルは変更されませんが、最も重要な予測値にフォーカスすることができます。デフォルトでは、上位 10 件の効果が表示されます。

有意確率:「有意確率」スライダーは、予測値の重要度に基づく表示のほか、さらにどの係数がビュー内に表示されるかを制御します。有意確率の値がスライダーの値より大きい係数は表示されません。このスライダーを使用してもモデルは変更されませんが、最も重要な係数に焦点を当てることができます。デフォルトでは値が 1.00 になるため、有意確率に基づいてフィルタリングされる係数はありません。

推定平均

これは有意な予測変数について表示されるグラフです。横軸の予測変数の各値に対して、縦軸に対象のモデル推定値を表示し、その他のすべての予測変数は一定に保ちます。対象に対する各予測変数の係数の効果が視覚化されるので有用です。

注**:** 有意な予測変数がない場合、推定平均値は生成されません。

モデル構築の要約

「モデル選択」設定で、「なし」以外のモデル選択アルゴリズムを選択すると、モデル作成プロセスの詳細が一部表示されます。

「変数増加ステップワイズ法」。: 変数増加ステップワイズ法が選択アルゴリズムである場合、テーブルにはステップワイズ・アルゴリズムの最後の 10 ステップが表示されます。ステップごとに、選択基準の値と

そのステップでのモデル内の効果が表示されます。これにより、各ステップのモデルへの寄与度を把握できます。各列で行を並べ替えることができるので、任意のステップでモデル内にどの効果があるかをより簡単に確認できます。

最適サブセット: 最適サブセットが選択アルゴリズムである場合、テーブルには上位 10 件のモデルが表示されます。モデルごとに、選択基準の値とモデル内の効果が表示されます。これにより、上位モデルの安定性が把握できます。相違点が少ない類似した効果が多い傾向があれば、「上位」モデルはかなり信頼できますが、非常に異なる効果を持つ傾向がある場合は、一部の効果が非常に類似している可能性があり、結合 (あるいは 1 つを削除) する必要があります。各列で行を並べ替えることができるので、任意のステップでモデル内にどの効果があるかをより簡単に確認できます。

設定

モデルがスコアリングされると、必ず予測値が計算されることに注意してください。新規フィールド名は、接頭辞として *$L-* が付いた対象フィールド名です。例えば、対象フィールド名が *sales* の場合、新規フィールド名は *$L-sales* になります。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- ネイティブの **SQL** に変更してスコアリング**:** これを選択すると、データベース内でモデルをスコアリングするためのネイティブ SQL が生成されます。

  注**:** このオプションの方が短時間で結果を得ることができますが、モデルの複雑度が増大すると、ネイティブ SQL のサイズと複雑度も、それに応じて増大します。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

Linear-AS ノード

IBM SPSS Modeler には、次に示す 2 つの異なるバージョンの線型ノードがあります。

- 線型**:** IBM SPSS Modeler Server 上で実行される従来のノードです。
- **Linear-AS**は、IBM SPSS Analytic Server に接続しているときに実行できます。

線型は、数値型入力フィールドの値に基づいてレコードを分類する一般的な統計手法です。線型は、予測された出力値と実際の出力値の違いを最小限にする直線または面に適合します。

要件**:** 線型モデルでは、数値型フィールドおよびカテゴリー型予測フィールドだけを使用できます。正確に、1 つの対象フィールド (役割を出力に設定) と 1 つ以上の予測フィールド (役割を入力に設定) を指定する必要があります。役割が「両方」または「なし」のフィールドは、非数値型フィールドのため、無視されます。(必要な場合、非数値型フィールドはフィールド作成ノードを使用して再コード化できます。)

利点**:** 線型モデルは比較的単純で、予測の生成のために解釈しやすい数式が取得できます。線型は、古くから確立されている統計手法なので、モデルのさまざまな特徴が確認されています。また、一般に線型モデルの学習速度は非常に高速です。線型ノードでは、自動フィールド選択を利用して、式から重要 (有意) でない入力フィールドを削除することができます。

注**:** 対象フィールドが連続した範囲でなく、「はい/いいえ」または「チャーンする/チャーンしない」のようなカテゴリー型の場合、ロジスティック回帰を代わりに使用できます。ロジスティック回帰でも、これらのフィールドを再コード化する必要性を排除して、文字列入力がサポートされます。 詳しくは、トピック 191 ページの『ロジスティック回帰ノード』を参照してください。

Linear-AS モデル

線型モデルは、対象と 1 つ以上の予測変数との線型の関係に基づいて連続型対象を予測します。

線型モデルは比較的単純で、スコアリング用に、解釈が容易な数式を提供します。これらのモデルのプロパティーについてはよく理解されていて、同じデータ・セットの他のモデル・タイプ (ニューラル・ネットワークまたはディシジョン・ツリーなど) に比べて、通常、非常に素早く作成できます。

例: 住宅所有者の保険金請求の調査を行うにはリソースが限られている保険会社が、請求のコストを推定するためのモデルを作成したいと考えます。このモデルをサービス・センターに提供することによって、担当者は顧客との電話中に請求情報を入力し、過去のデータに基づいて「予測される」請求のコストをすぐに計算することができます。

フィールド要件: 対象と 1 つ以上の入力が必要です。デフォルトでは、事前定義された役割が「両方」または「なし」のフィールドは使用されません。対象は連続型 (スケール) でなければなりません。予測変数 (入力) には測定レベルの制限はありません。カテゴリー・フィールド (フラグ型、名義型、および順序型) はこのモデルで因子として使用され、連続型フィールドは共変量として使用されます。

基本

定数項を含める このオプションは、x 軸が 0 であるときの y 軸上のオフセットを含めます。通常、切片はモデルに含まれます。データが原点を通ると仮定できる場合は、切片を除外できます。

2 要因の交互作用を考慮**:** このオプションは、モデルに対して、入力の可能な各ペアを比較して、一方の傾向が他方に影響するかどうかを確認するように指示します。影響する場合は、それらの入力が計画行列に含められる可能性が高くなります。

係数の推定値の信頼区間 **(%):** これは、係数ビューでモデル係数の推定値を計算するために使用する信頼区間です。0 より大きく 100 より小さい値を指定してください。デフォルトは 95 です。

カテゴリー予測のソート順**:** これらのコントロールは、「最後の」カテゴリーを決定するために、因子 (カテゴリー入力) に対するカテゴリーの順序を決定します。入力がカテゴリー型でない場合、またはカスタム参照カテゴリーが指定されている場合、ソート順設定は無視されます。

モデルの選択

モデル選択方法: モデル選択方法 (下記参照) のいずれかを選択するか、または主効果のモデル項として使用可能なすべての予測値を単に入力する「すべての予測値を含む」を選択します。デフォルトでは、「変数増加ステップワイズ法」が使用されます。

変数増加ステップワイズ法の選択: モデルの効果がない状態から、これ以上追加または削除できなくなるまで、ステップワイズ法の基準に従って徐々に効果を追加および削除します。

- 投入または除去の基準: これは、モデルに効果を加えるかどうか、またはモデルから効果を削除するかどうかを決定するときに使用する統計です。「情報量基準 **(AICC)**」はモデルを指定された学習セットの尤度に基づき、過度に複雑なモデルにペナルティーを科すよう調整されます。「**F** 統計量」はモデルのエラーの改善に対する統計検定に基づいています。「調整済み **R2** 乗」は学習セットの適合度に基づき、過度に複雑なモデルにペナルティーを科すよう調整されます。「オーバーフィット防止基準 **(ASE)**」は、オーバーフィット防止セットの適合度 (平均平方誤差 (ASE)) に基づきます。オーバーフィット防止セットは、モデルの学習に使用されない元のデータ・セットのおよそ 30% の無作為サブサンプルです。

  「**F** 統計量」以外の基準を選択した場合、各ステップでその基準での最も大きい正の増分に対応する効果がモデルに追加されます。その基準での減少に対応するモデルの効果はいずれも削除されます。

  「**F** 統計量」を基準として選択すると、「次の値より小さい **p** 値の効果を含む」で指定したしきい値よりも小さい、最小の *p* 値を持つ効果が、各ステップでモデルに追加されます。デフォルトは 0.05 です。「次の値より大きい **p** 値の効果を削除する」で指定したしきい値より大きい *p* 値を持つモデルの効果はいずれも削除されます。デフォルトは 0.10 です。
- 最終モデルの最大効果数をカスタマイズする: デフォルトでは、すべての使用可能な効果をモデルに投入できます。あるいは、ステップワイズ・アルゴリズムが指定した効果の最大数でステップを終了する場合は、アルゴリズムは効果の現在のセットで停止します。
- ステップの最大数をカスタマイズする: 特定のステップ数の後、ステップワイズ・アルゴリズムが停止します。デフォルトでは、これは使用可能な効果の数の 3 倍です。あるいは、ステップの最大数を正整数で指定します。

最適サブセットの選択: 「可能なすべての」モデル、または少なくとも変数増加ステップワイズ法より大きい、可能なモデルのサブセットをチェックし、最適サブセットの基準に従って最適サブセットを選択します。「情報量基準 **(AICC)**」はモデルを指定された学習セットの尤度に基づき、過度に複雑なモデルにペナルティーを科すよう調整されます。「調整済み **R2** 乗」は学習セットの適合度に基づき、過度に複雑なモデルにペナルティーを科すよう調整されます。「オーバーフィット防止基準 **(ASE)**」は、オーバーフィット防止セットの適合度 (平均平方誤差 (ASE)) に基づきます。オーバーフィット防止セットは、モデルの学習に使用されない元のデータ・セットのおよそ 30% の無作為サブサンプルです。

最大の基準値を持つモデルが最良のモデルとして選択されます。

注**:** 最適サブセットによる選択は、変数増加ステップワイズ法による選択に比べてより多くの計算リソースを使用します。最適サブセットが、ブースティング、バギング、または非常に大きいデータ・セットと組み合わせて実行されると、変数増加ステップワイズ法の選択を使用して作成された標準モデルよりも大幅に時間がかかる場合があります。

モデル・オプション

モデル名: 対象フィールドに基づいて自動的にモデル名を生成するか、またはカスタム名を指定できます。自動的に生成される名前は、対象フィールド名です。

モデルがスコアリングされると、必ず予測値が計算されることに注意してください。新規フィールド名は、接頭辞として *$L-* が付いた対象フィールド名です。例えば、対象フィールド名が *sales* の場合、新規フィールド名は *$L-sales* になります。

インタラクティブ出力

Linear-AS モデルを実行した後で、以下の出力が使用可能になります。

モデル情報

「モデル情報」ビューは、モデルについての重要な情報を提供します。テーブルは次のようなハイレベルなモデル設定を特定します。

- 「フィールド」タブで指定されている対象の名前
- 回帰重みフィールド
- モデル選択設定で指定されたモデル構築方法
- 予測値入力の数
- 最終モデル内の予測値の数
- 赤池情報量基準 (補正) (AICC): AICC は、-2 (制限) 対数尤度に基づいて混合モデルを選択し、比較するための指標です。値が小さいほどモデルが良好であることを示します。AICC は小さなサンプルサイズに対して AIC を「修正」します。標本サイズが大きくなるに従い、AICC は AIC に収束します。
- R2 乗: これは線型モデルの適合度です。決定係数とも呼びます。これは、回帰モデルによって説明される従属変数の変動の比率です。値の範囲は 0 から 1 までです。値が小さい場合は、モデルが十分にデータに適合していないことを示します。
- 調整済み R2 乗:

レコード要約

「レコード要約」ビューは、モデルに組み込まれたレコード (ケース) とモデルから除外されたレコード (ケース) の数および割合についての情報を提供します。

予測変数の重要度

一般にモデリングの作業では、最も重要な予測フィールドの編集に集中して取り組み、最も重要でない予測フィールドは削除するか無視してしまいたくなるものです。予測値の重要度グラフを使用すると、モデル推定時に各予測値の相対重要度が示されるので便利です。値が相対的であるため、表示されるすべての予測変数の値の合計は 1.0 となります。予測値の重要度は、モデルの精度に関連しません。予測が正確かどうかに関係なく、予測時における各予測の重要度にのみ関連します。

予測対観測

縦軸の予測値に対し横軸に観測値を示した分割散布図が表示されます。点は 45 度の線上にあるのが理想です。このビューで、モデルによるレコードの予測にとりわけ問題があるかがわかります。

設定

モデルがスコアリングされると、予測値が必ず計算されます。新規フィールド名は、対象フィールド名に、接頭辞の *$L-* が付けられます。例えば、対象フィールドの名前が *sales* の場合、新規フィールド名は *$L-sales* になります。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使

用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。

- データベースの外部でスコアリング**:** 選択した場合、このオプションは、データベースからデータをフェッチし、SPSS Modeler でそのデータをスコアリングします。

ロジスティック回帰ノード

ロジスティック回帰 (名義回帰) は、入力フィールドの値に基づいてレコードを分類する統計手法です。線型と似ていますが、数値型フィールドではなくカテゴリー・フィールドを対象フィールドとします。二項モデル (2 つの異なるカテゴリーがある対象用) と多項モデル (3 つ以上のカテゴリーのある対象用) の両方がサポートされます。

ロジスティック回帰では、入力フィールド値を各出力フィールド カテゴリーに対応する確率に関連付ける一連の方程式が作成されます。モデルを生成した後は、そのモデルを使用して新しいデータの確率を推定できます。レコードごとに、各出力カテゴリー候補の所属確率が算出されます。最も確率の高い対象カテゴリーが、そのレコードの予測出力値として割り当てられます。

二項式のサンプル**:** 競合他社に奪われる顧客の数に関して、電気通信プロバイダーが心配しているとします。サービス使用量データを使用して、二項モデルを作成し、どの顧客が他のプロバイダーに移りそうかを予測できれば、オファーをカスタマイズして、できるだけ多くの顧客を保持することができます。対象に 2 つの明確なカテゴリー (移行しそうかそうでないか) があるために、二項モデルを使用します。

注**:** 二項モデルの場合のみ、文字列フィールドは 8 文字に制限されます。必要な場合は、データ分類ノードまたは匿名化ノードを使用して、これより長い文字列を再コード化できます。

多項の例**:** ある通信プロバイダーは、サービス利用パターンによって顧客ベースを区分し、顧客を 4 つのグループに分類しました。グループのメンバーシップを予測するために人口統計データを使用することで、多項モデルを作成して、見込み顧客をグループに分類し、それから個々の顧客へのオファーをカスタマイズできます。

要件**:** 1 つ以上の入力フィールドと、2 つ以上のカテゴリーを含む 1 つのカテゴリー対象フィールドが必要です。二項モデルの場合、対象は尺度がフラグ型である必要があります。多項モデルの場合は、対象が 2 つ以上のカテゴリーを持つフラグ型または名義型フィールドであることが必要です。両方 またはなし が設定されているフィールドは無視されます。モデルで使用するフィールド・タイプは、完全にインスタンス化する必要があります。

利点**:** たいていの場合、ロジスティック回帰モデルは非常に正確です。ロジスティック回帰モデルでは、シンボル値と数値の入力フィールドを処理できます。すべての対象カテゴリーに対する予測確率が算出されるため、「次善の推量」を簡単に識別することができます。ロジスティック・モデルは、グループ・メンバーが真にカテゴリー的なフィールドの場合に最も効果的です。グループ・メンバーが連続した値の範囲の値 (例えば、「高い IQ」対「低い IQ」) に基づいている場合、値の範囲全体から提供されるより豊富な情報を活かすために、線型の使用を考慮する必要があります。ロジスティック・モデルでは、自動的なフィールド選択も実行できます。ただし、ツリー・モデルや特徴量選択モデルなどの他のアプローチのほうが大規模データ・セットでは迅速に実行できます。最後に、ロジスティック・モデルは多くのアナリストやデータ・マイニング技術者によく理解されているので、他のモデル作成技法に対する基準として、比較の対象に使用されることがあります。

大きなデータセットを処理する場合、詳細出力オプションの「尤度比検定」を無効にすることにより、パフォーマンスを大幅に改善することができます。 詳しくは、トピック 197 ページの『ロジスティック回帰の詳細出力』を参照してください。

重要: 一時ディスク領域が不十分である場合は、二項ロジスティック回帰モデルが構築に失敗し、エラーが表示される場合があります。大きなデータ・セット (10GB 以上) から構築する場合は、同じ量の空きディスク容量が必要です。一時ディレクトリーの場所を設定するには、環境変数 SPSSTMPDIR を使用します。

ロジスティック回帰ノードの「モデル」オプション

モデル名: ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

分割モデルを作成 : 分割フィールドとして指定される入力フィールドの各値の個別モデルを作成します。詳しくは、28 ページの『分割モデルの作成』を参照してください。

手続き : 二項モデルまたは多項モデルのどちらが作成されるかを指定します。ダイアログ・ボックス内で使用できるオプションは、どのタイプのモデル作成手順が選択されたかによって異なります。

- **2** 項。 対象フィールドが、はい/いいえ、オン/オフ、男性/女性 のように、2 つの異なる値 (二分) のフラグ型または名義型の場合に使用されます。
- 多項分布。 対象フィールドが 2 つ以上の値をとる名義型フィールドの場合に使用されます。「主効果」、「すべての因子による」、または「ユーザー設定」を指定できます。

回帰式に定数項を含む: 生成される方程式に定数項を含めるかどうかを指定します。ほとんどの場合、このオプションは選択したままにしておきます。

二項モデル

二項モデルには、次の方法とオプションが利用できます。

方法: ロジスティック回帰モデルの作成に使用する手法を指定します。

- **Enter :** デフォルトの方法で、すべての項が方程式に直接入力されます。モデル作成時にフィールド選択は実行されません。
- 変数増加ステップワイズ法: フィールド選択に対する変数増加ステップワイズ法は、名前が示すとおりステップごとに方程式を作成していきます。初期モデルは最も単純なモデルで、方程式にモデルの項はありません (定数を除く)。各ステップで、モデルにまだ追加されていない項を評価します。評価された項の中で最適な項がモデルの予測精度を大幅に改善する場合、その項が追加されます。 さらに、モデルの現在の項が再評価され、削除してもモデルの性能が低下しないかどうかが判断されます。低下しないと判断されると、これらの項は削除されます。この処理が繰り返されて、他の項の追加や削除が行われます。項を追加してもモデルの性能が改善されず、項を削除してもモデルの性能が低下しなくなった時点で、最終モデルが生成されます。
- 変数減少ステップワイズ法: 変数減少ステップワイズ法は、基本的に変数増加ステップワイズ法の反対です。この方法では、すべての項が予測フィールドとして初期モデルに含まれています。各ステップにおいて、モデル中の項が評価され、削除してもモデルの性能が大幅に低下しない項が削除されます。また、前に削除された項が再評価され、それらの項目を追加するとモデルの予測精度が大幅に改善される

かどうかが判断されます。大幅に改善される場合は、その項がモデルに追加されます。項を削除してもモデルの性能が大幅に低下せず、項を追加してもモデルの性能が改善されなくなった時点で、最終モデルが生成されます。

カテゴリー入力 **:**カテゴリーとして特定される、つまり尺度がフラグ型、名義型、または順序型であると特定されたフィールドを一覧します。各カテゴリー・フィールドについて、対比およびベース・カテゴリーを指定できます。

- フィールド名 **:** この列には、カテゴリー入力のフィールド名が入ります。この列に連続または数値入力を追加するには、「フィールドを追加」アイコンをクリックし、必要な入力を選択します。
- 対比 **:** カテゴリー・フィールドの回帰係数の解釈は、使用する対比によって異なります。対比により、どのように仮説の検定を設定して推定平均を計算するのかを決定します。例えば、カテゴリー・フィールドに、パターンやグループ分けなどの暗黙の順序があることを知っている場合は、その順序をモデル作成するために対比を使用できます。使用できる対比は次の通りです。

  指標。対比は、所属カテゴリーの有無を示します。これがデフォルトの方法となります。

  単純: 参照カテゴリーを除く予測値フィールドの各カテゴリーが、参照カテゴリーと比較されます。

  「差分」。最初のカテゴリーを除く予測値フィールドの各カテゴリーが、前のカテゴリーの平均効果と比較されます。逆 Helmert 対比とも呼ばれます。

  Helmert: 最後のカテゴリーを除く予測値フィールドの各カテゴリーが、後のカテゴリーの平均効果と比較されます。

  反復: 最初のカテゴリーを除く予測値フィールドの各カテゴリーが、その前のカテゴリーと比較されます。

  多項式 **:** 直交多項対比。直交多項式の対比。各カテゴリーが等間隔で配置されていると仮定されます。多項対比は数値フィールドのみで使用可能です。

  偏差: 予測値フィールドの各カテゴリーが、全体の効果と比較されます。
- 基本カテゴリー **:** 選択された対比の種類について参照カテゴリーを決定する方法を指定します。「最初」を選択して、アルファベットで分類された入力フィールドに最初のカテゴリーを使用するか、または、「最後」を選択して最後のカテゴリーを使用します。デフォルトの基本カテゴリーは、「カテゴリ入力」領域にリストされた変数に適用されます。

  注**:** 対比設定が差分、Helmert、反復、または多項である場合は、このフィールドを使用できません。

全体の回答の各フィールドの効果の推定は、参照カテゴリーに関連するその他のカテゴリーの尤度の増減として計算されます。このために、特定の応答を得やすいフィールドおよび値を特定しやすくなっています。

ベース・カテゴリーは 0.0 として出力に表示されます。これは、それをそれ自体と比較すると空の結果が得られるためです。他のすべてのカテゴリーは、ベース・カテゴリーに関係する式として表示されます。詳しくは、トピック 199 ページの『ロジスティック ナゲット・モデルの詳細』を参照してください。

多項モデル

多項モデルには、次の方法とオプションが利用できます。

方法: ロジスティック回帰モデルの作成に使用する手法を指定します。

- **Enter :** デフォルトの方法で、すべての項が方程式に直接入力されます。モデル作成時にフィールド選択は実行されません。
- ステップワイズ法**:** フィールド選択に対するステップワイズ法は、名前が示すとおりステップごとに方程式を作成していきます。初期モデルは最も単純なモデルで、方程式にモデルの項はありません (定数を除く)。各ステップで、モデルにまだ追加されていない項を評価します。評価された項の中で最適な項がモデルの予測精度を大幅に改善する場合、その項が追加されます。 さらに、モデルの現在の項が再評価され、削除してもモデルの性能が低下しないかどうかが判断されます。低下しないと判断されると、これらの項は削除されます。この処理が繰り返されて、他の項の追加や削除が行われます。項を追加してもモデルの性能が改善されず、項を削除してもモデルの性能が低下しなくなった時点で、最終モデルが生成されます。
- 変数増加法 **:**「変数増加法」は、モデル作成がステップに分かれている点で「ステップワイズ法」と似ています。この手法の初期モデルは、最も単純なモデルで、定数と項しかモデルに追加することはできません。各ステップで、モデルに含まれていない項が、モデルをどの程度改善するかに基づいて検定され、最も適したフィールドがモデルに追加されていきます。追加する項がなくなるか、候補の項を追加してもモデルの性能がそれほど向上しなくなった時点で、最終モデルが生成されます。
- 変数減少法。変数減少法は、基本的に変数増加法の反対の方法です。この手法では、初期モデルに予測フィールドとしてすべての項が含まれているため、項の削除だけが行われます。モデルの改善にほとんど寄与しないモデルの項が削除され、削除してもモデルの性能が低下しない項がなくなった時点で、最終モデルができあがります。
- 変数減少ステップワイズ法**:** 変数減少ステップワイズ法は、基本的にステップワイズ法の反対です。この方法では、すべての項が予測フィールドとして初期モデルに含まれています。各ステップにおいて、モデル中の項が評価され、削除してもモデルの性能が大幅に低下しない項が削除されます。また、前に削除された項が再評価され、それらの項目を追加するとモデルの予測精度が大幅に改善されるかどうかが判断されます。大幅に改善される場合は、その項がモデルに追加されます。項を削除してもモデルの性能が大幅に低下せず、項を追加してもモデルの性能が改善されなくなった時点で、最終モデルが生成されます。

注**:** 自動手法 (ステップワイズ法、変数増加法、変数減少法など) は、非常に適応性の高い学習手法であるため、学習データがオーバーフィットする傾向が強くなります。これらの方法を使用するときは、新しいデータまたはデータ区分ノードを使用して作成され提供されたサンプルを使用して、作成されたモデルの妥当性を検証することが非常に大切になります。

対象の基本カテゴリー**:** 参照カテゴリーの決定方法を指定します。 これは、対象の中の他のすべてのカテゴリーのための回帰式を推定するベースラインとして使用します。「最初」を選択してアルファベットで分類された現在の対象フィールドに最初のカテゴリーを使用するか、または、「最後」を選択して最後のカテゴリーを使用します。また、「指定」を使用して、特定のカテゴリーを選択し、一覧から必要な値を選択できます。得られた値は、データ型ノードでそれぞれのフィールドに定義できます。

多くの場合、利益をもたらさない製品など、最も興味のないカテゴリーをベース・カテゴリーに指定します。その他のカテゴリーは、相対的にこのベース・カテゴリーに関係するので、それら自体のカテゴリーにふさわしくなっています。このために、特定の応答を得やすいフィールドおよび値を特定しやすくなっています。

ベース・カテゴリーは 0.0 として出力に表示されます。これは、それをそれ自体と比較すると空の結果が得られるためです。他のすべてのカテゴリーは、ベース・カテゴリーに関係する式として表示されます。詳しくは、トピック 199 ページの『ロジスティック ナゲット・モデルの詳細』を参照してください。

モデル タイプ **:** モデルの項を定義する 3 つのオプションがあります。「主効果」を選択すると、モデルに入力フィールドが個別に含まれ、入力フィールド間の交互作用は検定されません (倍数効果)。「すべて

の因子による」を選択すると、モデルに入力フィールドの主効果の他に、すべての交互作用が含まれます。すべての因子によるモデルの方が複雑な関係を把握できますが、解釈が難しく、オーバーフィットの可能性も高くなります。考えられる組み合わせの数が大きくなる可能性があるため、すべての因子によるモデルの場合、自動フィールド選択手法 (強制投入方以外の手法) は無効にされます。「ユーザー設定」を選択すると、モデルには指定した項 (主効果と交互作用) だけが含まれます。このオプションを選択した場合、「モデルの項」リストを使用してモデルに項を追加、または削除します。

モデルの項 :「ユーザー設定」でモデルを構築する場合、モデル中の項を明示的に指定する必要があります。このリストには、モデルの現在の項のセットが表示されます。「モデルの項」リストの右側にあるボタンを使用して、モデルの項を追加、削除することができます。

- モデルに項を追加するには、「モデルの項の新規追加」ボタンをクリックします。
- 項を削除するには、該当する項を選択して「選択したモデルの項の削除」ボタンをクリックします。

ロジスティック回帰モデルへの項の追加

ユーザー設定のロジスティック回帰モデルを要求する場合、「ロジスティック回帰モデル」タブで「モデルの項の新規追加」ボタンをクリックすることにより、モデルに項を追加することができます。項を指定するための「新規項」ダイアログ・ボックスが表示されます。

追加する項のデータ型 :「利用可能フィールド」リストで選択した入力フィールドに応じて、さまざまな方法でモデルに項を追加することができます。

- 単一の交互作用 : すべての選択したフィールドの交互作用を表す項を挿入します。
- 主効果: 選択した各入力フィールドに対して、1 つの主効果の項 (フィールド自体) を挿入します。
- すべての **2** 要因の交互作用 : 選択した入力フィールドの考えられる各組み合わせに対して、2 要因の交互作用の項 (入力フィールドの生成物) を挿入します。例えば、「利用可能フィールド」リストから入力フィールド *A*、*B*、および *C* を選択した場合、この方法では項 *A * B*、*A * C*、および *B * C* が挿入されます。
- すべての **3** 要因の交互作用 : 選択した入力フィールドの考えられる各組み合わせに対して、3 要因の交互作用の項 (入力フィールドの生成物) を挿入します (一度に 3 つを取得)。例えば、「利用可能フィールド」リストから入力フィールド *A*、*B*、*C*、および *D* を選択した場合、この方法では項 *A * B * C*、*A * B * D*、*A * C * D*、および *B * C * D* が挿入されます。
- すべての **4** 要因の交互作用 : 選択した入力フィールドの考えられる各組み合わせに対して、4 要因の交互作用の項 (入力フィールドの生成物) を挿入します (一度に 4 つを取得)。例えば、「利用可能フィールド」リストから入力フィールド *A*、*B*、*C*、*D*、および *E* を選択した場合、この方法では項 *A * B * C * D*、*A * B * C * E*、*A * B * D * E*、*A * C * D * E*、および *B * C * D * E* が挿入されます。

利用可能なフィールド: モデルの項を構築するために利用できる入力フィールドが表示されます。

プレビュー: 「挿入」 をクリックした場合に、フィールドと項のデータ型に基づいて、モデルに追加される項が表示されます。

挿入 : (現在のフィールドおよび項のデータ型の選択内容に基づいて) モデルに項を挿入し、ダイアログ・ボックスを閉じます。

ロジスティック回帰ノードの「エキスパート」オプション

ロジスティック回帰をよく理解している場合は、エキスパート・オプションを使用して、学習過程を調整できます。エキスパート・オプションを利用するには、「エキスパート」タブで「モード」に「エキスパート」を設定してください。

スケール **(**多項モデルのみ**):** パラメーターの分散共分散行列の推定の訂正に使用する、分散の尺度値を指定できます。「**Pearson** の相関係数」 を選択すると、Peason のカイ 2 乗統計を使用して尺度値が推定されます。「逸脱」 を選択すると、逸脱関数 (尤度比カイ 2 乗) 統計を使用して尺度値が推定されます。また、ユーザー定義尺度値を独自に指定することもできます。尺度値は正の数値でなければなりません。

すべての確率を追加**:** このオプションを選択すると、出力フィールドの各カテゴリーの確率が、ノードで処理される各レコードに追加されます。このオプションを選択しないと、予測されたカテゴリーの確率だけが追加されます。

例えば、3 つのカテゴリーの多項モデルの結果を含むテーブルには、5 つの新しい列が含まれます。1 つの列には正しく予測された結果の確率が一覧され、次の列にはこの予測が当たるか外れるかの確率が表示され、さらに 3 つの列には、それぞれのカテゴリーの予測が当たるか外れるかの確率が表示されます。 詳しくは、トピック 199 ページの『ロジスティック・モデル・ナゲット』を参照してください。

注 : このオプションは、常に二項モデルで選択されます。

特異性許容度**:** 特異性のチェックに使用する許容範囲を指定します。

収束**:** これらのオプションを使用して、モデル収束のパラメーターを制御することができます。モデルを実行するときに、収束設定によって、どれだけうまく適合するかを調べるために、異なるパラメーターを繰り返し実行する回数が制御されます。パラメーターを使用する回数が多くなればなるほど、結果が近くなります (つまり結果が収束します)。 詳しくは、トピック『ロジスティック回帰の収束オプション』を参照してください。

出力**:** これらのオプションを使用して、ノードによって構築されたモデル・ナゲットの詳細出力に表示される追加の統計量を要求することができます。詳しくは、トピック 197 ページの『ロジスティック回帰の詳細出力』を参照してください。

ステップ基準**:** ここでは、推定手法「ステップワイズ法」、「変数増加法」、「変数減少法」、または「変数減少ステップワイズ法」を使用したフィールドの追加および削除の基準を制御できます (「強制投入法」を選択した場合、このボタンは無効になります)。詳しくは、トピック 197 ページの『ロジスティック回帰のステップ基準オプション』を参照してください。

ロジスティック回帰の収束オプション

ロジスティック回帰モデルの推定に使用する収束パラメーターを設定することができます。

最大反復回数**:** モデルを推定するときの最大反復数を指定します。

最大ステップ二分**:** 段階 2 分とは、ロジスティック回帰の推定過程で複雑性の処理に使用される手法です。通常は、デフォルト設定を使用します。

対数尤度収束**:** 対数尤度の相対変化がこの値未満になると、反復が停止します。値が 0 の場合、この基準は使用されません。

パラメーター収束**:** パラメーター推定値の絶対変化または相対変化がこの値未満になると、反復が停止します。値が 0 の場合、この基準は使用されません。

デルタ **(**多項モデルのみ**) :** 空の各セルに追加する値 (入力フィールド値および出力フィールド値の組み合わせ) を、0〜1 の間で指定できます。指定すると、データ内のフィールド値の可能な組み合わせがレコード数に比して多い場合に、推定アルゴリズムで処理しやすくなります。デフォルトは 0 です。

ロジスティック回帰の詳細出力

生成された回帰モデル・ナゲットの詳細出力に表示する出力オプションを選択します。詳細な出力を表示するには、生成されたモデル・ナゲットを参照して、「詳細」タブをクリックします。 詳しくは、トピック 201 ページの『ロジスティック・モデル・ナゲットの詳細出力』を参照してください。

2 項オプション

モデルのために生成する出力の種類を選択します。 詳しくは、トピック 201 ページの『ロジスティック・モデル・ナゲットの詳細出力』を参照してください。

表示。各ステップで結果を表示するか、すべてのステップが完了するまで待つかを選択します。

exp(B) の **CI :** 式の中の各係数の信頼区間 (ベータとして表示) を選択します。信頼区間のレベルを指定します。デフォルトは 95% です。

残差の診断 **:** 残差のケースワイズ診断テーブルを要求します。

- 外側の外れ値 **(**標準偏差**):** リストされた変数の絶対標準化値が指定値以上である残差のケースだけをリストします。デフォルト値は 2 です。
- すべてのケース。残差のケースワイズ診断テーブルの中のすべてのケースを含みます。

  注 : このオプションでは入力レコードのそれぞれを一覧するために、すべてのレコードに 1 行が割り当てられて、レポートで非常に大きなテーブルができることがあります。

分類カットオフ**:** ケースを分類するための分割点を決定できます。予測値が分類分割点を超えるケースは正に分類され、分割点より小さい予測値を持つケースは負に分類されます。デフォルトを変更するには、0.01 から 0.99 までの値を入力します。

多項オプション

モデルのために生成する出力の種類を選択します。 詳しくは、トピック 201 ページの『ロジスティック・モデル・ナゲットの詳細出力』を参照してください。

注 :「尤度比検定」オプションを選択すると、ロジスティック回帰モデルの構築時間が非常に長くなります。モデルの構築に時間がかかりすぎる場合は、選択を解除するか、代りにワルド統計量またはスコア統計量を使用します。 詳しくは、トピック『ロジスティック回帰のステップ基準オプション』を参照してください。

反復履歴頻度 **:** 詳細出力に反復の状態を出力するステップの間隔を選択します。

信頼区間: 信頼区間のレベルを指定します。デフォルトは 95% です。

ロジスティック回帰のステップ基準オプション

ここでは、推定手法「ステップワイズ法」、「変数増加法」、「変数減少法」、または「変数減少ステップワイズ法」を使用したフィールドの追加および削除の基準を制御できます

モデル中の項数 **(**多項モデルのみ**):** 変数減少法および変数減少ステップワイズ法のモデル中の項の最小数、および変数増加法およびステップワイズ法のモデル中の項の最大数を指定することができます。最小数に 0

より大きい値を指定した場合、統計基準に基づいて項が削除されるような場合でも、モデルには最低限その数だけの項が含まれます。変数増加法、ステップワイズ法、および強制投入法のモデルの場合、最小数の設定は無視されます。最大数を指定した場合、統計基準に基づいて項が選択された場合でも、一部の項がモデルから削除される可能性があります。「最大を指定」 の設定は、変数減少法、変数減少ステップワイズ法、および強制投入法のモデルでは無視されます。

投入基準 **(**多項モデルのみ**):** 「スコア」を選択すると、処理速度が最大化されます。「尤度比」 オプションを使用するといくぶん確実な推定が得られますが、計算するのに時間がかかります。デフォルトの設定はスコア統計量になっています。

削除基準 **:** 強力なモデルには 「尤度比」 を選択します。モデル構築に必要な時間を短縮するには、「ワルド」 を選択してみることもできます。ただし、データに完全分離または疑似完全分離がある場合は (分離はモデル・ナゲットの「詳細」タブで測定可能)、ワルド統計量は特に信頼度が低下するので使用しないでください。デフォルトの設定は尤度比統計になっています。二項モデルの場合は、追加オプションの「条件式 **(If-Then)**」があります。これは、条件パラメーター推定値に基づく尤度比統計の確率に基づく削除テストを行います。

基準の有意しきい値**:** このオプションを使用すると、各フィールドに関連付けられた統計的確率 (*p* 値) に基づいて選択基準を指定することができます。フィールドは、該当する *p* 値が 「投入」 値より小さい場合にのみモデルに追加され、*p* 値が 「削除」 値より大きい場合にのみ削除されます。「投入」 には 「削除」 よりも小さい値を指定してください。

投入または削除の要件 **(**多項モデルのみ**):** アプリケーションによっては、交互作用の項に含まれるフィールドに対する低位の項がモデルに含まれていないと、モデルへの交互作用の項の追加が数値的に意味がないことがあります。例えば、モデル中に *A* および *B* がないと、モデルに *A * B* を入れても意味がありません。これらのオプションでは、ステップワイズ法による項の選択時に、このような依存関係をどのように処理するかを指定することができます。

- 不連続効果の階層 **:** 関連フィールドに対する低位の効果 (主効果またはより少ないフィールドを包含する交互作用) がすべてモデル中にすでに存在している場合にだけ、上位の効果 (より多くのフィールドを包含する交互作用) がモデルに投入されます。また、低位の効果と同じフィールドを包含する上位の効果がモデル中に存在している場合、低位の効果は削除されません。このオプションは、カテゴリー型フィールドに対してのみ適用されます。
- すべての効果の階層 **:** このオプションは、すべての入力フィールドの適用されることを除いて、前述のオプションと同じように機能します。
- すべての効果の包含 **:** 効果中に含まれているすべての効果がモデル中にも現れている場合にだけ、モデル中に効果が現れます。このオプションは、「すべての効果の階層」オプションと似ていますが、連続型フィールドの処理が異なっています。ある効果が他の効果を含むためには、含まれる (低位の) 効果に、それを含む (上位の) 効果に包含されているすべて の連続型フィールドがなければなりません。また、低位の (含まれる) 効果のカテゴリー型フィールドが、上位の (含む) 効果のカテゴリー型フィールドのサブセットでなければなりません。例えば、*A* と *B* がカテゴリー型フィールドで、*X* が連続型フィールドの場合、項 *A * B * X* には、項 *A * X* および *B * X* が含まれます。
- なし**:** 項はモデルから個別に追加、削除されます。

ロジスティック・モデル・ナゲット

ロジスティック・モデル・ナゲットは、ロジスティック回帰ノードによって推定された式を表します。ロジスティック・モデル・ナゲットには、線型回帰モデルが取得したすべての情報と、モデル構造とパフォーマンスに関する情報が含まれます。このタイプの式は、Oracle SVM などの他のモデルからも生成できます。

ロジスティック・モデル・ナゲットを含むストリームを実行すると、そのモデルの予測と関連付けられた確率を含む 2 つの新規フィールドが追加されます。新規フィールド名は予測された出力フィールドの名前から派生し、予測されたカテゴリーのフィールドには接頭辞の *$L-*、関連付けられた確率のフィールドには接頭辞の *$LP-* が付けられます。例えば、出力フィールドの名前が *colorpref* の場合、新規フィールド名は *$L-colorpref* と *$LP-colorpref* になります。また、ロジスティック回帰ノードで「すべての確率を追加」オプションを選択している場合は、出力フィールドの各カテゴリーに対して各レコードの対応するカテゴリーに属する確率を含むフィールドが追加されます。これらの追加のフィールドの名前は、出力フィールドの値を基に作成され、接頭辞の *$LP-* が付けられます。例えば、*colorpref* の有効な値が、*Red*、*Green*、*Blue* の場合、次の 3 つの新規フィールドが追加されます。*$LP-Red*、*$LP-Green*、および *$LP-Blue*です。

フィルター・ノードの生成**:** 「生成」メニューを使用すると、モデルの結果を基にして入力フィールドを通過させるための新しいフィルター・ノードを生成することができます。モデルで使われないフィールドだけでなく、多重共線性のためモデルから除外されたフィールドも、生成されたノードによりフィルタリングされます。

ロジスティック ナゲット・モデルの詳細

多項モデルの場合、ロジスティック・モデル・ナゲットの「モデル」タブには、左側の領域にモデルの式が、右側に予測変数の重要度がそれぞれ表示されます。二項モデルの場合、タブには予測変数の重要度のみが表示されます。 詳しくは、トピック 44 ページの『予測変数の重要度』を参照してください。

モデルの式

多項モデルの場合、左側の領域にはロジスティック 回帰モデルに推定された実際の式が表示されます。ベースラインのカテゴリーを除き、対象フィールドの各カテゴリーに 1 つずつ式があります。式はツリー形式で表示されます。このタイプの式は、Oracle SVM などの他の特定のモデルからも生成できます。

方程式 **:** 一連の予測値から、対象カテゴリーの確率を作成するために用いられる回帰式を表示します。対象フィールドの最後のカテゴリーは、「ベースライン カテゴリー」と判断されます。表示されている式は、特定の予測値のセットに対するベースライン カテゴリーと相対的な他の対象カテゴリーのオッズを提供します。与えられた予測パターンに対する各カテゴリーの予測確率は、オッズ値から作成されます。

確率の算出方法は ?

各方程式は、ベースライン カテゴリーに関連する特定の対象カテゴリーのオッズを計算します。ロジットとも呼ばれる対数オッズは、ベースライン カテゴリーに対する指定した対象カテゴリーの確率で、結果には自然対数関数が適用されます。ベースライン カテゴリーの場合、それ自身に相対するカテゴリーのオッズは 1.0 になるため、対数オッズは 0 になります。これをすべての係数が 0 となるベースライン カテゴリーの暗黙の式ととらえることができます。

特定の対象カテゴリーの対数オッズから確率を作成するには、そのカテゴリーから算出されるロジット値を取得し、次の式を適用する必要があります。

$$P(group_i) = \exp(g_i) \ / \ \Sigma_k \exp(g_k)$$

ここで、*g* は算出された対数オッズ、*i* はカテゴリー・インデックス、そして *k* は 1 から対象カテゴリー数までの値を表しています。

予測変数の重要度

オプションで、モデルの推定時に各予測値の相対的重要度を示すグラフを「モデル」タブに表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。このグラフは、モデル生成前に「精度分析」タブで 「予測変数の重要度を計算」が選択されている場合にのみ使用できます。 詳しくは、トピック 44 ページの『予測変数の重要度』を参照してください。

注 : ロジスティック回帰の場合、他のタイプのモデルに比べて予測変数の重要度の計算に時間がかかるため、デフォルトでは「分析」タブの項目は選択されていません。このオプションを選択すると、特に大きなデータセットを含む場合にパフォーマンスの速度が遅くなる場合があります。

ロジスティック・モデル・ナゲットの要約

ロジスティック回帰モデルの要約には、モデル生成に使われたフィールドと設定が表示されます。また、モデル作成ノードに接続されている精度分析ノードを実行した場合、その精度分析情報もこのセクションに表示されます。モデル・ブラウザー使用法に関する一般情報については、 43 ページの『モデル・ナゲットの参照』を参照してください。

ロジスティック・モデル・ナゲットの設定

ロジスティック・モデル・ナゲットの「設定」タブでは、確信度、確率、傾向スコアおよびモデル・スコアリング中の SQL 生成のオプションを指定します。このタブは、モデル・ナゲットがストリームに追加された後にのみ表示され、モデルおよび対象の種類によって異なるオプションが表示されます。

多項モデル

多項式モデルでは、次のオプションを使用できます。

確信度の計算**:** スコアリング中に確信度を計算するかどうかを指定します。

傾向スコア **(**調整なし**)** を計算 **(**フラグ型対象のみ**):** フラグ型対象を含むモデルの場合にのみ、対象フィールドに `true` の結果が指定されている尤度を示す傾向スコア (調整なし) を要求することができます。これらは、標準の予測値と確信度値に追加されています。調整済み傾向スコアは使用できません。 詳しくは、トピック 35 ページの『モデル作成ノードの分析オプション』を参照してください。

すべての確率を追加**:** 出力フィールドの各カテゴリーの確率を、ノードで処理される各レコードに追加するかどうかを指定します。このオプションを選択しないと、予測されたカテゴリーの確率だけが追加されます。3 つのカテゴリーを含む名義型対象フィールドの場合、例えばスコアリング出力には 3 つのカテゴリーそれぞれの列があり、また予測されるカテゴリーの確率を示す 4 つめの列があります。例えばカテゴリー「赤」、「緑」および「青」の確率がそれぞれ 0.6、0.3、0.1 の場合、予測カテゴリーは 0.6 の確率の「赤」となります。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- ネイティブの **SQL** に変更してスコアリング**:** これを選択すると、データベース内でモデルをスコアリングするためのネイティブ SQL が生成されます。

  注**:** このオプションの方が短時間で結果を得ることができますが、モデルの複雑度が増大すると、ネイティブ SQL のサイズと複雑度も、それに応じて増大します。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

注**:** 多項式モデルでは、「すべての確率を追加」が選択されている場合、または名義型対象を含むモデルでは「確信度の計算」が選択されている場合に、SQL 生成を使用できません。確信度の計算を含む SQL 生成は、フラグ型対象フィールドを含む多項モデルにのみサポートされています。SQL 生成は、二項モデルでは使用できません。

二項モデル

二項モデルの場合、確信度および確率は常に有効です。これらのオプションを無効にするオプションは使用できません。SQL 生成は、二項モデルでは使用できません。二項モデルに変更できる設定のみ、傾向スコア (調整なし) を計算できます。前述の多項モデルと同じように、フラグ型対象フィールドを含むモデルにのみ適用されます。 詳しくは、トピック 35 ページの『モデル作成ノードの分析オプション』を参照してください。

ロジスティック・モデル・ナゲットの詳細出力

ロジスティック回帰 (名義回帰ともいいます) の詳細出力からは、推定されるモデルとそのパフォーマンスに関する詳細情報を得られます。詳細出力に含まれる情報は、技術的な情報がほとんどです。この出力を適切に解釈するには、ロジスティック回帰分析に関する広範な知識が必要です。

警告 **:** 結果に関する警告または潜在的な問題を示します。

処理したケースの要約 **:** モデル内の各シンボル値フィールドごとに分類された処理済みレコードの数が表示されます。

ステップの要約 **(**オプション**) :** 自動フィールド選択を使用したときに、モデル作成の各ステップで追加または削除された効果を一覧します。

注 : ステップワイズ法、変数増加法、変数減少法、変数減少ステップワイズ法のみで表示されます。

反復履歴 **(**オプション**):** 最初の推定値から *n* 回の反復ごとにパラメーター推定値の反復履歴を表示します。*n* は表示間隔の値です。デフォルトは反復ごとの表示です (n=1)。

モデル適合情報 **(**多項モデル**) :** すべてのパラメーター係数が 0 のモデル (切片のみ) に対する最終モデルの尤度比検定が表示されます。

分類 **(**オプション**) :** 予測された出力フィールド値と実際の出力フィールド値の行列がパーセンテージとともに表示されます。

カイ **2** 乗適合度 **(**オプション**) :** Pearson と尤度比のカイ 2 乗統計が表示されます。これらの統計によって、学習データに対するモデルの全体的な適合度が検定されます。

Hosmer-Lemeshow 適合度 **(**オプション**) :** ケースをリスクの 10 分位にグループ分けして、観察された確率を 各 10 分位の中で予測される確率と比較した結果を表示します。この適合度統計は、多項モデルで使用される従来の適合度統計よりもより強力です。特に、連続的共変量のあるモデルおよび標本サイズの小さい調査で役に立ちます。

擬似 **R 2** 乗 **(**オプション**) :** Cox と Snell、Nagelkerke、および McFadden のモデル適合の *R* 2 乗測定値が表示されます。これらの統計は、線型の *R* 2 乗統計といくつかの点で似ています。

単調性の指標 **(**オプション**) :** データの中の調和ペア、不調和ペア、および結合ペアの数を、 それぞれがあらわすペアの総数のパーセンテージとともに、表示します。このテーブルには、Somers の D、Goodman と Kruskal のガンマ、Kendall のタウ a、および一致指数 C も表示されます。

情報量基準 **(**オプション**):** 赤池情報量基準 (AIC) と Schwarz のベイズ情報量基準 (BIC) を表示します。

尤度比検定 **(**オプション**) :** モデル効果の係数が統計的に 0 ではないかどうかについての統計検定を示します。有意な入力フィールドは出力で非常に有意度が低いフィールドとなります (「有意」 とラベルがつけられます)。

パラメーター推定値 **(**オプション**) :** 式係数の推定値、それらの係数の検定値、係数から派生したオッズ比 (ラベル *Exp(B)*)、およびオッズ比の信頼区間が表示されます。

漸近分散共分散行列/相関行列 **(**オプション**) :** 漸近分散共分散または係数推定値の相関、あるいはその両方が表示されます。

観測および予測度数 **(**オプション**) :** 各共変量パターンに対して、各出力フィールド値の観測および予測された度数が表示されます。特に数値入力フィールドを持つモデルの場合、このテーブルは非常に大きくなる可能性があります。結果のテーブルが大きすぎて使用できない場合は、テーブルが省略され、警告が表示されます。

因子分析モデル・ナゲット

因子分析モデル・ナゲットには、データの複雑性を整理する強力なデータ分解手法が 2 種類あります。この 2 つは、よく似ていますが、異なる点もあります。

- 主成分分析 **(PCA)** : 入力フィールドの線型結合が検出されます。成分が互いに直交する (直角に交わる) 場合に、フィールドのセット全体の分散を把握するのに役立ちます。主成分分析では、共有される分散と一意の分散の両方を含むすべての分散に焦点が当てられます。
- 因子分析 : 一連の観測フィールド内の相関パターンを説明する基本概念 (因子) が識別されます。因子分析では、共有される分散だけに焦点が当てられます。特定フィールドに固有な分散は、モデル推定時に考慮されません。因子分析モデル・ナゲットでは、いくつかの因子分析方法を使用できます。

どちらの手法でも、元のフィールド・セットの情報を効果的に要約する少数の派生フィールドの検出が目標です。

要件: 主成分分析-因子分析モデルでは、数値型フィールドだけを使用できます。因子分析または主成分分析を推定するには、役割が入力フィールドに設定された 1 つ以上のフィールドが必要です。役割が 対象、両方、または なし のフィールドは、非数値型フィールドのため無視されます。

利点: 因子分析と主成分分析では、情報の内容を大きく損なうことなく、データの複雑性を効果的に低下させることができます。これらの手法では、元データの入力フィールドを使用するよりも高速に動作する強力なモデルを作成できます。

因子分析モデル・ナゲットの「モデル」オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

抽出方法 **:** データの分解方法を指定します。

- 主成分分析 **:** これはデフォルトで選択されています。この方法では、主成分分析を使用して、入力フィールドを要約する成分が検出されます。
- 重みなし最小 **2** 乗法 **:** この因子分析手法では、入力フィールド間のリレーションシップ (相関) のパターンを最もよく再現できる一連の因子が検出されます。
- 一般化最小 **2** 乗法 **:** この因子分析手法は重みなし最小 2 乗法と似ていますが、重みを使用して一意 (非共有) の分散がたくさんあるフィールドの影響を小さくする点が異なります。
- 最尤法 **:** この因子分析手法では、入力フィールドにおける関係 (相関) の形に関する仮説に基づいて、観測された関係パターンを作成したであろうと最も生成したと考えられる因子方程式が生成されます。具体的には、学習データが多変量正規分布に従っていることを前提としています。
- 主因子法 **:** この因子分析手法は、主成分分析手法と非常に似ていますが、共有される分散だけに焦点を当てる点が異なります。
- ***α*** 因子分析 **:** この因子分析手法では、分析時のフィールドは大量の潜在入力フィールドからのサンプルと見なされます。この方法では、因子の統計的な信頼性が最大化されます。
- イメージ因子法 **:** この因子分析手法では、データ推定を使用して共通の分散が分離され、それを説明する因子が検出されます。

因子分析モデル・ナゲットの「エキスパート」オプション

因子分析および主成分分析をよく理解している場合は、エキスパート・オプションを使用して、学習過程を調整できます。エキスパート・オプションを利用するには、「エキスパート」タブで「モード」に「エキスパート」を設定してください。

欠損値: デフォルトで、IBM SPSS Modeler ではモデルで使用されるすべてのフィールドに有効な値を持つレコードだけが使用されます。(これは、欠損値のリストごとの削除とも呼ばれます。)欠損値が大量にある場合は、この方法では多くのレコードが除外され、データ不足で適切なモデルを作成できなくなることがあります。このような場合、「完全なレコードのみ使用」 オプションを選択解除できます。IBM SPSS Modeler は、フィールドの一部に欠損値のあるレコードなど、モデルを推定するためにできる限り多くの情報を使用します(これは、欠損値のペアごとの削除とも呼ばれます)。ただし、状況によっては、このようにして不完全なレコードを使用すると、モデルの推定に計算上の問題が発生することがあります。

フィールド**:** モデルの推定に、入力フィールドの相関行列と分散共分散行列のどちらを使用するかを指定します。デフォルトでは、「相関行列」が選択されています。

収束のための最大反復回数**:** モデルを推定するときの最大反復数を指定します。

因子抽出 **:** 入力フィールドから因子数を抽出するには、2 種類の方法があります。

- 固有値下限：指定された基準よりも大きい固有値を持つすべての因子またはコンポーネントを保持します。固有値は、各因子 (成分) が入力フィールドのセットにおける分散を要約する能力を示します。相関行列を使用する場合は、モデルでは指定値よりも大きな固有値を持つすべての因子 (成分) が保持されます。分散共分散行列を使用する場合は、指定値に平均固有値を掛けた値が基準となります。この計算により、このオプションを両方の行列に同じ意味で使用することができます。
- 最大数：固有値の降順に、指定された数の因子またはコンポーネントを保持します。つまり、大きい順に *n*個の固有値に対応する因子またはコンポーネントが保持されます。ここで、*n* は指定された基準を表します。デフォルトの抽出基準は、5 因子/コンポーネントです。

成分行列または因子行列形式：因子行列の形式を制御します (主成分分析モデルでは成分行列)。

- 値のソート：このオプションを選択すると、モデル出力の因子負荷が数値でソートされます。
- 表示する値の下限：このオプションを選択すると、行列中のパターンを見やすくするために、行列中の指定された閾値未満の得点は表示されません。

回転: このオプションにより、モデルの回転方法を制御することができます。詳しくは、トピック『因子分析モデル・ナゲットの「回転」オプション』を参照してください。

因子分析モデル・ナゲットの「回転」オプション

多くの場合、保持した因子のセットを数学的に回転させると、有用性が高まります。特に、解釈が容易になります。次のいずれかを選択します。

- 回転なし：デフォルトのオプションです。回転は使用されません。
- バリマックス：因子ごとに負荷の高いフィールド数を最小化する直交回転方法です。因子の解釈が単純化されます。
- 直接オブリミン: 斜交 (非直交) 回転法の 1 つ。デルタが 0 (デフォルト) の場合、斜交解が得られます。デルタが負になるに従って、因子の斜交度は下ります。デフォルト値の 0 を無効にするには、0.8 以下の数を入力してください。
- クォーティマックス：各フィールドの説明に必要な因子数を最小化する直交回転法です。観測されたフィールドの解釈が単純化されます。
- エカマックス：因子を単純化するバリマックス法と、フィールドを単純化するクォーティマックス法を組み合わせた回転法です。因子負荷が高いフィールドの数と、フィールドの説明に必要な因子の数が最小化されます。
- プロマックス: 因子を相関させることを可能にする、斜交回転です。直接オブリミン回転法よりも高速に計算できるため、大きなデータセットの場合に役立ちます。カッパによって、解の斜交度 (因子を相関させる度合) が制御されます。

因子分析モデル・ナゲット

因子分析モデル・ナゲットは、因子分析ノードで作成された因子分析および主成分分析 (PCA) モデルを表します。因子分析モデル・ナゲットには、学習済みのモデルが取得したすべての情報と、モデルのパフォーマンスと特性に関する情報が含まれます。

因子式モデルを含むストリームを実行すると、ノードによって、モデル内の各因子または各成分に対応する新規フィールドが追加されます。新規フィールド名はモデル名から派生し、接頭辞の *$F-* と接尾辞の *-n* が付けられます。ここで、*n* は因子または成分の番号です。例えば、*Factor* という名前で 3 つの因子を含むモデルの場合、新規フィールド名は *$F-Factor-1*、*$F-Factor-2*、および *$F-Factor-3* になります。

因子モデルにコード化された内容をより詳しく理解するには、さらにいくつかの下流を分析します。因子モデルの結果を表示するための便利な方法として、記述統計ノードを使用して、因子と入力フィールド間の相関を表示する方法があります。これにより、どの入力フィールドがどの因子に大きな負荷をかけているかが示され、因子が潜在的な意味または解釈を持っているかどうかを知ることができます。

また、詳細出力内で利用できる情報を使用して、因子モデルを評価することもできます。詳細出力を表示するには、モデル・ナゲットの「詳細」タブをクリックしてください。詳細出力には、多くの詳細情報が含まれており、因子分析と主成分分析に関する広範な知識を得られます。 詳しくは、トピック『因子分析モデル・ナゲットの詳細出力』を参照してください。

因子分析モデル・ナゲットの式

因子モデル・ナゲットの「モデル」タブで、各因子の因子得点方程式が表示されます。因子または成分の得点を計算するには、各入力フィールド値にその係数を掛け、結果を合計します。

因子分析モデル・ナゲットの要約

因子モデルの「要約」タブで、モデル生成に使用されフィールドと設定についての追加情報とともに、因子分析モデル内に保持された因子の数が表示されます。 詳しくは、トピック 43 ページの『モデル・ナゲットの参照』を参照してください。

因子分析モデル・ナゲットの詳細出力

因子分析の詳細出力からは、推定されるモデルとそのパフォーマンスに関する詳細情報が得られます。詳細出力に含まれる情報は、技術的な情報がほとんどです。この出力を適切に解釈するには、因子分析に関する広範な知識が必要です。

警告 **:** 結果に関する警告または潜在的な問題を示します。

共通性 **:** 因子または成分によって説明された各フィールドの分散の比率が表示されます。「初期」は、完全な因子のセットを使用して初期の共通性を提供します (モデルは初めに入力フィールドと同じ数の因子を持っています)。「抽出」は、保持されている因子のセットを基にして共通性を提供します。

説明された分散の合計 **:** モデル内の因子によって説明された分散の合計が表示されます。「初期の固有値」には、初期因子の完全なセットによって説明された分散が表示されます。「抽出後の負荷量平方和」には、モデル内に保持されている因子によって説明された分散が表示されます。「回転後の負荷量平方和」には、回転後の因子によって説明された分散が表示されます。斜交回転の場合は、「回転後の負荷量平方和」に、負荷量平方和のみが表示され、分散のパーセンテージは表示されないことに注意してください。

因子行列 **(**または成分行列**) :** 入力フィールドと回転のない因子との相関が表示されます。

回転後の因子行列または回転後の成分行列 **:** 直角回転の場合の、入力フィールドと回転後の因子との相関が表示されます。

パターン行列 **:** 斜交回転の場合の、入力フィールドと回転後の因子との偏相関が表示されます。

構造行列 **:** 斜交回転の場合の、入力フィールドと回転後の因子との単純な相関が表示されます。

因子相関行列 **:** 斜交回転の場合の、因子間の相関が表示されます。

判別分析ノード

判別分析により、所属グループ用の予測モデルが作成されます。このモデルは、各グループを最も適切に判別する予測変数の線型結合に基づいた 1 つの判別関数 (複数のグループの場合は、判別関数のセット) から構成されます。各関数は、所属グループが判明しているケースのサンプルから生成されます。各関数は、予測変数の測定は存在するが所属グループが不明な新規ケースに適用することができます。

例: 電話会社は、判別分析を使用し、顧客を利用データに基づいてグループ分けすることができます。これにより、将来性のある顧客と最も価値あるグループに収まりそうな顧客をスコアリングできるようになります。

要件**:** 1 つの入力フィールドと 1 つの対象フィールドが必要です。ターゲットは、文字列または整数のストレージを持つカテゴリー・フィールド (測定の尺度がフラグ型またはセット型) である必要があります。(ストレージは、必要に応じて、フィルター・ノードまたはフィールド作成ノードを使用して変換することができます。) 両方 またはなし が設定されているフィールドは無視されます。モデルで使用するフィールド・タイプは、完全にインスタンス化する必要があります。

利点**:** 判別分析とロジスティック回帰は両方とも、分類モデルに適しています。ただし、判別分析のほうが入力フィールドについての想定が多い傾向があります。例えば、正規分布され、連続型となる必要があります。これらの要件が満たされると、特に標本サイズが小さい場合に、よりよい結果が生じます。

判別分析ノードのモデル関連のオプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

分割モデルを作成 **:** 分割フィールドとして指定される入力フィールドの各値の個別モデルを作成します。詳しくは、 28 ページの『分割モデルの作成』を参照してください。

方法: 予想値をモデルに入力するのに、次のオプションが利用できます。

- **Enter :** デフォルトの方法で、すべての項が方程式に直接入力されます。モデルの予測精度を大幅に改善しない項は、追加されません。
- ステップワイズ法**:** 初期モデルは最も単純なモデルで、方程式にモデルの項はありません (定数を除く)。各ステップで、モデルにまだ追加されていない項を評価します。評価された項の中で最適な項がモデルの予測精度を大幅に改善する場合、その項が追加されます。

注 : ステップワイズ法には、学習データがオーバーフィットする強い傾向があります。このような方法を使用する場合は、提供されたテスト・サンプルまたは新しいデータを使用して、作成されたモデルの妥当性を検証することが非常に大切になります。

判別分析ノードのエキスパート関連のオプション

判別分析をよく理解している場合は、エキスパート・オプションを使用して学習過程を微調整できます。エキスパート・オプションを利用するには、「エキスパート」タブで 「モード」 に 「エキスパート」 を設定してください。

事前確率: このオプションは、所属グループの事前の知識に応じて分類係数を調整するかどうかを決定します。

- すべてのグループが等しい: すべてのグループについて同じ事前確率が想定されます。係数に対する影響はありません。
- グループ・サイズから計算: サンプル内のグループ・サイズの観測結果により、所属グループの事前確率が決定されます。例えば、分析に含まれる観測結果の 50% が最初のグループ、25% が 2 番目のグループ、残りの 25% が 3 番目のグループに分類される場合、分類係数は、他の 2 つのグループに対して最初のグループ内の所属性の尤度を増やすように調整されます。

共分散行列の使用: グループ内の共分散行列または個別グループ共分散行列を使用してケースを分類することができます。

- グループ内: プールされたグループ内共分散行列は、ケースの分類に使用します。
- グループ別: グループ別共分散行列は分類に使用します。分類は (元の変数ではなく) 判別関数に基づいて行うため、このオプションは必ずしも 2 次の判別と等価ではありません。

出力**:** これらのオプションによって、ノードに構築されたモデル・ナゲットの詳細出力に表示される付加統計量を要求することができます。詳しくは、トピック『判別分析ノードの出力関連のオプション』を参照してください。

ステップ基準**:** これらのオプションにより、ステップワイズ推定方法を使用したフィールドの追加と除去のための基準を制御することができます (「強制投入法」 を選択した場合、このボタンは無効になります)。詳しくは、トピック 208 ページの『判別分析ノードのステップ関連のオプション』を参照してください。

判別分析ノードの出力関連のオプション

生成されたロジスティック回帰モデル・ナゲットの詳細出力に表示する出力オプションを選択します。詳細な出力を表示するには、生成されたモデル・ナゲットを参照して、「詳細」タブをクリックします。詳しくは、トピック 209 ページの『判別分析モデル・ナゲットの詳細出力』を参照してください。

記述統計: 使用可能なオプションは、平均 (標準偏差を含む)、1 変量の分散分析、Box の *M* 検定です。

- *Means (*平均値 *(*信頼性分析*)).* 独立変数の合計、グループ平均値、および標準偏差を表示します。
- *Univariate ANOVAs (1* 変量の分散分析 *(*判別分析*)).* 一元配置分散分析を実行して、独立変数ごとにグループ平均値の等質性を検定します。
- *Box* の *M* 検定: グループの共分散行列の等質性を調べる検定。サンプル数が十分に多いが p 値が有意でない場合は、行列が異なるという証拠が不十分であることを意味します。この検定は、多変量正規性からの逸脱に対して敏感です。

関数係数: 使用可能なオプションは、Fisher の分類係数と非標準化係数です。

- *Fisher's (Fisher* の分類関数の係数*).* 分類に直接使用できる、Fisher の分類関数の係数を表示します。分類関数の一連の係数をグループごとに個別に求め、最大判別得点 (分類関数の値) を持つグループにケースを割り当てます。
- *Unstandardized (*標準化されていない *(*判別分析*)).* 標準化していない判別関数の係数を表示します。

行列: 独立変数の使用可能な係数の行列は、グループ内相関行列、グループ内共分散行列、グループ別共分散行列、全共分散行列です。

- グループ内相関: 相関を計算する前にすべてのグループの個別の共分散行列を平均化することによって得られるプールされたグループ内相関行列を表示します。
- グループ内共分散: プールされたグループ内共分散行列を表示します。全共分散行列とは異なる場合があります。この行列は、すべてのグループの個別の共分散行列を平均化することによって得られます。
- グループ別共分散: 各グループの個別の共分散行列を表示します。

- 全共分散: すべてのケースから得た共分散行列を、1 つのサンプルから取り出したかのように表示します。

分類: 次の出力が分類結果に伴って表示されます。

- ケースごとの結果: 実際のグループ、予測グループ、事後確率、および判別得点のコードをケースごとに表示します。
- 集計表: 判別分析に基づいて各グループに正しくまたは誤って割り当てられたケースの数。「混同行列」と呼ぶこともあります。
- *Leave-one-out* 分類: 分析における各ケースを、そのケース以外のすべてのケースから派生した関数で分類します。「U 手法」とも呼びます。
- 領域マップ: 関数の値に基づいてケースをグループに分類するために使用する境界のプロット。これらの数値は、ケースの分類先グループに対応します。各グループの平均値は、その境界の内側にアスタリスクで示します。判別関数が 1 つしかない場合は、このマップを表示しません。
- 結合されたグループ: 最初の 2 つの判別関数の値を使用して全グループ散布図を作成します。関数が 1 つしかない場合は、代わりにヒストグラムが表示されます。
- グループ別: 最初の 2 つの判別関数の値のグループ別散布図を作成します。関数が 1 つしかない場合は、代わりにヒストグラムを表示します。

ステップワイズ法**:** 「ステップの要約」 には、各ステップ実行後の利用可能なすべての変数の統計量が表示され、「ペアごとの距離の **F** 値」 には、グループ内のペアごとの *F* 比率の行列が表示されます。*F* 比率は、グループ間の Mahalanobis 距離の有意性検定に使用できます。

判別分析ノードのステップ関連のオプション

方法: 新しい変数を投入または削除する際に使用される統計方法を選択します。有効な選択肢は、Wilks のラムダ、解明不明の分散、Mahalanobis の距離、最小 *F* 比率、Rao の *V* です。 Rao の *V* を使用すると、投入する変数に対して *V* 単位の最小増分を指定することができます。

- *Wilks* のラムダ: ステップワイズ判別分析における変数選択法の 1 つ。変数が Wilks のラムダを低下させる程度に基づいて式に投入する変数を選択します。各ステップでは、Wilks のラムダが最小になる変数を投入します。
- 解明不明の分散: 各ステップで、グループ間の説明されない分散の合計を最小化する変数を投入します。
- *Mahalanobis* の距離: 独立変数のケースの値と全ケースの平均との差異の程度を示す指標。マハラノビスの距離が大きい場合は、ケースにおいて 1 つ以上の独立変数に極値が存在することを示します。
- 最小 *F* 比率: グループ間のマハラノビスの距離から計算した F 比の最大化に基づく、ステップワイズ分析での変数選択法。
- *Rao* の *V*: グループ平均値の差の指標。Lawley-Hotelling のトレースとも呼びます。各ステップで、Rao の V における増加を最大化する変数を投入します。このオプションを選択した後、分析に投入する変数が持つべき最小値を入力してください。

「基準」。有効な選択肢は、「ステップワイズのための **F** 値」と「ステップワイズのための **F** 値確率」です。変数の投入用の値と削除用の値を入力します。

- ステップワイズのための *F* 値: F 値が「投入」の値より大きい場合に変数をモデルに投入し、「削除」の値より小さい場合に変数を除去します。「投入」は「削除」より大きくなければならず、いずれの値も正でなければなりません。さらに多くの変数をモデルに投入するには、「投入」の値を下げてください。さらに多くの変数をモデルから除去するには、「除去」の値を上げてください。

- ステップワイズのための *F* 値確率: F 値の有意水準が「投入」の値より小さい場合に変数をモデルに投入し、有意水準が「削除」の値より大きい場合に変数を除去します。「投入」は「削除」より小さくなければならず、いずれの値も正でなければなりません。さらに多くの変数をモデルに投入するには、「投入」の値を上げてください。さらに多くの変数をモデルから除去するには、「除去」の値を下げてください。

判別分析モデル・ナゲット

判別分析モデル・ナゲットは、判別分析ノードによって推定された式を表します。判別分析モデル・ナゲットには、判別分析モデルが取得したすべての情報と、モデル構造とパフォーマンスに関する情報が含まれます。

判別分析モデル・ナゲットを含むストリームを実行すると、そのモデルの予測と関連付けられた確率を含む 2 つの新規フィールドが追加されます。新規フィールドの名前は予測された出力フィールドの名前から派生し、予測されたカテゴリーのフィールドには接頭辞の *$D-*、関連付けられた確率のフィールドには接頭辞の *$DP-* が付けられます。例えば、出力フィールドの名前が *colorpref* の場合、新規フィールド名は *$D-colorpref* と *$DP-colorpref* になります。

フィルター・ノードの生成**:** 「ノードの生成」メニューを使用すると、モデルの結果を基にして入力フィールドを通過させるための新しいフィルター・ノードを生成することができます。

予測変数の重要度

オプションで、モデルの推定時に各予測値の相対的重要度を示すグラフを「モデル」タブに表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。このグラフは、モデル生成前に「精度分析」タブで 「予測変数の重要度を計算」が選択されている場合にのみ使用できます。 詳しくは、トピック 44 ページの『予測変数の重要度』を参照してください。

判別分析モデル・ナゲットの詳細出力

判別分析の詳細出力からは、推定されるモデルとそのパフォーマンスに関する詳細情報が得られます。詳細出力に含まれる情報は、技術的な情報がほとんどです。この出力を適切に解釈するには、判別分析に関する広範な知識が必要です。 詳しくは、トピック 207 ページの『判別分析ノードの出力関連のオプション』を参照してください。

判別分析モデル・ナゲットの設定値

判別分析モデル・ナゲットの「設定」 タブでは、モデルのスコアリングの際に傾向スコアを取得できます。このタブは、フラグ型対象のモデルの場合にのみ使用でき、また、モデル・ナゲットがストリームに追加された後にのみ使用できます。

未調整傾向スコアを計算**:** フラグ型対象 (yes または no の予測を返す) を持つモデルの場合、対象フィールドに指定された真 (true) の結果の尤度を示す傾向スコアが必要な場合があります。また、スコアリング時に生成することができるその他の予測および確信度値があります。

調整済み傾向スコアを計算 **:** 未調整傾向スコアは、学習データにのみ基づき、このデータがオーバーフィットする多くのモデルの傾向によって過度に楽観的な場合があります。調整済み傾向は、テストまたは検証用データ区分に対してモデルのパフォーマンスを評価することによって補正しようとします。このオプションでは、モデルの生成前にデータ区分フィールドをストリーム内で定義し、調整済み傾向スコアがモデル作成ノードで有効化されている必要があります。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

判別分析モデル・ナゲットの要約

判別分析モデル・ナゲットの「要約」タブには、モデル生成に使われたフィールドと設定が表示されます。また、モデル作成ノードに接続されている精度分析ノードを実行した場合、その精度分析情報もこのセクションに表示されます。モデル・ブラウザー使用法に関する一般情報については、 43 ページの『モデル・ナゲットの参照』を参照してください。

一般化線型ノード

一般化線型モデルは、一般の線型モデルを拡張し、従属変数が指定されたリンク関数経由で因数と共変量に直線的に関係付けられるようにします。さらにこのモデルでは、非正規分布の従属変数を使用することができます。一般化線型モデルは、正規分布した回答の線型、バイナリ データのためのロジスティック・モデル、計数データのための対数線型モデル、間隔を決めて検閲される延命データのための補数対数-対数モデルなどの広く使用される統計モデルに加えて、一般的なモデルの定式を通じて多くのほかの統計モデルも対象とします。

例: 海運会社では一般化線型モデルを使用して、異なる期間に建設された複数の種類の船の損害数にポアソン回帰を当てはめることができ、構築されたモデルによって損害を受けやすい船の種類を判断することができます。

自動車保険会社では一般化線型モデルを使用して、自動車に対する損害請求にガンマ回帰を当てはめることができ、構築されたモデルによって請求に最も寄与する因子を判断することができます。

医療研究者は、一般化線型モデルを使用して区間打ち切り生存率データに補ログ マイナス・ログを当てはめ、病状が再発する時間を予測します。

一般化線型モデルは、入力フィールドの値を出力フィールドの値に関係付ける方程式を作成することで機能します。モデルを生成した後は、そのモデルを使用して新しいデータの値を推定できます。レコードごとに、各出力カテゴリー候補の所属確率が算出されます。最も確率の高い対象カテゴリーが、そのレコードの予測出力値として割り当てられます。

要件**:** 1 つ以上の入力フィールドと、2 つ以上のカテゴリーを含む 1 つの対象フィールド (測定の尺度が連続型 またはフラグ型 のフィールド) が必要です。モデルで使用するフィールド・タイプは、完全にインスタンス化する必要があります。

強度**:** 一般化線型モデルは非常に柔軟性がありますが、モデル構造を選択するプロセスは自動化されていないので、「ブラック ボックス」型のアルゴリズムには必要ないことですが、使用するデータをある程度熟知している必要があります。

一般化線型ノードの「フィールド」オプション

モデル作成ノードの「フィールド」タブ (31 ページの『モデル作成ノードのフィールド・オプション』を参照) に表示される通常の対象、入力、およびデータ区分のユーザー指定オプションに加えて、一般化線型ノードには次の特別な機能があります。

重みフィールドを使用: スケール・パラメーターは、応答の分散に関連する推定モデル・パラメーターです。尺度重み付けは、観測ごとに異なる「既知の」値です。尺度重み付け変数が指定された場合、応答の分散と関連性を持つ尺度パラメーターは、各観測ごとに尺度重み付け変数によって分割されます。尺度の重み値が 0 以下または欠損値のレコードは、分析に使用されません。

対象フィールドが試行セットで生じたイベント数を表す **:** 回答が一定の試行回数のセット内で発生したイベント数の場合、対象フィールドにはイベント数が含まれ、この試行回数を含んだ追加の変数を選択できます。ただし、試行数がすべての被験者に対して同じである場合は、固定値を使用して試行を指定することができます。試行回数は、各レコードのイベント数以上である必要があります。また、イベント数は非負整数、試行数は正の整数であることが必要です。

一般化線型ノードの「モデル」オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

分割モデルを作成 **:** 分割フィールドとして指定される入力フィールドの各値の個別モデルを作成します。詳しくは、 28 ページの『分割モデルの作成』を参照してください。

モデル タイプ **:** 作成するモデルのタイプには 2 つのオプションがあります。「主効果のみ」 を選択すると、モデルに入力フィールドのみが個別に含まれ、入力フィールド間の交互作用は検定されません (倍数効果)。「主効果およびすべての **2** 要因の交互作用」 には、入力フィールド主効果に加えてすべての 2 要因の交互作用が含まれます。

オフセット。 オフセット項は、「構造的」な予測フィールドです。その係数はモデルにより推定されませんが、値が 1 であると見なされます。したがって、オフセットの値は単純に対象の線型予測フィールドに追加されます。このことはポアソン回帰モデルでは特に有用であり、各ケースには興味深いイベントへのさまざまな公開レベルがある可能性があります。

例えば、個々のドライバーの事故率をモデリングする場合に、3 年間で過失責任事故が 1 回のドライバーと 25 年間で過失責任事故が 1 回のドライバーの間では、重要な違いがあります。運転手の経験をオフセット項として加味する場合、事故の発生数は対数リンクを持つポアソン応答または負の 2 項応答としてモデル化できます。

分布およびリンクの種類のその他の組み合わせには、オフセット変数のその他の変換が必要です。

注 *:* 変数オフセット・フィールドが使用された場合、指定されたフィールドは入力にも使用されるべきではありません。上流のソースでオフセット・フィールドの役割を 「なし」 と設定するか、必要な場合はデータ型ノードを入力します。

フラグ型対象に対するベース カテゴリー**:**

二者択一の回答については、従属変数のための参照カテゴリーを選択できます。このことでパラメーター推定値や保存済みの値などの一定の出力に影響を与えることができますが、モデルの適合度を変更してはなりません。例えば、二者択一の回答の値が 0 と 1 だとします。

- デフォルトでは、手続きは最後の (最高値の) カテゴリー、つまり 1 を参照カテゴリーにします。この状況で、モデルに保存された確率でケースが値 0 になる機会を推定します。また、パラメーター推定値はカテゴリー 0 の尤度への関連として解釈される必要があります。
- 最初の (最低値の) カテゴリー、つまり 0 を参照カテゴリーに指定する場合は、モデルに保存された確率で、このケースが値 1 になる機会を推定します。
- カスタム カテゴリーを指定し、変数にラベルを定義した場合は、リストから値を選択して参照カテゴリーを設定できます。これは、モデル指定の途中で特定の変数がどのようにコーディングされたか正確にわからないときに便利です。

モデルに切片を含む: 通常、モデルには切片が含まれています。データが原点を通ると仮定できる場合は、切片を除外できます。

一般化線型ノードの「エキスパート」オプション

一般化線型モデルをよく理解している場合は、エキスパート・オプションを使用して、学習過程を調整できます。エキスパート・オプションを利用するには、「エキスパート」タブで 「モード」 に 「エキスパート」 を設定してください。

対象フィールドの分布およびリンク関数

分布 **:**

このセクションで、従属変数の分布を指定します。非正規分布と非恒等式リンク関数を指定する機能は、一般化線型モデルの従来の機能を越える、本質的な機能の向上です。分布とリンク関数には多くの組み合わせの可能性があり、その中のいくつかは特定のデータセットに適切な場合があるので、この選択が先験的な理論考察または一番適合するように見える組み合わせによって導き出される可能性があります。

- **2** 項。 この分布は、二者択一の回答またはイベント数を表す変数に対してのみ、適しています。
- ガンマ**:** この分布は、正の値が大きくなるほどゆがむ正のスケール値を持つ変数に適しています。データの値が 0 以下または欠損している場合は、対応するケースが分析に使用されません。
- 逆ガウス**:** この分布は、正の値が大きくなるほどゆがむ正のスケール値を持つ変数に適しています。データの値が 0 以下または欠損している場合は、対応するケースが分析に使用されません。
- 負の **2** 項**:** この分布は *k* の成功を観測するために必要な試行回数として考えることができ、負ではない整数値の変数に適しています。データの値が非正数、0 より小さい、または欠損している場合は、対応するケースが分析に使用されません。負の 2 項分布の補助パラメーターの固定値は、0 以上の数値です。補助パラメーターが 0 に設定されている場合、この分布の使用とポアソン分布の使用が同じ結果となります。
- 正規。これは、値が対称で、中心 (平均) 値に関してベル型の分布であるスケール変数に適しています。従属変数は、数値でなければなりません。
- ポワソン。この分布は一定期間の対象のイベントの発生回数として考えることができ、負ではない整数値の変数に適しています。データの値が非正数、0 より小さい、または欠損している場合は、対応するケースが分析に使用されません。
- **Tweedie :**この分布はガンマ分布のポアソン混合によって表すことができる変数に適しています。分布の「混合」とは連続型 (負でない十数値） と離散型 (単一値 0 の 正の確率質量) の特性を結合するこ

とです。従属変数は 0 またはそれ以上のデータ値を持った数値である必要があります。データの値が 0 より小さい、または欠損している場合は、対応するケースが分析に使用されません。Tweedie 分布のパラメーターの固定値は 1 以上 2 以下のどんな数字でもかまいません。

- 多項分布。 この分布は順序型応答を表す変数に適しています。従属変数は数値または文字列で、少なくとも 2 つの明確な有効データ値を持っている必要があります。

リンク関数 :

リンク関数は、モデルを推定できるようにする従属変数の変形です。使用できる関数は次のとおりです。

- 恒等式: $f(x)=x$。従属変数は変換されません。このリンクは、どの分布でも共に使用できます。
- 補ログ・マイナス・ログ: $f(x)=\log(-\log(1-x))$。これは、二項分布とのみ使用するのが適しています。
- 累積コーチット: $f(x) = \tan(\pi (x - 0.5))$。回答の各カテゴリーの累積確率に適用されます。これは、多項分布にのみ使用するのが適しています。
- 累積補ログ・マイナス・ログ: $f(x)=\ln(-\ln(1-x))$。回答の各カテゴリーの累積確率に適用されます。これは、多項分布にのみ使用するのが適しています。
- 累積統計のロジット : $f(x)=\ln(x / (1-x))$、回答の各カテゴリーの累積統計確率に適用されます。これは、多項分布にのみ使用するのが適しています。
- 累積負の対数-対数: $f(x)=-\ln(-\ln(x))$。回答の各カテゴリーの累積確率に適用されます。これは、多項分布にのみ使用するのが適しています。
- 累積統計のプロビット : $f(x)=\Phi^{-1}(x)$、回答の各カテゴリーの累積統計確率に適用され、Φ^{-1} は逆標準正規の累積分布関数です。これは、多項分布にのみ使用するのが適しています。
- 対数: $f(x)=\log(x)$。このリンクは、どの分布でも共に使用できます。
- 対数-補数: $f(x)=\log(1-x)$。 これは、二項分布とのみ使用するのが適しています。
- ロジット。$f(x)=\log(x/(1-x))$。これは、二項分布とのみ使用するのが適しています。
- 負の 2 項: $f(x)=\log(x / (x+k^{-1}))$。$k$ は、負の 2 項分布の補助パラメーターです。これは、負の二項分布とのみ使用するのが適しています。
- 負ログ・マイナス・ログ: $f(x)=-\log(-\log(x))$。これは、二項分布とのみ使用するのが適しています。
- オッズべき乗: $f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$ (α が 0 以外の場合)。$f(x)=\log(x)$ (α が 0 の場合)。α には数値を指定する必要があり、その数値は実数である必要があります。これは、二項分布とのみ使用するのが適しています。
- プロビット。$f(x)=\Phi^{-1}(x)$。Φ^{-1} は逆標準正規累積分布関数です。これは、二項分布とのみ使用するのが適しています。
- べき乗: $f(x)=x^{\alpha}$、$\alpha \neq 0$。$f(x)=\log(x)$ (α が 0 の場合)。α には数値を指定する必要があり、その数値は実数である必要があります。このリンクは、どの分布でも共に使用できます。

パラメーター: 特定の分布オプションを選択すると、このグループのコントロールを使用してパラメーター値を指定することができます。

- 負の 2 項分布のパラメーター : 負の 2 項分布の場合、値を指定するか、システムが推定値を提供できるように選択できます。
- **Tweedie** のパラメーター : Tweedie 分布の場合、固定値に 1.0 ～ 2.0 の値を指定します。

  パラメーター推定値: このグループ内のコントロールにより、推定方法を指定し、パラメーター推定値に最初の値を提供できるようになります。

  – 方法: パラメーター推定方法を 1 つ選択できます。Newton-Raphson、Fisher スコアリング、または複合型の中から選択します。複合型では、Newton-Raphson 方法へ切り替わる前に、Fisher スコ

アリングの反復が実行されます。複合型の Fisher スコアリングフェーズ中で Fisher 反復の最大回数に達する前に収束が達成された場合、アルゴリズムは Newton-Raphson 方法で続行されます。

– スケール・パラメーター方法**:** スケール・パラメーター推定方法を 1 つ選択できます。最尤法は、モデル効果と共同で尺度パラメーターを推定します。このオプションは、回答が負の 2 項分布、ポアソン分布、または 2 項分布の場合は有効でないことに注意してください。逸脱度および Pearson カイ 2 乗のオプションは、これらの統計からスケール・パラメーターを推定します。または、スケール・パラメーターに固定値を指定することもできます。

- 共分散行列 **:** モデルに基づく推定量は、ヘッセ行列の一般化逆行列の負の値です。頑健推定量 (Huber/White/サンドウィッチ推定量とも呼ばれる) は「修正された」モデルに基づく推定量で、分散やリンク関数の指定が不適切な場合でも、精度の高い共分散の推定を行うことができます。

反復: これらのオプションを使用して、モデル収束のパラメーターを制御することができます。 詳しくは、トピック『一般化線型モデルの反復』を参照してください。

出力**:** これらのオプションによって、ノードに構築されたモデル・ナゲットの詳細出力に表示される付加統計量を要求することができます。詳しくは、トピック『一般化線型モデルの詳細出力』を参照してください。

特異性許容度**:** 特異な (または非可逆的な) 行列に直線的に依存する列があり、これが推定アルゴリズムに重大な問題を引き起こす可能性があります。特異性に近似する行列でさえ貧弱な結果を導く可能性があるので、手続きは、この行列を特異性の許容範囲内と判断して処理します。正の値を指定します。

一般化線型モデルの反復

一般化線型モデルの推定に使用する収束パラメーターを設定することができます。

反復作業。 使用可能なオプションは次のとおりです。

- 最大反復回数**:** アルゴリズムが実行できる反復の最大回数。負でない整数を指定してください。
- 最大ステップ二分**:** 各反復で、対数-尤度が増えるか最大段階 2 分に達するまで、ステップ・サイズが係数 0.5 単位で減らされます。正の整数を指定します。
- データ・ポイントの区切りを確認**:** 選択すると、パラメーター推定値が一意の値を持っていることを確認するテストが、アルゴリズムにより実行されます。区切りは、手続きが各ケースを正しく分類するモデルを作成できるときに作成されます。このオプションは、バイナリー・フォーマットの 2 項回答に使用できます。

収束基準**:** 使用可能なオプションは次のとおりです。

- パラメーター収束**:** 選択すると、パラメーター推定値内の絶対または相対的な変化が指定された値より少ない (正数であることが必要) 反復の後に、アルゴリズムが停止します。
- 対数尤度収束**:** 選択すると、対数-尤度関数内の絶対または相対的な変化が指定された値より少ない (正数であることが必要) 反復の後に、アルゴリズムが停止します。
- **Hessian** 収束**:** 絶対的指定の場合は、Hessian 収束に基づいた統計が指定された正数より小さいと、収束とみなされます。相対的指定の場合は、統計が指定された正数値と対数-尤度の絶対値の積より小さいと、収束とみなされます。

一般化線型モデルの詳細出力

生成された線型モデル・ナゲットの詳細出力に表示する出力オプションを選択します。詳細な出力を表示するには、生成されたモデル・ナゲットを参照して、「詳細」タブをクリックします。詳しくは、トピック 216 ページの『GenLin モデル・ナゲットの詳細出力』を参照してください。

出力できる内容は以下のとおりです。

- ケース処理の要約: 分析対象となるケースおよび分析対象から除外されるケースの数と割合、および「相関データの集計」表が表示されます。
- 「記述統計」。記述統計量に加え、従属変数、共変量、および因子に関する要約情報が表示されます。
- 「モデル情報」。データ・セット名、従属変数またはイベント変数と試行変数、オフセット変数、スケール重み変数、確率分布、およびリンク関数が表示されます。
- 「適合度統計量」。逸脱度とスケール逸脱度、Pearson のカイ 2 乗と尺度付き Pearson カイ 2 乗、対数尤度、赤池情報量基準 (AIC)、有限サンプル修正 AIC (AICC)、ベイズ情報量基準 (BIC) 、一致 AIC (CAIC) が表示されます。
- 「モデル要約統計量」。モデル適合度のオムニバス検定に関する尤度比統計量や、効果ごとのタイプ I またはタイプ III の対比に関する統計量を含むモデル適合度検定が表示されます。
- 「パラメーター推定値」。パラメーター推定値およびそれに対応する検定統計量と信頼区間が表示されます。オプションで、生のパラメーター推定値に加えて指数化されたパラメーター推定値も表示できます。
- 「パラメーター推定値の共分散行列」。推定パラメーター分散共分散行列が表示されます。
- 「パラメーター推定値の相関行列」。推定パラメーター相関行列が表示されます。
- 「対比係数 **(L)** 行列」。デフォルトの効果の対比係数が表示されます。また、「EM 平均」タブで要求されている場合は、推定周辺平均の対比係数も表示されます。
- 「一般の推定可能関数」。対比係数 (L) 行列を生成するための行列が表示されます。
- 反復履歴**:** パラメーター推定値と対数-尤度の反復履歴頻度が表示され、傾斜ベクトルおよび Hessian 行列の最後の評価が表示されます。反復履歴テーブルにより、0 番目の反復 (最初の推定値) で始まる *n* 番目の反復ごとにパラメーター推定値が表示されます。*n* は、表示間隔の値です。反復履歴頻度が要求された場合は、*n* にかかわらず、最後の反復が常に表示されます。
- **LaGrange** 乗数検定 **:** 尺度パラメーターの有効性を査定するためのラグランジュの未定係数法検定の統計量を表示します。これは、逸脱または Pearson カイ 2 乗を使用して計算されるか、正規、ガンマ、および逆ガウス分布のために固定値に設定されます。負の 2 項分布の場合は、固定値の補助パラメーターが検定対象となります。

モデル効果**:** 使用可能なオプションは次のとおりです。

- 分析の種類 **:**作成する分析の種類を指定します。タイプ I 分析は通常、モデル内の予測フィールドの順序付けに先見的な理由がある場合に適しています。一方タイプ III は、それよりも一般的に適用可能です。ワルドまたは尤度比統計は、カイ 2 乗統計の選択に基づいて計算されます。
- 信頼区間。50 より大きい 100 未満の信頼度レベルを指定します。ワルド区間は、パラメーターが漸近正規分布であるという想定に基づいています。プロファイル尤度区間はより正確ですが、計算上負荷がかかる場合があります。プロファイル尤度区間の許容レベルは、区間の計算に使用する反復アルゴリズムを停止するために使用する基準です。
- 対数尤度関数。対数尤度関数の表示形式を制御します。関数全体には、パラメーター推定値に対する定数の追加の項が含まれます。これはパラメーター推定値には何の効力もなく、ある種のソフトウェア製品による表示用に残されています。

GenLin モデル・ナゲット

GenLin モデル・ナゲットは、一般化線型ノードによって推定された式を表します。これらには、モデルが取得したすべての情報と、モデル構造とパフォーマンスに関する情報が含まれます。

GenLin モデル・ナゲットを含むストリームを実行すると、ノードによって、対象フィールドの性質ごとに異なる内容の新規フィールドが追加されます。

- フラグ型対象 : 予測されたカテゴリーと関連する確率、および各カテゴリーの確率を含むフィールドが追加されます。最初の 2 つの新規フィールドの名前は予測された出力フィールドの名前から派生し、予測されたカテゴリーのフィールドには接頭辞の *$G-*、関連付けられた確率のフィールドには接頭辞の *$GP-* が付けられます。例えば、出力フィールドの名前が *default* の場合、新規フィールド名は *$G-default* と *$GP-default* になります。後ろ 2 つの追加のフィールドの名前は、出力フィールドの値を基に作成され、接頭辞の *$GP-* が付けられます。例えば、*default* の有効な値が *Yes* および *No* の場合、新規フィールドの名前は、*$GP-Yes* および *$GP-No* となります。
- 連続型対象 : 予測された平均と標準誤差を含むフィールドが追加されます。
- 一連の繰り返し回数内のイベント数を表す連続型対象 : 予測された平均と標準誤差を含むフィールドが追加されます。
- 順序型対象: 順序セットの各値に対する、予測されたカテゴリーと関連する確率を含むフィールドが追加されます。フィールド名は予測された順序セットの値から派生し、予測されたカテゴリーのフィールドには接頭辞の *$G-*、関連付けられた確率のフィールドには接頭辞の *$GP-* が付けられます。

フィルター・ノードの生成: 「ノードの生成」メニューを使用すると、モデルの結果を基にして入力フィールドを通過させるための新しいフィルター・ノードを生成することができます。

予測変数の重要度

オプションで、モデルの推定時に各予測値の相対的重要度を示すグラフを「モデル」タブに表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。このグラフは、モデル生成前に「精度分析」タブで 「予測変数の重要度を計算」が選択されている場合にのみ使用できます。 詳しくは、トピック 44 ページの『予測変数の重要度』を参照してください。

GenLin モデル・ナゲットの詳細出力

一般化線型モデルの詳細出力からは、推定されるモデルとそのパフォーマンスに関する詳細情報が得られます。詳細出力に含まれる情報は、技術的な情報がほとんどです。この出力を適切に解釈するには、このタイプの分析に関する広範な知識が必要です。 詳しくは、トピック 214 ページの『一般化線型モデルの詳細出力』を参照してください。

GenLin モデル・ナゲットの設定

GenLin モデル・ナゲットの「設定」タブを使用すると、モデルのスコアリング時に傾向スコアを取得できます。また、モデルのスコアリング中に SQL 生成を行うこともできます。このタブは、フラグ型対象のモデルの場合にのみ使用でき、また、モデル・ナゲットがストリームに追加された後にのみ使用できます。

未調整傾向スコアを計算: フラグ型対象 (yes または no の予測を返す) を持つモデルの場合、対象フィールドに指定された真 (true) の結果の尤度を示す傾向スコアが必要な場合があります。また、スコアリング時に生成することができるその他の予測および確信度値があります。

調整済み傾向スコアを計算 : 未調整傾向スコアは、学習データにのみ基づき、このデータがオーバーフィットする多くのモデルの傾向によって過度に楽観的な場合があります。調整済み傾向は、テストまたは検証用データ区分に対してモデルのパフォーマンスを評価することによって補正しようとします。このオプションでは、モデルの生成前にデータ区分フィールドをストリーム内で定義し、調整済み傾向スコアがモデル作成ノードで有効化されている必要があります。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

GenLin モデル・ナゲットの要約

GenLin モデル・ナゲットの「要約」タブには、モデル生成に使われたフィールドと設定が表示されます。また、モデル作成ノードに接続されている精度分析ノードを実行した場合、その精度分析情報もこのセクションに表示されます。モデル・ブラウザー使用法に関する一般情報については、 43 ページの『モデル・ナゲットの参照』を参照してください。

一般化線型混合モデル

GLMM ノード

一般化線型混合モデル (GLMM) を作成します。

一般化線型混合モデル

一般化線型混合モデルが線型モデルを拡張したことにより、以下のようになりました。

- 目標は指定したリンク関数を介して因子および共変量に線形に関連します。
- 対象は非正規分布をする場合があります。
- 観測を相関させることができます。

一般化線型混合モデルには、単純な線型から、非正規分布の縦断的データを取り扱う複雑なマルチレベル・モデルまで、さまざまなモデルがあります。

例: 地区の教育委員会は、実験的な教授法が数学の点数を向上させる効果的であるかどうかを判断するために一般化線形混合モデルを使用することができます。生徒は同じ教師によって教えられているので、同じ教室の生徒を相関させる必要があり、同じ学校内の教室にも相関があり場合があるため、変動の異なるソースを考慮して学校やクラスのレベルで変量効果を含めることができます。

医学研究者は、新しい抗てんかん薬が、てんかん発作の患者の割合を減らすことができるかどうかを判断するために一般化線形混合モデルを使用することができます。同一患者から繰り返し測定を行う場合通常、正の相関関係があるため、いくつかの変量効果を持つ混合モデルが適切となります。対象フィールドである発作の数は、正の整数値をとるため、ポアソン分布と対数リンクを持つ一般化線形混合モデルが適切となる場合があります。

テレビ、電話、およびインターネットサービスのケーブル・プロバイダーの経営陣は、潜在的な顧客についての詳細を知るために一般化線形混合モデルを使用することができます。考えられる回答は名義型尺度であ

るため、同社のアナリストは、所定の調査応答者の回答の中のサービスの種類（テレビ、電話、インターネット）全体のサービスの使い方についての質問への回答との相関関係をキャプチャするために、ランダム切片と一般化ロジット混合モデルを使用します。

「データ構造」タブは、観測値が相関しているときに、データセット内のレコード間の構造的な関係を指定することができます。データセット内のレコードが独立した観察を表している場合、このタブでは何も指定する必要はありません。

被験者。指定したカテゴリー・フィールドの値を組み合わせることで、データ・セット内の被験者を一意的に定義する必要があります。例えば、1 つの病院内の被験者を定義するには、患者 *ID* 変数が 1 つあれば十分ですが、複数の病院間で患者の ID 番号が重複する場合は、病院 *ID* と患者 *ID* を組み合わせて使用することが必要になります。反復測定では、被験者ごとに複数の観測値が記録されるため、各被験者がデータセット内の複数のレコードを占めることがあります。

被験者は、その他の被験者から独立している見なすことができる観測単位です。例えば、医学研究では患者の血圧の測定値は、他の患者の測定値から独立していると見なすことができます。被験者の定義は、被験者ごとに測定を繰り返す場合、これらの観測間の相関関係をモデル化したい場合に重要になります。例えば、病院に連続して訪問する際に一人の患者の血圧測定値は相関していると期待できる場合があります。

「データ構造」タブの「被験者」として指定されたすべてのフィールドを使用して、残差共分散構造の被験者を定義し、変量効果ブロックの変量効果共分散構造の被験者を定義するフィールドのリストを提供します。

反復測定。ここで指定するフィールドは、反復観測値を特定するために使用されます。例えば、変数「週」 は医療研究の 10 週間にわたる観測を指定、また 「月」 および 「日」 を同時に使用して 1 年間にわたって行われる日常の観測を示します。

共分散グループの定義。ここで指定するカテゴリー・フィールドは、反復効果共分散パラメーターの独立したセットを定義します。グループ化フィールドの交差分類により定義される各カテゴリーに対して 1 つです。すべての被験者は、同じ共分散のタイプです。同じ共分散グループ内の被験者は、パラメーターに同じ値を持つことになります。

空間共分散座標。反復共分散タイプに対して空間共分散タイプのいずれかが選択されている場合、このリストの変数に、反復する観測の座標を指定します。

反復共分散タイプ。残差に対する共分散構造を指定します。使用できる構造は次のとおりです。

- 1 次自己回帰 (AR1)
- 自己回帰の移動平均 (1,1) (ARMA11)
- 複合対称
- 対角
- 計測された単位
- 空間: べき乗
- 空間: 指数
- 空間: ガウス
- 空間: 線型
- 空間: 線型-対数
- 空間: 球形
- Toeplitz

- 非構造化
- 分散成分

対象**:** これらの設定は、リンク関数を介してターゲット、その分布、および予測因子との関係を定義します。

対象 **:** 対象値を示します。対象は必須です。これは、任意の尺度を持つことができ、対象の尺度は適切な分布とリンク関数を制限します。

- 分母に試行数を使用する。対象の応答が一連の試行内で発生するイベント数である場合、対象フィールドにはイベント数が含まれます。また試行回数を含んでいる追加フィールドを選択できます。例えば、新しい農薬をテストするときは、さまざまな濃度の農薬をアリのサンプルに噴霧して死んだアリの数と各サンプルのアリの数を記録します。この場合、死んだアリの数を記録するフィールドは、対象（イベント）フィールドとして指定されなければならず、各サンプル中のアリの数を記録するフィールドは、試行フィールドとして指定する必要があります。アリの数は、各サンプルに対して同じである場合、試行回数は、固定値を使用して指定することができます。

  試行回数は、各レコードのイベント数以上である必要があります。また、イベント数は非負整数、試行数は正の整数であることが必要です。
- 参照カテゴリーのカスタマイズ。カテゴリー対象に参照カテゴリーを選択できます。このことでパラメーター推定値などの一定の出力に影響を与えることができますが、モデルの適合度を変更してはなりません。例えば、対象値がデフォルトで0、1、および 2 となる場合、手順では最後の（最も高い値を持つ）カテゴリー、または 2 を参照のカテゴリーにします。この場合、パラメーター推定はカテゴリー 2 の尤度に相対 してカテゴリー 0 または 1 の尤度に関連すると解釈されます。カスタム カテゴリーを指定して、対象がラベルを定義している場合、リストから値を選択して基準のカテゴリーを設定することができます。これは、モデル指定の途中で特定のフィールドがどのようにコーディングされたか正確にわからないときに便利です。

対象の分布および線型モデルとの関連 **(**リンク**)** 。予測値の値を指定することで、モデルは指定した形状に従う対象値の分布、および指定したリンク関数を使用して予測値と線型に関連する対象値を予期します。いくつかの共通モデルへのショートカットが提供されます。または、ショートカットのリストにない分布とリンク関数の特定の組み合わせを使用する場合は、「カスタム」設定を選択します。

- 線型モデル。同一リンクを持つ正規分布を指定します。これは、線型または分散分析モデルを使用して対象が予測される際に有用です。
- ガンマ回帰。対数リンクを持つガンマ分布を指定します。これは、対象に含まれる値がすべて正の値で、値が大きくなるほどゆがむ場合に使用されます。
- 対数線型。対数リンクを持つポワソン分布を指定します。これは、対象が一定期間内の出現回数を表すときに使用されます。
- 負の **2** 項回帰。対数リンクを持つ負の 2 項分布を指定します。これは、対象と分母が *k* の成功を観測するために必要な試行回数を表すときに使用されます。
- 多項ロジスティック回帰。多項分布を指定します。これは対象が複数カテゴリーの応答である場合に使用されます。累積ロジットリンク（順序型結果）または一般化ロジット・リンク（マルチカテゴリー名義型回答）を使用します。
- **2** 値ロジスティック回帰。ロジット・リンクを持つ 2 項分布を指定します。これは対象がロジスティック回帰モデルで予測される 2 値応答である場合に使用されます。
- **2** 値プロビット。プロビット・リンクを持つ 2 項分布を指定します。これは対象が基礎の正規分布を使用した 2 値応答である場合に使用されます。

- 打ち切り。補ログ・マイナス・ログ・リンクを持つ 2 項分布を指定します。これは終了イベントのない観測がある場合の生存分析で有用です。

分布　このセクションで、対象の分布を指定します。非正規分布と非恒等式リンク関数を指定する機能は、線形混合モデルを超える、一般化線型混合モデルの本質的な機能の向上です。分布とリンク関数には多くの組み合わせの可能性があり、その中のいくつかは特定のデータセットに適切な場合があるので、この選択が先験的な理論考察または一番適合するように見える組み合わせによって導き出される可能性があります。

2 項　この分布は、二者択一の回答またはイベント数を表す対象に対してのみ、適しています。

ガンマ
: この分布は、より大きな正数値の方向へ歪められる正のスケール値を持つ対象に適しています。データの値が 0 以下または欠損している場合は、対応するケースが分析に使用されません。

逆ガウス分布
: この分布は、より大きな正数値の方向へ歪められる正のスケール値を持つ対象に適しています。データの値が 0 以下または欠損している場合は、対応するケースが分析に使用されません。

多項分布
: この分布は、複数カテゴリーの応答を表す対象に適しています。モデルの形式は、対象の尺度によって異なります。

  名義型対象は、モデル パラメーターの異なるセットが、（参照カテゴリーを除く）対象のカテゴリーごとに推定される名義型多項モデルとなります。特定の予測のためのパラメーター推定値は、参照カテゴリーに相対的に、予測値と対象の各カテゴリーの尤度との関連性を示します。

  順序型対象は、従来の切片項が、対象カテゴリーの累積確率に関連する閾値パラメーターのセットと置き換えられる順序型多項モデルとなります。

負の **2** 項分布
: 負の 2 項回帰は、対数リンクを含む負の 2 項分布を使用します。対象が高い分散度を持つ出現回数を表す場合に使用します。

正常　これは、値が対称で、中心 (平均) 値に関してベル型の分布である連続型対象に適しています。

ポアソン
: この分布は一定期間の対象のイベントの発生回数として考えることができ、負ではない整数値の変数に適しています。データの値が非正数、0 より小さい、または欠損している場合は、対応するケースが分析に使用されません。

リンク関数
: リンク関数は、モデルを推定できるようにする対象の変形です。使用できる関数は次のとおりです。

  恒等式
  : $f(x)=x$。対象は変換されません。このリンクは、多項分布を除き、どの分布でも共に使用できます。

  補数対数-対数
  : $f(x)=\log(-\log(1-x))$。これは、二項分布とまたは多項分布とのみ使用するのが適しています。

コーチット
: $f(x) = \tan(\pi (x - 0.5))$。これは、二項分布とまたは多項分布とのみ使用するのが適しています。

Log
: $f(x)=\log(x)$。このリンクは、多項分布を除き、どの分布でも共に使用できます。

対数**-**補数
: $f(x)=\log(1-x)$。 これは、二項分布とのみ使用するのが適しています。

ロジット
: $f(x)=\log(x/(1-x))$。これは、二項分布とまたは多項分布とのみ使用するのが適しています。

負の対数**-**対数
: $f(x)=-\log(-\log(x))$。これは、二項分布とまたは多項分布とのみ使用するのが適しています。

プロビット
: $f(x)=\Phi^{-1}(x)$。Φ^{-1} は逆標準正規累積分布関数です。これは、二項分布とまたは多項分布とのみ使用するのが適しています。

べき乗
: $f(x)=x^{\alpha}$ (α が 0 以外の場合)。$f(x)=\log(x)$ (α が 0 の場合)。α には数値を指定する必要があり、その数値は実数である必要があります。このリンクは、多項分布を除き、どの分布でも共に使用できます。

固定効果**:** 固定効果の因子は、一般的に、その関心の値がすべてのデータセットで表現されるフィールドとして考えられ、スコアリングに使用することができます。デフォルトでは、ダイアログ内の他の場所で指定されていない定義済みの入力の役割を持つフィールドは、モデルの固定効果部分に入力されます。カテゴリー型 (フラグ型、名義型、順序型) フィールドは、モデルでの因子として使用され、連続型フィールドは共変量として使用されます。

ソース・リスト内の1つ以上のフィールドを選択し、効果リストにドラッグして、効果をモデルに入力します。作成する効果の種類は、選択項目をドロップするホットスポットによって異なります。

- 主相互作用**:** ドロップされたフィールドは、効果リストの一番下にある別の主効果として表示されます。
- **2** 要因。ドロップされたフィールドのすべての可能なペアが、2 要因の交互作用として効果リストの下部に表示されます。
- **3** 要因。ドロップされたフィールドのすべての可能なトリプレットが、3 要因の交互作用として効果リストの下部に表示されます。
- ***.** ドロップされたすべてのフィールドの組み合わせは、効果リストの下部に単一の相互作用として表示されます。

効果ビルダーの右側にあるボタンを使用すると、さまざまな操作を行うことができます。

表 *10.* 効果ビルダー・ボタンの説明

アイコン	説明
	削除したい条件を選択し、削除ボタンをクリックして、固定効果モデルから用語を削除します。

表 *10.* 効果ビルダー・ボタンの説明 *(*続き*)*

アイコン	説明
	順序を変更する条件を選択し、上向きまたは下向きの矢印をクリックして、固定効果モデル内の項目を並べ替えます。
	「カスタム項の追加」ボタンをクリックし、「『カスタム項の追加』」ダイアログを使用して、入れ子になった項をモデルに追加します。

定数項を含める。通常、切片はモデルに含まれます。データが原点を通ると仮定できる場合は、切片を除外できます。

カスタム項の追加**:** この手順で使用するモデルの入れ子になった項を構築することができます。入れ子になった項は、値が別の因子の水準と相互作用しない因子または共変量の影響をモデル化するのに便利です。例えば、食料品店チェーンは、複数の店舗の場所で、顧客の支出の習慣に従う場合があります。各顧客はそれぞれ、これらの場所の 1 つにだけ頻繁に訪れるため、「顧客」 効果は「店舗の場所」 効果内で入れ子にすることはできます。

また、そのような同一の共変量を含む多項式の項のような相互作用の効果を含める、または入れ子になった項に複数レベルの入れ子を追加することができます。

制限**:** ネスト項目には、次のような制限があります。

- 交互作用内のすべての因子は固有である必要があります。したがって、*A* が因子の場合、*A*A* の指定は無効です。
- 1 つのネスト効果内の因子はすべて、固有のものである必要があります。したがって、*A* が因子の場合、*A(A)* の指定は無効です。
- 共変量内に効果を入れ子にすることはできません。*A* が因子であり、*X* が共変量である場合、*A(X)* を指定しても無効になります。

入れ子になった項の構築

1. もう一つの因子内に入れ子になっている要因および共変量を選択し、矢印ボタンをクリックします。
2. 「**(**内**)**」 をクリックします。
3. 前の因子または共変量が入れ子になっている因子を選択し、矢印のボタンをクリックします。
4. 「項目を追加」をクリックします。

また、相互作用の効果を含める、または入れ子になった項に複数レベルの入れ子を追加することができます。

変量効果**:** ランダム効果の因子は、値がデータファイル内の値のより大きな母集団から無作為標本を検討することができるフィールドです。これらは、対象の過剰な変動を説明するのに便利です。デフォルトでは、「データ構造」タブで複数の被験者を選択した場合、変量効果ブロックが、最も内側の被験者を超えて、被験者ごとに作成されます。例えば、「データ構造」タブで学校、クラス、生徒を選択した場合、以下の変量効果が自動的に作成されます。

- 変量効果 1:被験者は学校です (効果なし、定数項のみ)
- 変量効果 2:被験者は学校 * クラスです (効果なし、定数項のみ)

以下の方法で変量効果ブロックの作業が可能です。

1. 新しいブロックを追加するには、「ブロックの追加...」をクリックします。「『変量効果ブロック』」ダイアログが開きます。
2. 既存のブロックを編集するには、編集するブロックを選択して、「ブロックの編集」をクリックします。「『変量効果ブロック』」ダイアログが開きます。
3. 1 つ以上のブロックを削除するには、削除するブロックを選択して「削除」ボタンをクリックします。

変量効果ブロック**:** ソース・リスト内の1つ以上のフィールドを選択し、効果リストにドラッグして、効果をモデルに入力します。作成する効果の種類は、選択項目をドロップするホットスポットによって異なります。カテゴリー型 (フラグ型、名義型、順序型) フィールドは、モデルでの因子として使用され、連続型フィールドは共変量として使用されます。

- 主相互作用**:** ドロップされたフィールドは、効果リストの一番下にある別の主効果として表示されます。
- **2** 要因。ドロップされたフィールドのすべての可能なペアが、2 要因の交互作用として効果リストの下部に表示されます。
- **3** 要因。ドロップされたフィールドのすべての可能なトリプレットが、3 要因の交互作用として効果リストの下部に表示されます。
- ***.** ドロップされたすべてのフィールドの組み合わせは、効果リストの下部に単一の相互作用として表示されます。

効果ビルダーの右側にあるボタンを使用すると、さまざまな操作を行うことができます。

表 *11.* 効果ビルダー・ボタンの説明

アイコン	説明
	削除したい条件を選択し、削除ボタンをクリックして、モデルから用語を削除します。
	順序を変更する条件を選択し、上向きまたは下向きの矢印をクリックして、モデル内の項目を並べ替えます。
	「カスタム項の追加」ボタンをクリックし、「 222 ページの『カスタム項の追加』」ダイアログを使用して、入れ子になった項をモデルに追加します。

定数項を含める。デフォルトでは、切片はランダム効果モデルに含まれません。データが原点を通ると仮定できる場合は、切片を除外できます。

このブロックにパラメーター予測を表示 **(Display parameter predictions for this block)**: ランダム効果パラメーター推定値を表示する場合に指定します。

共分散グループの定義。ここで指定するカテゴリー・フィールドは、ランダム効果共分散パラメーターの独立したセットを定義します。グループ化フィールドの交差分類により定義される各カテゴリーに対して 1

つです。グループ化フィールドの異なるセットを各変量効果ブロックに指定することができます。すべての被験者は、同じ共分散のタイプです。同じ共分散グループ内の被験者は、パラメーターに同じ値を持つことになります。

被験者の組み合わせ。「データ構造」タブで事前設定された被験者の組み合わせから、ランダム効果の被験者を指定できます。例えば、学校、クラス、生徒が「データ構造」タブで被験者として指定されている場合、被験者の組合せのドロップダウン・リストには、「なし」、「学校」、「学校 * クラス」および「学校 * クラス * 生徒 」がオプションとして表示されます。

ランダム効果共分散タイプ。残差に対する共分散構造を指定します。使用できる構造は次のとおりです。

- 1 次自己回帰 (AR1)
- 不均質自己回帰 (ARH1)
- 自己回帰の移動平均 (1,1) (ARMA11)
- 複合対称
- 不均質複合対称 (CSH)
- 対角
- 計測された単位
- Toeplitz
- 非構造化
- 分散成分

重みおよびオフセット: 分析の重み付け。スケール・パラメーターは、応答の分散に関連する推定モデル・パラメーターです。分析の重み付けは、観測ごとに異なる「既知の」値です。「分析の重み付け」フィールドを指定した場合は、応答の分散に関連するスケール・パラメータが、観測ごとに分析の重み付けの値で分割されます。分析の重み値が 0 以下または欠損値のレコードは、分析に使用されません。

オフセット。 オフセット項は、「構造的」な予測フィールドです。その係数はモデルにより推定されませんが、値が 1 であると見なされます。したがって、オフセットの値は単純に対象の線型予測フィールドに追加されます。このことはポアソン回帰モデルでは特に有用であり、各ケースには興味深いイベントへのさまざまな公開レベルがある可能性があります。

例えば、個々のドライバーの事故率をモデリングする場合に、3 年間で過失責任事故が 1 回のドライバーと 25 年間で過失責任事故が 1 回のドライバーの間では、重要な違いがあります。運転手の経験をオフセット項として加味する場合、事故の発生数は対数リンクを持つポアソン応答または負の 2 項応答としてモデル化できます。

分布およびリンクの種類のその他の組み合わせには、オフセット変数のその他の変換が必要です。

一般的な作成オプション: これらの選択は、モデルの作成に使用されるより高度な条件を指定します。

ソート順
: これらのコントロールは、「最後の」カテゴリーを決定するために、対象と因子 (カテゴリー入力) に対するカテゴリーの順序を決定します。対象がカテゴリー型でない場合、またはカスタム参照カテゴリーが「 219 ページの『対象』」設定で指定されている場合、対象のソート順設定は無視されます。

停止規則
: アルゴリズムが実行する反復の最大回数を指定できます。アルゴリズムは、内部ループと外部ルー

プで構成される 2 重の反復プロセスを使用します。反復の最大回数について指定された値は、両方のループに適用されます。負でない整数を指定してください。デフォルトは 100 です。

事後推定設定

これらの設定により、表示のためにモデルの出力がどのように計算されるかが決定されます。

確信度レベル **(%)**

これはモデル係数の区間推定値の計算に使用される信頼度のレベルです。0 より大きく、100 より小さい値を指定します。デフォルトは 95 です。

自由度

有意差検定に対する自由度の計算方法を指定します。サンプル サイズが十分に大きい場合、またはデータが均衡している場合、あるいはモデルが (計測された単位または対角性など) 単純な共分散タイプを使用している場合は、「残差法」を指定します。これがデフォルトの設定です。サンプル サイズが小さい場合、またはデータが不均衡である場合、あるいはモデルが (無構造化など) 複雑な共分散タイプを使用する場合は、「**Satterthwaite** 近似値」を指定します。サンプル サイズが小さく、かつ制限された最尤法 (REML) モデルを使用する場合は、「**Kenward-Roger** 近似値」を選択します。

固定効果および係数の検定

これはパラメーター推定値共分散行列を計算する方法です。モデルの想定に反していると考えられる場合、堅牢な推定量を選択してください。

推定**:** モデル作成アルゴリズムは、内部ループと外部ループで構成される 2 重の反復プロセスを使用します。次の設定は、内部ループに適用されます。

「パラメーター収束」

パラメーター推定値の最大絶対変化または最大相対変化が、指定した値 (負以外でなければなりません) より小さい場合に収束とみなされます。指定した値が 0 の場合、この基準は使用されません。

「対数尤度収束」

対数尤度関数の絶対変化または相対変化が、指定した値 (負以外でなければなりません) より小さい場合に収束とみなされます。指定した値が 0 の場合、この基準は使用されません。

「**Hessian** 収束」

「絶対」を指定した場合は、Hessian に基づく統計量が、指定した値よりも小さい場合に収束とみなされます。「相対」を指定した場合は、指定した値と対数尤度の絶対値の積よりも統計が小さい場合に収束とみなされます。指定した値が 0 の場合、この基準は使用されません。

「**Fisher** スコア法の最大ステップ数」

負でない整数を指定してください。値 0 は、Newton-Raphson 法を指定します。1 以上の値は、反復回数が *n* に達するまで、Fisher スコア法のアルゴリズムを使用し、それ以降は Newton-Raphson 法を使用することを指定します。ここで、*n* は、指定した整数です。

「特異性許容度」

この値は、特異性の確認時に許容度として使用されます。正の値を指定してください。

注**:** デフォルトでは、許容度が 1E-6 の最大「絶対」変化が検査される「パラメータ収束」が使用されます。この設定は、バージョン 22 より前のバージョンで取得される結果とは異なる結果を生成する場合があります。バージョン 22 より前のバージョンの結果を再現するには、「パラメータ収束」基準に「相対」を使用し、デフォルトの許容値を 1E-6 のままにしてください。

一般**:** モデル名: 対象フィールドに基づいて自動的にモデル名を生成するか、またはカスタム名を指定できます。自動的に生成される名前は、対象フィールド名です。複数の対象がある場合、モデル名はそれらのフィールド名が順番にアンパサンドで区切られた形式となります。例えば、対象フィールドが *field1 field2 field3* の場合、モデル名は *field1 & field2 & field3* となります。

スコアリングで使用可能にする：モデルをスコアリングする場合、このグループで選択された項目を作成する必要があります。すべての対象フィールドの予測された値とカテゴリー型対象の確信度は、モデルをスコアリングする場合必ず計算されます。計算される確信度は、予測値の確率 (最も高い予測確率) または最も高い予測確率と 2 番目に高い予測確率との差を基準とする場合があります。

- カテゴリー型対象の予測確率： カテゴリー型対象の予測確率を生成します。カテゴリーごとにフィールドが作成されます。
- フラグ型対象の傾向スコア**:** フラグ型対象フィールド (yes または no の予測を返す) を持つモデルの場合、対象フィールドに指定された真 (true) の結果の尤度を示す傾向スコアが必要な場合があります。モデルは、傾向スコア (調整なし) を生成します。データ区分が有効な場合、モデルは検定データ区分に基づいて、調整済み傾向スコアも生成します。

推定平均**:** このタブでは、因子と因子の相互作用のレベルの推定周辺平均を表示することができます。推定周辺平均 は、多項モデルには使用できません。

項　全体がカテゴリー・フィールドから構成されている固定効果のモデル項が表示されます。モデルに推定周辺平均を生成させる各項を確認します。

対比タイプ
対比フィールドのレベルに使用する対比のタイプを指定します。

なし　対比は生成されません。

ペアごと
指定の因子の全レベルの組み合わせについてペアごとに比較します。これは、交互作用に対して行える唯一の対比です。

偏差　因子の各レベルを全体の平均と比較します。

Simple
因子の各レベル (最後のレベルを除く) を最後のレベルと比較します。「最後のレベル」は、「作成オプション」で指定した因子のソート順によって決まります。これらの対比の種類のすべてが直交であるわけではありません。

対比フィールド
選択した対比タイプを使用して、1 つの因子 (比較されるレベル) を指定します。「なし」 が対比の種類に選択されている場合、対比フィールドは選択できません (または必要ありません)。

連続型フィールド
連続型フィールドのリストは、連続型フィールドを使用する固定効果の項から抽出されます。推定周辺平均を計算する場合、共変量は指定された値に固定されます。平均値を選択するか、カスタム値を指定します。

次を使用して多重比較の調整
複数の対比で仮説検定を実行すると、全体の有意水準は、含められた対比の有意水準から調整されます。これにより、調整方法を選択できます。

最小有意差
この方法では、線型対比のいくつかが帰無仮説の値とは異なるという仮説を拒否する全体的な確率は制御されません。

Sequential Bonferroni (逐次 Bonferroni)
個々の仮説を棄却する点であまり保守的ではないが、同じ全体の有意水準を維持する逐次ステップダウン棄却 Bonferroni 手続き。

Sequential Sidak (逐次 Sidak)
個々の仮説を棄却する点であまり保守的ではないが、同じ全体の有意水準を維持する逐次ステップダウン棄却 Sidak 手続き。

最小有意差法は、sequential Bonferroni 法より保守的でない sequential Sidak 法より保守的でありません。つまり、最小有意差法は、少なくとも sequential Bonferroni と同じ仮設数を拒否する sequential Sidak と同じ仮設数を拒否します。

次についての推定平均を表示する
対象の元のスケール、あるいはリンク関数変換のどちらに基づいて推定周辺平均を計算するかを指定します。

元の目標スケール
対象の推定周辺平均を計算します。対象がイベント/試行のオプションを使用して指定された場合、イベントの数ではなくイベント/試行の割合の推定周辺平均を計算します。

リンク関数の変換
線形予測値の推定周辺平均を計算します。

モデル・ビュー**:** デフォルトでは、「モデル要約」ビューが表示されます。別のモデル・ビューを表示するには、ビューのサムネイルから選択します。

モデル要約*:* このビューはスナップショットで、モデルとその適合度についての要約が一目でわかります。

テーブルこのテーブルは、『対象の設定』で指定された対象、確率分布、リンク関数を識別します。対象が、イベントや試行によって定義されている場合、セルはイベントのフィールドと試行のフィールド試行の固定数を表示できるよう分割されます。また、有限のサンプルの補正赤池情報量基準 (AICC) およびベイズ情報量基準 (BIC) が表示されます。

- 補正赤池 : -2 (制限) 対数尤度に基づいて混合モデルを選択し、比較するための指標。値が小さいほどモデルが良好であることを示します。AICC は、小さな標本サイズに適応するように AIC を「修正」します。標本サイズが大きくなるに従い、AICC は AIC に収束します。
- ベイジアン : -2 対数尤度に基づいてモデルを選択し、比較するための指標。値が小さいほどモデルが良好であることを示します。BIC もパラメータが過多のモデル (例えば、大量の入力がある複雑なモデル) にペナルティを科しますが、AIC よりも厳密にそれを行います。

グラフ。対象がカテゴリーの場合、グラフには最終のモデルの精度が表示されます。これは正確な分類の割合です。

データ構造*:* このビューでは、指定されたデータ構造の概要を提供し、被験者と反復測定が正しく指定されていることを確認するのに役立ちます。最初の被験者の観測情報は、各被験者フィールドと反復測定のフィールド、および対象ごとに表示されます。さらに、それぞれの被験者フィールドと反復測定フィールドのレベルの数が表示されます。

予測対観測*:* イベント/試行として指定された対象を含む連続型対象について、縦軸に予測値を、横軸に観測値を示す分割散布図を表示します。点は 45 度の線にあるのが理想です。このビューはレコードがモデルによって特に不正に予測されているかどうかを示します。

分類*:* カテゴリー型対象の場合、観測値と予測値のクロス分類と、すべての正分類パーセントをヒート・マップに表示します。

テーブルのスタイル。さまざまな表示スタイルがあり、「スタイル」ドロップダウン・リストからアクセスできます。

- 行パーセント。セルの行パーセント (行合計のパーセントで表現されるセルの度数) が表示されます。これがデフォルトです。
- セルの度数。セルのセル度数が表示されます。ヒート・マップの色の濃さは、行パーセントに基づいています。
- ヒート・マップ。セルの値は表示しません。陰影付けのみ表示します。
- 圧縮。セルの行または列のヘッダー、セルの値を表示しません。この方法は、対象にカテゴリー数が多い場合に役立ちます。

欠損値。対象に欠損値があるレコードがある場合、レコードはすべての有効な行の下の 「**(**欠損値**)**」行に表示されます。欠損値のあるレコードは、すべての正分類パーセントには貢献しません。

複数の対象。複数のカテゴリー対象がある場合、各対象は別々のテーブルに表示され、「対象」ドロップダウン・リストを使用して表示する対象を制御します。

大型テーブル。表示する対象に 100 を超えるカテゴリーがある場合、テーブルは表示されません。

固定効果*:* このビューには、モデルの各固定効果のサイズが表示されます。

スタイル: さまざまな表示スタイルがあり、「スタイル」ドロップダウン・リストから選択できます。

- ダイアグラム。「固定効果」設定で指定された順番で先頭から最後まで効果がソートされたグラフです。ダイアグラムで繋がった線は、効果の有意確率に基づいて重みがつけられます。線の太いほど効果の有意確率は大きくなります (*p* 値は小さくなります)。これがデフォルトです。
- テーブル: これは、モデル全体の効果および個別のモデル効果を示す分散分析テーブルです。各効果は、固定効果の設定で指定された順序で上から下にソートされているチャートです。

有意確率。「有意」スライダーを使用して、ビューで表示する効果を制御します。有意確率の値がスライダーの値より大きい効果は表示されません。このスライダーを使用してもモデルは変更されませんが、最も重要な効果に焦点を当てることができます。デフォルトでは値が 1.00 になるため、有意確率に基づいてフィルタリング処理される効果はありません。

固定係数*:* このビューには、モデルの各固定係数の値が表示されます。因子 (カテゴリー型予測フィールド) はモデル内で指標コード化されるため、因子を含む効果には通常複数の関連する係数があります。一方は冗長係数に対応するカテゴリーを除いたものとなります。

スタイル: さまざまな表示スタイルがあり、「スタイル」ドロップダウン・リストから選択できます。

- ダイアグラムこれは、最初に定数項を表示し、次に「固定効果」の設定で指定された順序に従って上から下に効果をソートしたグラフです。因子を含む効果内で、係数はデータ値の昇順でソートされます。ダイアグラムで繋がった線は色分けされ、係数の有意確率に基づいて重みがつけられます。線の太いほど係数の有意確率は大きくなります (*p* 値は小さくなります)。これはデフォルトのスタイルです。
- テーブル: 各モデル係数の値、有意差検定、および信頼区間が表示されます。定数項の後、効果が固定効果の設定で指定された順序で上から下にソートされているチャートです。因子を含む効果内で、係数はデータ値が小さい順に並べ替えられます。

多項。多項分布が効果にある場合、「多項」ドロップダウン・リストを使用して表示する対象カテゴリーを制御します。リスト内の値のソート順は、「作成オプション」の設定の仕様によって決定されます。

指数。2 値ロジスティック回帰 (2 項分布およびロジット・リンク)、名義ロジスティック回帰 (多項分布およびロジット・リンク)、負の 2 項回帰 (負の 2 項分布および対数リンク)、対数線型モデル (ポワソン分布および対数リンク) など、特定のモデル・タイプに対する指数係数推定値と信頼区間を表示します。

有意確率。「有意」スライダーを使用して、ビューに表示する係数を制御します。有意確率の値がスライダーの値より大きい係数は表示されません。このスライダーを使用してもモデルは変更されませんが、最も重要な係数に焦点を当てることができます。デフォルトでは値が 1.00 になるため、有意確率に基づいてフィルタリングされる係数はありません。

変量効果共分散*:* 変量効果共分散行列 (**G**) を表示します。

スタイル: さまざまな表示スタイルがあり、「スタイル」ドロップダウン・リストから選択できます。

- 共分散値。「固定効果」設定で指定された順番で先頭から最後まで効果がソートされた、共分散行列のヒート・マップです。corrgram の色は、キーに示されているセルの値に対応します。これはデフォルトです。
- **Corrgram**。共分散行列のヒート・マップです。
- 圧縮。これは行と列のヘッダーを除いた共分散行列のヒート・マップです。

ブロック。複数のランダム効果ブロックがある場合は、「ブロック」ドロップダウン・リストから表示するブロックを選択します。

グループ: 変量効果のブロックにグループの指定がある場合は、「グループ」ドロップダウン・リストで表示するグループ・レベルを選択します。

多項。多項分布が効果にある場合、「多項」ドロップダウン・リストを使用して表示する対象カテゴリーを制御します。リスト内の値のソート順は、「作成オプション」の設定の仕様によって決定されます。

共分散パラメーター*:* このビューには、残差とランダム効果の共分散パラメーターの推定値と関連する統計情報が表示されます。これらは高度ですが、共分散構造が適しているかどうかに関する情報を提供する基本的な結果です。

要約表。これは、残差 (**R**) およびランダム効果 (**G**) 共分散行列のパラメーター数、固定効果 (**X**) およびランダム効果 (**Z**) デザイン行列のランク (列数)、データ構造を定義する被験者フィールドにより定義される被験者数に関するクイック・リファレンスです。

共分散パラメーター・テーブル。選択した効果について、推定値、標準誤差、信頼区間が各共分散パラメーターに表示されます。表示されるパラメーターの数は、効果および変量効果ブロックの共分散構造、ブロックの効果の数によって異なります。非対角パラメーターが重大ではないことが表示された場合、単純な共分散構造を使用することができる場合があります。

効果。ランダム効果ブロックがある場合は、「効果」ドロップダウン・リストで表示する残差またはランダム効果ブロックを選択します。残差効果は常に使用可能です。

グループ: 残差または変量効果のブロックにグループの指定がある場合は、「グループ」ドロップダウン・リストで表示するグループ・レベルを選択します。

多項。多項分布が効果にある場合、「多項」ドロップダウン・リストを使用して表示する対象カテゴリーを制御します。リスト内の値のソート順は、「作成オプション」の設定の仕様によって決定されます。

推定平均*:* 有意な効果*:* 3 要因の交互作用から始まり、2 要因の交互作用、および最終的に主効果となる、10 個の「最も有意な」固定全因子効果について表示されるグラフです。グラフは横軸上の主効果（または交互作用で最初に表示されている効果）の各値について縦軸上の対象のモデル推定値を表示します。相互作用の 2 番目に記載されている効果の各値について別の線が生成されます、3 要因の交互作用で 3 番目に表示される効果の値ごとグラフが表示されます。他のすべての予測値は一定です。対象フィールドに対する各予測フィールドの係数の効果について、役立つ視覚化を提供します。予測値が重要でない場合、推定平均値は生成されません。

信頼度。これは、「作成オプション」の一部として指定された信頼度レベルを使用して、周辺平均に対する信頼限界の上限と下限を表示します。

推定平均*:* カスタム効果*:* ユーザが要求したすべての因子の効果を固定するためのテーブルおよびグラフです。

スタイル: さまざまな表示スタイルがあり、「スタイル」ドロップダウン・リストから選択できます。

- ダイアグラムこのスタイルの場合は、横軸の主効果 (または交互作用で最初にリストされている効果) のそれぞれの値について縦軸の対象のモデル推定値が折れ線グラフで表示されます。交互作用の 2 番目にリストされている効果のそれぞれの値について別の線が描画されます。3 要因の交互作用で 3 番目にリストされている効果のそれぞれの値について別のグラフが作成されます。それ以外の予測変数はすべて一定に保たれます。

  対比が要求された場合、別のグラフが、対比のフィールドのレベルを比較するために表示されます。相互作用の場合、グラフが対比フィールド以外の効果の各レベルの組み合わせで表示されます。ペアごとの対比の場合、距離のネットワーク グラフです。つまり、ネットワーク内のノード間の距離は、サンプル間の差異に対応する比較のテーブルをグラフィカルに表現したものです。黄色の線は、統計的に有意な差に対応し、黒線が非有意差に対応しています。ネットワーク内の線の上にマウスを乗せると、線で接続されたノード間の相違の調整済み有意度とツールヒントが表示されます。

  偏差の対比の場合、棒グラフに縦軸上のターゲットのモデル推定値と横軸に対比のフィールドの値が表示されます。相互作用の場合、グラフは対比フィールド以外の効果の各水準の組み合わせごとに表示されます。バーは、対比のフィールドのレベルと全体の平均値との差を示し、黒の水平線で表されます。

  単純な対比の場合、棒グラフに縦軸上のターゲットのモデル推定値と横軸に対比のフィールドの値が表示されます。相互作用の場合、グラフは対比フィールド以外の効果の各水準の組み合わせごとに表示されます。バーは、対比のフィールド (最後を除く) と最後のレベルとの間の差を示し、黒の水平線で表されます。
- テーブルこのスタイルの場合は、対象のモデル推定値、その標準誤差、および効果のフィールドの各レベルの組み合わせの信頼区間の表が表示されます。それ以外の予測変数はすべて一定に保たれます。

  対比が要求された場合、別のテーブルが推定、標準誤差、有意差検定、およびそれぞれのコントラストのための信頼区間とともに表示されます。相互作用の場合、対比フィールド以外の効果の各レベルの組み合わせの行が別途表示されます。さらに、全体的なテスト結果を含むテーブルが表示されます。相互作用の場合、対比のフィールド以外の効果の各レベルの組み合わせごとに全体的な検定があります。

信頼度。これは、「作成オプション」の一部として指定された信頼度レベルを使用して、周辺平均に対する信頼限界の上限と下限の表示を切り替えます。

レイアウト。これは、ペアワイズ対比ダイアグラムのレイアウトを切り替えます。サークル レイアウトは、ネットワーク レイアウトよりも対比は少ないですが、行の重複を回避できます。

設定**:** モデルをスコアリングする場合、このタブで選択された項目を作成する必要があります。すべての対象フィールドの予測された値とカテゴリー型対象の確信度は、モデルをスコアリングする場合必ず計算されます。計算される確信度は、予測値の確率 (最も高い予測確率) または最も高い予測確率と 2 番目に高い予測確率との差を基準とする場合があります。

- カテゴリー型対象の予測確率： カテゴリー型対象の予測確率を生成します。カテゴリーごとにフィールドが作成されます。
- フラグ型対象の傾向スコア**:** フラグ型対象フィールド (yes または no の予測を返す) を持つモデルの場合、対象フィールドに指定された真 (true) の結果の尤度を示す傾向スコアが必要な場合があります。モデルは、傾向スコア (調整なし) を生成します。データ区分が有効な場合、モデルは検定データ区分に基づいて、調整済み傾向スコアも生成します。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

GLE ノード

GLE モデルは、指定したリンク関数を介して因子および共変量に線型に関連する従属変数を識別します。さらにこのモデルでは、非正規分布の従属変数を使用することができます。一般化線型モデルは、正規分布した回答の線型、バイナリ・データのためのロジスティック・モデル、カウント・データのための対数線型モデル、間隔を決めて検閲される延命データのための補対数-対数モデルなどの広く使用される統計モデルに加えて、一般的なモデルの定式を通じて多くのほかの統計モデルも対象とします。

例: 海運会社では一般化線型モデルを使用して、異なる期間に建設された複数の種類の船の損害数にポアソン回帰を当てはめることができ、構築されたモデルによって損害を受けやすい船の種類を判断することができます。

自動車保険会社では一般化線型モデルを使用して、自動車に対する損害請求にガンマ回帰を当てはめることができ、構築されたモデルによって請求に最も寄与する因子を判断することができます。

医療研究者は、一般化線型モデルを使用して区間打ち切り生存率データに補ログ マイナス・ログを当てはめ、病状が再発する時間を予測します。

GLE モデルは、入力フィールドの値を出力フィールドの値に関係付ける方程式を作成することで機能します。モデルを生成した後は、そのモデルを使用して新しいデータの値を推定できます。

カテゴリ型対象の場合、レコードごとに、各出力カテゴリ候補の所属確率が算出されます。最も確率の高い対象カテゴリーが、そのレコードの予測出力値として割り当てられます。

要件: 1 つ以上の入力フィールドと、2 つ以上のカテゴリを含む 1 つの対象フィールド (測定の尺度が連続型、カテゴリ型、またはフラグ型 のフィールド) が必要です。モデルで使用するフィールド・タイプは、完全にインスタンス化する必要があります。

ターゲット

これらの設定は、リンク関数を介してターゲット、その分布、および予測因子との関係を定義します。

対象: 対象は必須です。対象には任意の尺度を設定でき、対象の尺度によって適切な分布とリンク関数が変わります。

- 定義済みの対象を使用: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の対象設定を使用するには、このオプションを選択します。
- カスタム対象を使用: 対象を手動で割り当てる場合は、このオプションを選択します。
- 分母に試行数を使用: 対象の応答が一連の試行内で発生するいくつかのイベントである場合、対象フィールドにはイベント数が含まれます。また試行回数を含んでいる追加フィールドを選択できます。例えば、新しい農薬をテストするときは、さまざまな濃度の農薬をアリのサンプルに噴霧して死んだアリの数と各サンプルのアリの数を記録します。この場合、死んだアリの数を記録するフィールドは、対象（イベント）フィールドとして指定されなければならず、各サンプル中のアリの数を記録するフィールドは、試行フィールドとして指定する必要があります。アリの数は、各サンプルに対して同じである場合、試行回数は、固定値を使用して指定することができます。

  試行回数は、各レコードのイベント数以上である必要があります。また、イベント数は非負整数、試行数は正の整数であることが必要です。
- 参照カテゴリをカスタマイズ: カテゴリー対象に参照カテゴリーを選択できます。このことでパラメーター推定値などの一定の出力に影響を与えることができますが、モデルの適合度を変更してはなりません。例えば、対象値がデフォルトで0、1、および 2 となる場合、手順では最後の（最も高い値を持つ）カテゴリー、または 2 を参照のカテゴリーにします。この場合、パラメーター推定はカテゴリー 2 の尤度に相対 してカテゴリー 0 または 1 の尤度に関連すると解釈されます。カスタム カテゴリーを指定して、対象がラベルを定義している場合、リストから値を選択して参照カテゴリーを設定することができます。これは、モデル指定の途中で特定のフィールドがどのようにコーディングされたか正確にわからないときに便利です。

目標分布と線型モデルとのリレーション **(**リンク**)**: 予測値の値を指定することで、モデルは指定した形状に従う対象値の分布、および指定したリンク関数を使用して予測値と線型に関連する対象値を予期します。いくつかの共通モデルへのショートカットが提供されます。または、ショートカットのリストにない分布とリンク関数の特定の組み合わせを使用する場合は、「カスタム」設定を選択します。

- 線型モデル: 同一リンクを持つ正規分布を指定します。これは、線型または分散分析モデルを使用して対象が予測される際に有用です。
- ガンマ回帰: 対数リンクを持つガンマ分布を指定します。これは、対象に含まれる値がすべて正の値で、値が大きくなるほどゆがむ場合に使用されます。
- 対数線型: 対数リンクを持つポアソン分布を指定します。これは、対象が一定期間内の出現回数を表すときに使用されます。
- 負の二項回帰: 対数リンクを持つ負の二項分布を指定します。これは、対象と分母が *k* の成功を観測するために必要な試行回数を表すときに使用されます。
- **Tweedie** 回帰: 恒等式、対数、またはべき乗のリンク関数を使用する Tweedie 分布を指定します。これは、ゼロと正の実数値が混在する応答をモデル化する場合に役立ちます。これらの分布は、複合ポアソン 分布、複合ガンマ 分布、およびポアソン ガンマ 分布とも呼ばれます。

- 多項ロジスティック回帰: 多項分布を指定します。これは対象が複数カテゴリーの応答である場合に使用されます。累積ロジットリンク（順序型結果）または一般化ロジット・リンク（マルチカテゴリー名義型回答）を使用します。
- 二項ロジスティック回帰: ロジット・リンクを持つ二項分布を指定します。これは対象がロジスティック回帰モデルで予測される 2 値応答である場合に使用されます。
- **2** 値プロビット: プロビット・リンクを持つ二項分布を指定します。これは対象が基礎の正規分布を使用した 2 値応答である場合に使用されます。
- 調査された生存推定値間隔: 補対数-対数リンクを持つ二項分布を指定します。これは終了イベントのない観測がある場合の生存分析で有用です。
- カスタム: 分布とリンク関数の組み合わせを独自に指定します。

分布

これを選択すると、対象の分布が指定されます。非正規分布と非同一リンク関数を指定する機能は、線型モデルから一般化線型モデルに移行する上での重要な改善点です。分布とリンク関数には多くの組み合わせの可能性があり、その中のいくつかは特定のデータセットに適切な場合があるので、この選択が先験的な理論考察または一番適合するように見える組み合わせによって導き出される可能性があります。

- 自動: 使用すべき分布がわからない場合はこのオプションを選択します。ノードによってデータが分析され、最適な分布方式が推定されて適用されます。
- 二項分布: この分布は 2 値応答またはイベント数を表す対象にのみ適しています。
- ガンマ分布: この分布は、正の値が大きくなるほどゆがむ正のスケール値を持つ対象に適しています。データの値が 0 以下または欠損している場合は、対応するケースが分析に使用されません。
- 逆ガウス分布: この分布は、正の値が大きくなるほどゆがむ正のスケール値を持つ対象に適しています。データの値が 0 以下または欠損している場合は、対応するケースが分析に使用されません。
- 多項分布: この分布は、複数カテゴリーの応答を表す対象に適しています。モデルの形式は、対象の尺度によって異なります。

  名義型対象は、モデル パラメーターの異なるセットが、（参照カテゴリーを除く）対象のカテゴリーごとに推定される名義型多項モデルとなります。特定の予測のためのパラメーター推定値は、参照カテゴリーに相対的に、予測値と対象の各カテゴリーの尤度との関連性を示します。

  順序型対象は、従来の切片項が、対象カテゴリーの累積確率に関連する閾値パラメーターのセットと置き換えられる順序型多項モデルとなります。
- 負の二項分布: 負の二項回帰では、対数リンクを含む負の二項分布を使用します。対象が高い分散度を持つ出現回数を示す場合に使用する必要があります。
- 正規分布: これは、中心 (平均) 値の周りで値が対称に、ベル型の分布になる連続型対象に適しています。
- ポアソン分布: この分布は一定期間の対象のイベントの発生回数として考えることができ、負ではない整数値の変数に適しています。データの値が非正数、0 より小さい、または欠損している場合は、対応するケースが分析に使用されません。
- **Tweedie**: この分布はガンマ分布のポアソン混合によって表すことができる変数に適しています。分布の「混合」とは、連続型分布 (負でない実数値) と離散型分布 (単一値 0 の 正の確率質量) のプロパティーを結合することです。従属変数は 0 またはそれ以上のデータ値を持った数値である必要があります。データの値が 0 より小さい、または欠損している場合は、対応するケースが分析に使用されません。Tweedie 分布のパラメーターの固定値は 1 以上 2 以下のどんな数字でもかまいません。

リンク関数

リンク関数は、モデルを推定できるようにする対象の変換の 1 つです。使用できる関数は次のとおりです。

- 自動: 使用すべきリンクがわからない場合はこのオプションを選択します。ノードによってデータが分析され、最適なリンク関数が推定されて適用されます。
- 同一: $f(x)=x$。対象は変換されません。このリンクは、多項分布を除き、どの分布でも共に使用できます。
- 補対数-対数: $f(x)=\log(-\log(1-x))$。これは、二項分布とまたは多項分布とのみ使用するのが適しています。
- コーチット: $f(x) = \tan(\pi\ (x - 0.5))$。これは、二項分布とまたは多項分布とのみ使用するのが適しています。
- 対数: $f(x)=\log(x)$。このリンクは、多項分布を除き、どの分布でも共に使用できます。
- 対数-補数: $f(x)=\log(1-x)$。これは、二項分布とのみ使用するのが適しています。
- ロジット: $f(x)=\log(x\ /\ (1-x))$。これは、二項分布とまたは多項分布とのみ使用するのが適しています。
- 負の対数-対数: $f(x)=-\log(-\log(x))$。これは、二項分布とまたは多項分布とのみ使用するのが適しています。
- プロビット: $f(x)=\Phi^{-1}(x)$。Φ^{-1} は、累積標準正規分布関数の逆関数です。これは、二項分布とまたは多項分布とのみ使用するのが適しています。
- べき乗: $f(x)=x^{\alpha}$ ($\alpha \neq 0$ の場合)。 $f(x)=\log(x)$ (α が 0 の場合)。α には数値を指定する必要があり、その数値は実数である必要があります。このリンクは、多項分布を除き、どの分布でも共に使用できます。

Tweedie のパラメータ: 「**Tweedie** 回帰」ラジオ ボタンを選択した場合、または「分布」方法として「Tweedie」を選択した場合にのみ使用できます。1 と 2 の間の値を選択します。

モデル効果

固定効果の因子は、一般的に、その関心の値がすべてのデータセットで表現されるフィールドとして考えられ、スコアリングに使用することができます。デフォルトでは、ダイアログ内の他の場所で指定されていない定義済みの入力の役割を持つフィールドは、モデルの固定効果部分に入力されます。カテゴリー型 (フラグ型、名義型、順序型) フィールドは、モデルでの因子として使用され、連続型フィールドは共変量として使用されます。

ソース・リスト内の1つ以上のフィールドを選択し、効果リストにドラッグして、効果をモデルに入力します。作成する効果の種類は、選択項目をドロップするホットスポットによって異なります。

- メイン: ドロップされたフィールドが、別の主効果として効果リストの下部に表示されます。
- **2** 要因: ドロップされたフィールドのすべての可能なペアが、2 要因の交互作用として効果リストの下部に表示されます。
- **3** 要因: ドロップされたフィールドのすべての可能なトリプレットが、3 要因の交互作用として効果リストの下部に表示されます。
- *: すべてのドロップされたフィールドの組み合わせが、1 つの交互作用として効果リストの下部に表示されます。

効果ビルダーの右側にあるボタンを使用すると、さまざまな操作を行うことができます。

表 12. 効果ビルダー・ボタンの説明

アイコン	説明
	削除したい条件を選択し、削除ボタンをクリックして、固定効果モデルから用語を削除します。
	順序を変更する条件を選択し、上向きまたは下向きの矢印をクリックして、固定効果モデル内の項目を並べ替えます。
	「カスタム項の追加」ボタンをクリックし、「カスタム項の追加」ダイアログを使用して、入れ子になった項をモデルに追加します。

定数項を含める: 通常、切片はモデルに含まれます。データが原点を通ると仮定できる場合は、切片を除外できます。

カスタム項の追加

この手順で使用するモデルの入れ子になった項を構築することができます。入れ子になった項は、値が別の因子の水準と相互作用しない因子または共変量の影響をモデル化するのに便利です。例えば、食料品店チェーンは、複数の店舗の場所で、顧客の支出の習慣に従う場合があります。各顧客は、それらの店舗のうち 1 つの店舗にのみ頻繁に通うため、顧客効果は店舗の場所効果内にネストされている ということができます。

また、そのような同一の共変量を含む多項式の項のような相互作用の効果を含める、または入れ子になった項に複数レベルの入れ子を追加することができます。

制限。ネスト項目には、次のような制限があります。

- 交互作用内のすべての因子は固有である必要があります。したがって、*A* が因子の場合、*A*A* の指定は無効です。
- 1 つのネスト効果内の因子はすべて、固有のものである必要があります。したがって、*A* が因子の場合、*A(A)* の指定は無効です。
- 共変量内に効果を入れ子にすることはできません。*A* が因子であり、*X* が共変量である場合、*A(X)* を指定しても無効になります。

入れ子になった項の構築

1. もう一つの因子内に入れ子になっている要因および共変量を選択し、矢印ボタンをクリックします。
2. 「**(**内**)**」 をクリックします。
3. 前の因子または共変量が入れ子になっている因子を選択し、矢印のボタンをクリックします。
4. 「項目を追加」 をクリックします。

また、相互作用の効果を含める、または入れ子になった項に複数レベルの入れ子を追加することができます。

重みとオフセット

分析の重み付け: スケール・パラメーターは、応答の分散に関連する推定モデル・パラメーターです。分析の重み付けは、観測ごとに異なる「既知の」値です。「分析の重み付け」フィールドを指定した場合は、応答の分散に関連するスケール・パラメータが、観測ごとに分析の重み付けの値で分割されます。分析の重み付けの値が 0 以下であるか、または欠損しているレコードは、分析で使用されません。

オフセット: オフセット項は、「構造的」な予測値です。その係数はモデルにより推定されませんが、値が 1 であると見なされます。したがって、オフセットの値は単純に対象の線型予測フィールドに追加されます。このことはポアソン回帰モデルでは特に有用であり、各ケースには興味深いイベントへのさまざまな公開レベルがある可能性があります。

例えば、個々のドライバーの事故率をモデリングする場合、3 年間で過失責任事故が 1 回のドライバーと、25 年間で過失責任事故が 1 回のドライバーの間には、重要な差があります。運転手の経験をオフセット項として加味する場合、事故の発生数は対数リンクを持つポアソン応答または負の 2 項応答としてモデル化できます。

分布およびリンクの種類のその他の組み合わせには、オフセット変数のその他の変換が必要です。

作成オプション

これらの選択は、モデルの作成に使用されるより高度な条件を指定します。

ソート順: これらのコントロールは、「最後の」カテゴリーを決定するために、対象と因子 (カテゴリー入力) に対するカテゴリーの順序を決定します。対象がカテゴリー型でない場合、またはカスタム参照カテゴリーが「 232 ページの『ターゲット』」設定で指定されている場合、対象のソート順設定は無視されます。

事後推定設定: これらの設定により、表示のためにモデルの出力がどのように計算されるかが決定されます。

- 信頼係数 **%**: これはモデル係数の区間推定値の計算に使用される信頼度のレベルです。 0 より大きく、100 より小さい値を指定します。デフォルトは 95 です。
- 自由度: 有意確率の検定に対する自由度の計算方法を指定します。サンプルサイズが十分大きい場合、またはデータが均衡である場合、またはモデルが尺度化識別または対角性など単純な共変量タイプを使用する場合、「すべての検定に固定 **(**残差方法**)**」 を指定します。これはデフォルトです。サンプルサイズが小さい場合、またはデータが不均衡である場合、またはモデルが非構造化など複雑な共変量タイプを使用する場合、「検定全体で異なる **(Satterthwaite** 近似値**)**」 を指定します。
- 固定効果および係数の検定: これはパラメーター推定値共分散行列を計算する方法です。モデルの想定に反していると考えられる場合、堅牢な推定量を選択してください。

影響がある外れ値を検出する: このオプションは、多項分布を除くすべての分布について、影響がある外れ値を特定する場合に選択します。

トレンド分析を実行する: このオプションは、分布図についてトレンド分析を実行する場合に選択します。

推定

方法: 使用する最尤推定方法を選択します。選択可能なオプションは次のとおりです。

- Fisher スコア法
- Newton-Raphson
- Hybrid

最大 **Fisher** 反復回数 **(Maximum Fisher iterations)**: 負でない整数を指定します。値 0 は、Newton-Raphson 法を指定します。1 以上の値は、反復回数が *n* に達するまで、Fisher スコア法のアルゴリズムを使用し、それ以降は Newton-Raphson 法を使用することを指定します。ここで、*n* は、指定した整数です。

スケール パラメータ方法: スケール パラメータの推定方法を選択します。選択可能なオプションは次のとおりです。

- 最尤推定
- 固定値。使用される値も設定します。
- 逸脱
- Pearson カイ 2 乗

負の二項分布法: 負の二項分布補助パラメータの推定方法を選択します。選択可能なオプションは次のとおりです。

- 最尤推定
- 固定値。使用される値も設定します。

パラメータ収束 パラメーター推定値の最大絶対変化または最大相対変化が、指定した値 (負以外でなければなりません) より小さい場合に収束とみなされます。指定した値が 0 の場合、この基準は使用されません。

対数尤度収束 対数尤度関数の絶対変化または相対変化が、指定した値 (負以外でなければなりません) より小さい場合に収束とみなされます。指定した値が 0 の場合、この基準は使用されません。

Hessian 収束 「絶対」を指定した場合は、Hessian に基づく統計量が、指定した値よりも小さい場合に収束とみなされます。「相対」を指定した場合は、指定した値と対数尤度の絶対値の積よりも統計が小さい場合に収束とみなされます。指定した値が 0 の場合、この基準は使用されません。

最大反復回数: アルゴリズムが実行する反復の最大回数を指定できます。アルゴリズムは、内部ループと外部ループで構成される 2 重の反復プロセスを使用します。反復の最大回数について指定された値は、両方のループに適用されます。負でない整数を指定してください。デフォルトは 100 です。

特異性許容度: この値は、特異性の確認時に許容度として使用されます。正の値を指定してください。

注**:** デフォルトでは、許容度が 1E-6 の最大「絶対」変化が検査される「パラメータ収束」が使用されます。この設定は、バージョン 17 より前のバージョンで取得される結果とは異なる結果を生成する場合があります。バージョン 17 より前のバージョンの結果を再現するには、「パラメータ収束」基準に「相対」を使用し、デフォルトの許容値を 1E-6 のままにしてください。

モデルの選択

モデル選択または正規化を使用する: このペインのコントロールをアクティブ化するには、このチェックボックスを選択します。

方法: モデル選択方法を選択するか、(「リッジ」を使用する場合は) 使用する正規化を選択します。以下のオプションから選択できます。

- **Lasso**: L1 正規化とも呼ばれます。大量の予測値がある場合は、変数増加ステップワイズ法よりもこの方法の方が高速です。この方法では、パラメータを小さくする (つまりペナルティを課す) ことでオーバーフィットを防止します。この方法では、一部のパラメータをゼロまで小さくして、変数選択 Lasso を実行できます。

- リッジ: L2 正規化とも呼ばれるこの方法では、パラメータを小さくする (つまりペナルティを課す) ことでオーバーフィットを防止します。この方法では、すべてのパラメータを同じ割合で小さくしますが、none は除外されます。この方法は変数選択方法ではありません。
- **Elastic Net**: L1 + L2 正規化とも呼ばれるこの方法では、パラメータを小さくする (つまりペナルティを課す) ことでオーバーフィットを防止します。この方法では、一部のパラメータをゼロまで小さくして、変数選択を実行できます。
- 変数増加ステップワイズ法: モデルに効果がない状態から、これ以上追加または削除できなくなるまで、ステップワイズ法の基準に従って徐々に効果を追加および削除します。

2 要因の交互作用を自動的に検出する: 2 要因の交互作用を自動的に検出するには、このオプションを選択します。

ペナルティ パラメータ

以下のオプションは、「方法」で「Lasso」または「Elastic Net」を選択した場合にのみ選択できます。

自動的にペナルティ パラメータを選択する: 設定するパラメータ ペナルティがわからない場合は、このチェック ボックスを選択します。これにより、ノードによってペナルティが特定されて適用されます。

Lasso ペナルティ パラメータ: Lasso モデル選択方法で使用されるペナルティ パラメータを入力します。

Elastic Net ペナルティ パラメータ **1**: Elastic Net モデル選択方法で使用される L1 ペナルティ パラメータを入力します。

Elastic Net ペナルティ パラメータ **2**: Elastic Net モデル選択方法で使用される L2 ペナルティ パラメータを入力します。

変数増加ステップワイズ法

以下のオプションは、「方法」で「変数増加ステップワイズ法」を選択した場合にのみ使用できます。

次の値以上の **p** 値の効果を含む: 効果が計算に算入される条件となる効果の最小確率値を指定します。

次の値より大きい **p** 値の効果を削除する: 効果が計算に算入される条件となる効果の最大確率値を指定します。

最終モデルの効果の最大数をカスタマイズする: 「効果の最大数」オプションをアクティブ化するには、このチェック ボックスを選択します。

効果の最大数: 変数増加ステップワイズ法を作成方法として使用する場合の、効果の最大数を指定します。

最大ステップ数をカスタマイズする: 「ステップの最大数」オプションをアクティブ化するには、このチェック ボックスを選択します。

ステップの最大数: 変数増加ステップワイズ法を作成方法として使用する場合の、ステップの最大数を指定します。

モデル・オプション

モデル名: 対象フィールドに基づいて自動的にモデル名を生成することも、カスタム名を指定することもできます。自動的に生成される名前は、対象フィールド名です。複数の対象がある場合、モデル名はそれらの

フィールド名が順番にアンパサンドで区切られた形式となります。例えば、対象フィールドが field1、field2、および field3 の場合、モデル名は *field1 & field2 & field3* となります。

予測変数の重要度を計算: 重要度の適切な測定基準を作成するモデルの場合、モデル推定時に各予測値の相対重要度を示すグラフを表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。モデルによっては、特に大きなデータセットを使用する場合、予測変数の重要度の計算に時間がかかることがあります。そのため、一部のモデルではデフォルトでオフになっています。

詳しくは、 44 ページの『予測変数の重要度』を参照してください。

GLE モデル ナゲット

GLE モデル ナゲットの出力

GLE モデルを作成すると、以下の情報が出力に表示されます。

「モデル情報」テーブル

「モデル情報」テーブルは、モデルについての重要な情報を提供します。テーブルは次のようなハイレベルなモデル設定を特定します。

- データ型ノードまたは GLE ノードのいずれかの「フィールド」タブで選択された対象フィールドの名前。
- モデル化された対象カテゴリ パーセンテージと参照対象カテゴリ パーセンテージ。
- 確率分布と関連するリンク関数。
- 使用するモデル作成方法。
- 入力された予測値の数と最終モデルでの数。
- 分類精度の割合。
- モデル タイプ。
- モデルの精度のパーセンテージ (対象が連続型の場合)。

レコード要約

集計表には、モデルを適合させるために使用されたレコード数と除外されたレコード数が表示されます。詳細として、包含されたレコードと除外されたレコードの数と割合、および重み付きなしの数値 (度数の重みを使用した場合) が表示されます。

予測変数の重要度

「予測変数の重要度」グラフは、モデル内の上位 10 個の入力 (予測値) の重要度を棒グラフとして表示します。

グラフ内に 10 個を超えるフィールドがある場合は、グラフの下のスライダーを使用して、グラフ内に含まれる予測値の選択を変更できます。スライダー上のインディケーター・マークは固定幅であり、スライダー上の各マークは 10 個のフィールドを表します。スライダーに沿ってインディケーター・マークを移動して、予測変数の重要度の順序で並べられた次の 10 個または前の 10 個のフィールドを表示できます。

グラフをダブルクリックして、グラフ設定を編集するための別個のダイアログ・ボックスを開くことができます。例えば、グラフのサイズ、使用されるフォントのサイズと色などの項目を修正できます。この別個の編集ダイアログ・ボックスを閉じると、「出力」タブに表示されるグラフに変更が適用されます。

「残差と予測」プロット

このプロットは、外れ値を特定するため、あるいは非線型誤差分散や非定数誤差分散を診断する場合に使用できます。理想的なプロットでは、ゼロの線の周辺に点が無作為に散在します。

予想されるパターンは、線型予測変数の予測値全体にわたる標準化最大対数尤度比残差の分布の平均値がゼロで、かつ範囲が一定であるというものです。予想されるパターンは、ゼロを通る水平な線となります。

GLE モデル ナゲットの設定

GLE モデル ナゲットの「設定」タブでは、モデル スコアリング時の未調整傾向のオプションおよび SQL 生成のオプションを指定します。このタブは、モデル・ナゲットがストリームに追加された後にのみ使用できます。

未調整傾向スコアの計算: フラグ型対象のみを含むモデルについては、対象フィールドに true の結果が指定されている尤度を示す未調整傾向スコアを要求できます。これらは、標準の予測値と確信度値に追加されています。調整済み傾向スコアは使用できません。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL の生成方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

Cox ノード

Cox 回帰は時間事象データのための予測モデルを構築します。モデルは、対象となるイベントが予測値変数の特定の値のために特定の時間 *t* に発生した確率を予測する生存関数を生成します。生存関数の形状および予測値の回帰係数は観測サブジェクトから推測されます。その後モデルは予測値変数の計測がある新しいケースに適用できます。検閲されたサブジェクトからの情報、つまり観測時に対象となるイベントを経験しないサブジェクトからの情報は、モデルの推定に非常に役立ちます。

例: 顧客離れを減らすという目標の一環として、ある通信会社は「解約するまでの期間」のモデル作成に注目し、解約して他社のサービスに切り替える顧客と関連する要因を特定します。最終的に、顧客の無作為サンプルが選択され、顧客としての期間 (アクティブな顧客かどうかに関係なく) やさまざまな人口統計フィールドがデータベースから取り出されます。

要件: 1 つまたは複数の入力フィールド、1 つの対象フィールドが必要で、Cox ノード内に生存時間フィールドを指定する必要があります。「偽 (false)」の値が生存期間を表し、「真 (true)」の値が対象となるイベントが発生したことを表すよう、対象フィールドをコーディングします。対象フィールドは文字列または整数のストレージを持つフラグ型の測定の尺度である必要があります。(ストレージは、必要に応じて、フィルター・ノードまたはフィールド作成ノードを使用して変換することができます。) 両方 またはなし

が設定されているフィールドは無視されます。モデルで使用するフィールド・タイプは、完全にインスタンス化する必要があります。生存時間は数値型フィールドにすることができます。

注: Cox 回帰モデルのスコアリング時、カテゴリー変数の空文字列をモデル構築の入力として使用すると、エラーが報告されます。空文字列を入力として使用しないでください。

日付/時間: 日付/時間フィールドがある場合、そのフィールドを使用して研究日および観察日へのエントリーの日付間の差異に基づいて、生存時間を含むフィールドを作成する必要があります。

Kaplan-Meier 分析。 Cox 回帰は、入力フィールドなしで実行することができます。これは、Kaplan-Meier 分析と同じです。

Cox ノードのフィールド・オプション

生存時間: 数値型フィールド (測定の尺度が連続型の 1 つ) を選択して、ノードを実行可能にします。生存時間は、予測されるレコードの寿命を示します。例えば、顧客が解約するまでの時間のモデル作成を行う場合、顧客がどれくらいの期間組織に属するかを記録するフィールドとなります。顧客が参加または解約した日付は、モデルに影響を与えません。顧客の保有期間のみが関連します。

生存時間は、単位のない期間であるとみなされます。入力フィールドが生存時間に一致することを確認する必要があります。例えば、月ごとに解約を測定する場合、年間売り上げでなく月間売り上げを入力として使用します。データに期間ではなく開始日と終了日がある場合、これらの日付を Cox ノードの上流で期間に最コード化する必要があります。

このダイアログ・ボックスのその他のフィールドは、IBM SPSS Modeler では使用する標準的なものです。 詳しくは、トピック 31 ページの『モデル作成ノードのフィールド・オプション』を参照してください。

Cox ノードの「モデル」オプション

モデル名: ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

分割モデルを作成 : 分割フィールドとして指定される入力フィールドの各値の個別モデルを作成します。詳しくは、 28 ページの『分割モデルの作成』を参照してください。

方法: 予想値をモデルに入力するのに、次のオプションが利用できます。

- **Enter :** デフォルトの手法で、すべての項がモデルに直接入力されます。モデル作成時にフィールド選択は実行されません。
- ステップワイズ法: フィールド選択に対するステップワイズ法は、名前が示すとおりステップごとにモデルを作成していきます。初期モデルは最も単純なモデルで、モデルにモデルの項はありません (定数を除く)。各ステップで、モデルにまだ追加されていない項を評価します。評価された項の中で最適な項がモデルの予測精度を大幅に改善する場合、その項が追加されます。 さらに、モデルの現在の項が再評価され、削除してもモデルの性能が低下しないかどうかが判断されます。低下しないと判断されると、これらの項は削除されます。この処理が繰り返されて、他の項の追加や削除が行われます。項を追加してもモデルの性能が改善されず、項を削除してもモデルの性能が低下しなくなった時点で、最終モデルが生成されます。

- 変数減少ステップワイズ法**:** 変数減少ステップワイズ法は、基本的にステップワイズ法の反対です。この方法では、すべての項が予測フィールドとして初期モデルに含まれています。各ステップにおいて、モデル中の項が評価され、削除してもモデルの性能が大幅に低下しない項が削除されます。また、前に削除された項が再評価され、それらの項目を追加するとモデルの予測精度が大幅に改善されるかどうかが判断されます。大幅に改善される場合は、その項がモデルに追加されます。項を削除してもモデルの性能が大幅に低下せず、項を追加してもモデルの性能が改善されなくなった時点で、最終モデルが生成されます。

注 : 自動手法 (ステップワイズ法および変数減少法を含む) は、非常に適応性の高い学習手法なので、学習データがオーバーフィットする傾向が強くなります。これらの方法を使用するときは、新しいデータまたはデータ区分ノードを使用して作成され提供されたサンプルを使用して、作成されたモデルの妥当性を検証することが非常に大切になります。

グループ: グループ・フィールドを指定すると、ノードはフィールドの各カテゴリーの個別モデルを計算します。グループ・フィールドは、文字列または整数のストレージを持つカテゴリー・フィールド (フラグ型または名義型) である場合があります。

モデル タイプ **:** モデルの項を定義する 2 つのオプションがあります。「主効果」 を選択すると、モデルに入力フィールドが個別に含まれ、入力フィールド間の交互作用は検定されません (倍数効果)。「ユーザー設定」を選択すると、モデルには指定した項 (主効果と交互作用) だけが含まれます。このオプションを選択した場合、「モデルの項」リストを使用してモデルに項を追加、または削除します。

モデルの項 **:**「ユーザー設定」でモデルを構築する場合、モデル中の項を明示的に指定する必要があります。このリストには、モデルの現在の項のセットが表示されます。「モデルの項」リストの右側にあるボタンを使用して、モデルの項を追加、削除することができます。

- モデルに項を追加するには、「モデルの項の新規追加」ボタンをクリックします。
- 項を削除するには、該当する項を選択して「選択したモデルの項の削除」ボタンをクリックします。

Cox 回帰モデルへの項の追加

ユーザー設定のモデルを要求する場合、「モデル」タブで 「モデルの項の新規追加」 ボタンをクリックすることにより、モデルに項を追加することができます。項を指定するための新しいダイアログ・ボックスが表示されます。

追加する項のデータ型 **:**「利用可能フィールド」リストで選択した入力フィールドに応じて、さまざまな方法でモデルに項を追加することができます。

- 単一の交互作用 **:** すべての選択したフィールドの交互作用を表す項を挿入します。
- 主効果: 選択した各入力フィールドに対して、1 つの主効果の項 (フィールド自体) を挿入します。
- すべての **2** 要因の交互作用 **:** 選択した入力フィールドの考えられる各組み合わせに対して、2 要因の交互作用の項 (入力フィールドの生成物) を挿入します。例えば、「利用可能フィールド」リストから入力フィールド *A*、*B*、および *C* を選択した場合、この方法では項 *A * B*、*A * C*、および *B * C* が挿入されます。
- すべての **3** 要因の交互作用 **:** 選択した入力フィールドの考えられる各組み合わせに対して、3 要因の交互作用の項 (入力フィールドの生成物) を挿入します (一度に 3 つを取得)。例えば、「利用可能フィールド」リストから入力フィールド *A*、*B*、*C*、および *D* を選択した場合、この方法では項 *A * B * C*、*A * B * D*、*A * C * D*、および *B * C * D* が挿入されます。
- すべての **4** 要因の交互作用 **:** 選択した入力フィールドの考えられる各組み合わせに対して、4 要因の交互作用の項 (入力フィールドの生成物) を挿入します (一度に 4 つを取得)。例えば、「利用可能フィ

ールド」リストから入力フィールド *A*、*B*、*C*、*D*、および *E* を選択した場合、この方法では項 *A* * *B* * *C* * *D*、*A* * *B* * *C* * *E*、*A* * *B* * *D* * *E*、*A* * *C* * *D* * *E*、および *B* * *C* * *D* * *E* が挿入されます。

利用可能なフィールド**:** モデルの項を構築するために利用できる入力フィールドが表示されます。リストには正しい入力フィールドでないフィールドが含まれている場合があるため、すべてのモデルの項に入力フィールドのみが含まれていることを確実にする必要があります。

プレビュー: 「挿入」 をクリックした場合に、上で選択したフィールドと項のデータ型に基づいて、モデルに追加される項が表示されます。

挿入 **:** (現在のフィールドおよび項のデータ型の選択内容に基づいて) モデルに項を挿入し、ダイアログ・ボックスを閉じます。

Cox ノードの「エキスパート」オプション

収束**:** これらのオプションを使用して、モデル収束のパラメーターを制御することができます。 モデルを実行するときに、収束設定によって、どれだけうまく適合するかを調べるために、異なるパラメーターを繰り返し実行する回数が制御されます。パラメーターを使用する回数が多くなればなるほど、結果が近くなります (つまり結果が収束します)。 詳しくは、トピック『Cox ノードの収束基準』を参照してください。

出力**:** これらのオプションによって、ノードに構築された生成モデルの詳細出力に表示される生存曲線などの付加統計量およびプロットを要求することができます。詳しくは、トピック『Cox ノードの詳細出力オプション』を参照してください。

ステップ基準**:** これらのオプションにより、ステップワイズ推定方法を使用したフィールドの追加と除去のための基準を制御することができます (「強制投入法」 を選択した場合、このボタンは無効になります)。詳しくは、トピック 244 ページの『Cox ノードのステップ基準』を参照してください。

Cox ノードの収束基準

最大反復回数: モデルの最大反復数を指定し、解決方法を探す手順の期間を制御することができます。

対数尤度収束**:** 対数尤度の相対変化がこの値未満になると、反復が停止します。値が 0 の場合、この基準は使用されません。

パラメーター収束**:** パラメーター推定値の絶対変化または相対変化がこの値未満になると、反復が停止します。値が 0 の場合、この基準は使用されません。

Cox ノードの詳細出力オプション

統計: exp(*B*) の信頼区間および推定値の相関を含む、モデル・パラメーターの統計量を取得することができます。「各ステップで」、または「最終ステップ」で、これらの統計量を要求できます。

ベースライン関数の表示。 ベースライン ハザード関数および共変量の平均の累積生存を表示することができます。

作図

作図することにより、推定モデルを評価し、また結果を解釈できます。累積生存関数、ハザード関数、ログマイナスログ関数、および 1 マイナス累積生存関数を作図できます。

- 生存*:* 累積生存関数を線型スケールで表示します。
- ハザード*:* 累積ハザード関数を線型スケールで表示します。

- ログマイナスログ。 ln(-ln) 変換が推定に適用された後、累積生存推定を表示します。
- 累積死亡関数 : 線型スケールで 1 マイナス累積生存関数を作図します。

各値の個別の線を作図します。 このオプションは数値型フィールドに対してのみ使用できます。

作図に使用する値: これらの関数は予測の値に依存するため、予測に定数値を使用して時間に対して関数を作図する必要があります。デフォルトでは、定数値として各予測の平均を使用しますが、グリッドを使用した作図に独自の値を入力することができます。カテゴリー入力の場合、指標のコード化を使用するため、各カテゴリー (最後のカテゴリーを除いて) に回帰係数が存在します。カテゴリー入力には、指標の対比に対応したカテゴリーのケースの比率に等しい、各指標の対比の平均値があります。

Cox ノードのステップ基準

削除基準 : 強力なモデルには 「尤度比」 を選択します。モデル構築に必要な時間を短縮するには、「ワルド」 を選択してみることもできます。条件パラメーター推定値に基づく尤度比統計の確率に基づく削除テストを行う、「条件」 の追加オプションがあります。

基準の有意しきい値: このオプションを選択すると、各フィールドに関連付けられた統計確率 (*p* 値) に基づいて選択基準を指定することができます。フィールドは、該当する *p* 値が 「投入」 値より小さい場合にのみモデルに追加され、*p* 値が 「削除」 値より大きい場合にのみ削除されます。「投入」 には 「削除」 よりも小さい値を指定してください。

Cox ノードの設定オプション

将来の生存を予測: 将来の時間を指定します。生存、つまり少なくとも最終的なイベントが発生していない (現在からの) 時間の長さで各ケースが生存しているかどうかが、各時間値の各レコードに予測され、時間値ごとに 1 つの予測が行われます。生存は、対象フィールドの「偽 (false)」の値です。

- 一定の間隔: 生存期間値は、指定された「時間区分」と「スコアリングの時間数」から生成されます。例えば、各時間に 2 の間隔で 3 つの期間が要求された場合、生存は 2、4、6 の将来の時間に予測されます。各レコードは同じ時間値で評価されます。
- 時間フィールド: 生存時間が選択された時間フィールドの各レコードに提供されます (フィールドごとに 1 つの予測が生成されます)。そのため、各レコードはさまざまな時間に評価されます。

過去の生存期間: レコードの生存時間を指定します。例えば、既存の顧客の保有期間をフィールドとして指定します。将来の生存の尤度をスコアリングは、過去の生存時間の条件式です。

注 : 将来および過去の生存時間の値は、モデルを学習するために使用されるデータの生存時間の範囲内にある必要があります。時間が範囲外となるレコードは、ヌルとしてスコアリングされます。

すべての確率を追加: 出力フィールドの各カテゴリーの確率を、ノードで処理される各レコードに追加するかどうかを指定します。このオプションを選択しないと、予測されたカテゴリーの確率だけが追加されます。確率は、それぞれの将来の時間に対して計算されます。

累積ハザード関数を計算: 累積ハザードの値が各レコードに追加されるかどうかを指定します。累積ハザードは、それぞれの将来の時間に対して計算されます。

Cox モデル・ナゲット

Cox 回帰モデルは、Cox ノードによって推定された式を表します。これらには、モデルが取得したすべての情報と、モデル構造とパフォーマンスに関する情報が含まれます。

生成された Cox 回帰モデルを含むストリームを実行すると、そのモデルの予測と関連付けられた確率を含む 2 つの新規フィールドが追加されます。新規フィールド名は、予測される出力フィールド名から派生し、予測カテゴリーの *$C-* および関連する確率の *$CP-* の接頭辞、将来の時間間隔の数または時間間隔を定義する時間フィールドの名前の接尾辞が付きます。例えば、解約 *(churn)* という出力フィールドで一定の間隔で定義された 2 つの将来の時間区分がある場合、新規フィールド名は、*$C-churn-1*、*$CP-churn-1*、*$C-churn-2*、および *$CP-churn-2* となります。将来の時間が時間フィールド 保有期間 *(tenure)* で定義されている場合、新規フィールド名は *$C-churn_tenure* および *$CP-churn_tenure* となります。

Cox ノードで 「すべての確率を追加」 設定オプションを選択した場合、各レコードの生存確率および失敗の確率が含まれる 2 つの追加フィールドがそれぞれの将来の時間に追加されます。これらの追加フィールドは出力フィールド名に基づいて、生存の確率の場合は *$CP-<false value>-*、イベントが発生する確率の場合は *$CP-<true value>-* の接頭辞が付き、将来の時間間隔の数の接尾辞が付いた名前になります。例えば、「偽 (false)」の値が 0 で「真 (true)」の値が 1 である出力フィールドで、一定の間隔で定義された 2 つの将来の時間区分である場合、新規フィールド名は *$CP-0-1*、*$CP-1-1*、*$CP-0-2*、および *$CP-1-2* となります。将来の時間が 1 つの時間フィールドの保有期間 *(tenure)* で定義されている場合、将来の時間間隔が 1 つであるため、新規フィールド名は *$CP-0-1* および *$CP-1-1* となります。

Cox ノードの 「累積ハザード関数を計算」 設定オプションを選択した場合、各レコードの累積ハザード関数を含む追加フィールドがそれぞれの将来の時間に追加されます。これらの追加フィールドは、出力フィールド名に基づいて、*$CH-* の接頭辞が付き、将来の時間間隔の数または時間間隔を定義する時間フィールド名の接頭辞が付いた名前になります。例えば、解約 *(churn)* という出力フィールドで一定の間隔で定義された 2 つの将来の時間区分がある場合、新規フィールド名は、*$CH-churn-1*、*$CH-churn-2* となります。将来の時間が時間フィールドの保有期間 *(tenure)* で定義されている場合、新規フィールド名は *$CH-churn-1* となります。

Cox 回帰の出力設定

SQL 生成用のコントロールを除き、ナゲットの「設定」タブには、モデル・ノードの「設定」タブと同じコントロールが含まれています。ナゲット・コントロールのデフォルト値は、モデル・ノードに設定された値によって決まります。 詳しくは、トピック 244 ページの『Cox ノードの設定オプション』を参照してください。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

Cox 回帰の詳細出力

Cox 回帰の詳細出力からは、生存曲線など、推定されるモデルとそのパフォーマンスに関する詳細情報が得られます。詳細出力に含まれる情報は、技術的な情報がほとんどです。この出力を適切に解釈するには、Cox 回帰分析に関する広範な知識が必要です。

第 11 章 クラスタリング・モデル

クラスタリング・モデルは、類似したレコードのグループを識別し、そのグループに従ってレコードにラベルを付けます。この操作には、グループやその特性に関する事前の知識は必要ありません。実際には、検索するグループ数が正確にわからない場合もあります。これが、クラスタリング・モデルと他のマシン学習技法との違いであり、クラスタリング・モデルには、モデルが予測する定義済みの出力フィールドや対象フィールドはありません。クラスタリング・モデルは、モデルの分類性能を判定する外部標準がないので、教師なし学習モデルと呼ばれることがよくあります。これらのモデルには、正、誤 という回答はありません。モデルの価値は、データのグループ構成を把握し、それらのグループについて役に立つ説明を提供できるかどうかで決まります。

クラスタリング手法は、レコード間およびクラスター間の距離の測定に基づいています。レコードは、同じクラスターに属するレコード間の距離を最小にするようにして、クラスターに割り当てられます。

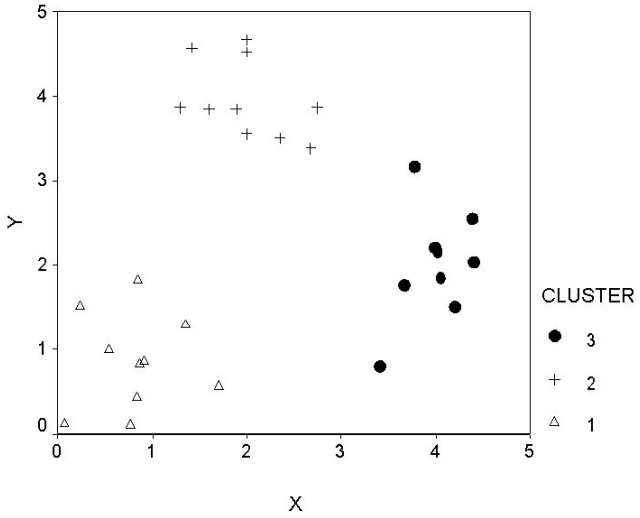


図 *44.* 簡単なクラスタリング・モデル

以下のクラスター化方法が用意されています。

K-Means ノードで、データ・セットが異なるグループ (つまりクラスター) へ、クラスタリングされます。この方法で、固定数のクラスターを定義し、クラスターにレコードを繰り返し割り当てて、これ以上調整してもモデルが改善されなくなるまで、クラスターの中心を調整します。*K*-means では、結果を予測するのではなく、入力フィールドのセット内のパターンを明らかにするために、「教師なし学習」として知られるプロセスが使用されます。

TwoStep ノードで、2 段階のクラスター化手法が使用されます。最初のステップでは、データを 1 度通過させて、未処理の入力データを管理可能な一連のサブクラスターに圧縮します。2 番目のステップでは、階層クラスター化手法を使用して、サブクラスターをより大きなクラスターに結合させていきます。TwoStep には、学習データに最適なクラスター数を自動的に推定するという利点があります。また、フィールド・タイプの混在や大規模データ・セットも効率よく処理できます。

Kohonen ノードは、ニューラル・ネットワークの一種であり、データ・セットをクラスター化して異なるグループを形成する目的で使用できます。ネットワークの学習が完了すると、類似のレコードは出力マップで互い近くに表示され、違いの大きいレコードほど離れたところに表示されます。強度の高いユニットを識別するために生成されたモデル内で、各ユニットが獲得した観察の数値を調べることができます。これは、適切なクラスター数についてのヒントになる場合があります。

Hierarchical Density-Based Spatial Clustering (HDBSCAN)© は、教師なし学習を使用してデータ・セットのクラスター (つまり、密度の高い領域) を検出します。SPSS Modeler の HDBSCAN ノードは、HDBSCAN ライブラリーのコア機能およびよく使用されるパラメーターを公開します。このノードは Python で実装されており、最初にグループの性質が分からない場合にデータ・セットを異なるグループにクラスター化するために使用できます。

クラスタリング・モデルは、クラスターやセグメントを作成するためによく利用されます。このクラスターやセグメントは、後の分析で入力として使用されます。一般的な例として、マーケット・セグメントがあります。これは、マーケティング担当者がマーケット全体を等質のサブグループに細分化するために使用します。各セグメントには、そのセグメントを対象としたマーケティングの成果に影響する特性があります。データ・マイニングを使用してマーケティング戦略を最適化している場合は、適切なセグメントを識別し、そのセグメント情報を予測モデルで使用することで、モデルを大幅に改善できます。

Kohonen ノード

Kohonen ネットワークは、クラスタリングを実行するニューラル・ネットワークの一種で、**knet**、または自己組織化マップとしても知られています。この種のネットワークを使用すると、開始時にグループの性質がわからない場合に、データセットを異なるグループにクラスター化することができます。グループまたはクラスター内のレコードは互いに似た傾向があり、異なるグループのレコードとは似ていないように、レコードがグループ化されます。

基本ユニットはニューロンで、次の 2 つの層で編成されています。入力層および出力層 (出力マップと呼ばれることもあります)。すべての入力ニューロンがすべての出力ニューロンに接続されます。これらの接続には、それぞれに関連付けられた強さまたは重みがあります。学習中、各ユニットは各レコードを「勝ち取る」ために互いに競争します。

出力マップは、ニューロンの 2 次配列グリッドで、ユニット間の接続はありません。

入力データが入力層に入り、値が出力層に伝達されます。最も強い応答の出力ニューロンはウィナーと呼ばれ、入力に対する応答となります。

最初は、重みはすべて無作為です。あるユニットがレコードを勝ち取ると、そのレコードの予測値のパターンとの適合性を高めるために、重みが (隣接と呼ばれる近くのユニットの重みとともに) 調整されます。入力レコードがすべて表示され、それに従って重みが更新されます。変化がほとんどなくなるまで、この処理が何回も繰り返されます。学習が進行するにつれて、グリッド ユニットの重みがクラスターの 2 次元「マップ」を構成するように調整されます (自己組織化マップと呼ばれる理由です)。

ネットワークの学習が完了すると、類似のレコードは出力マップで互い近くに表示され、違いが非常に大きいレコードほど離れたところに表示されます。

他の IBM SPSS Modeler の学習方法とは異なり、Kohonen ネットワークは対象フィールドを使用しません。このタイプの学習は、対象フィールドがないことから、教師なし学習と呼ばれます。Kohonen ネット

ワークは、結果を予測するのではなく、一連の入力フィールドのパターンを明らかにします。通常、最終的な Kohonen ネットワークは、多数の観測値を要約した少数のユニット (強いユニット) と、どの観測値とも対応しない複数のユニット (弱いユニット) で構成されます。強いユニット (グリッド内の近接ユニットを含むこともある) はクラスターの中心候補を表します。

Kohonen ネットワークは、次元分解にも用いられます。2 次元グリッドの空間的な特徴により、*k* 個の元の予測値から、元の予測フィールドの類似関係を保持する 2 つの派生特性への関連付けが行われます。場合によっては、因子分析や主成分分析と同様の利点があります。

出力グリッドのデフォルト・サイズの算出方法が、前のバージョンの IBM SPSS Modeler から変更されたことに注意してください。新しい手法では、全般的に学習が速く、効果的に一般化できる、小さい出力層が生成されます。デフォルト・サイズで満足する結果を得られない場合は、「エキスパート」タブで出力グリッドのサイズを増やしてください。 詳しくは、トピック 250 ページの『Kohonen ノードの「エキスパート」オプション』を参照してください。

要件: Kohonen ネットワークの学習には、役割が入力 に設定された 1 つ以上のフィールドが必要です。対象、両方、またはなしが役割に設定されたフィールドは無視されます。

利点: Kohonen ネットワーク・モデルを作成する場合、グループに属するデータは必要ありません。検索するグループ数も必要ありません。Kohonen ネットワークは多数のユニットから始めて、学習が進行するにつれて、ユニットがデータの自然クラスターを形成していきます。モデル・ナゲットの各ユニットが獲得した観測値数から強いユニットを識別することにより、適切なクラスター数がわかります。

Kohonen ノードの「モデル」オプション

モデル名: ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

既存モデルの学習を継続: Kohonen ノードを実行するたびに、完全に新しいネットワークが作成されます。このオプションを選択すると、ノードによって正常に生成された最後のネットワークで学習が続行されます。

フィードバック グラフを表示: このオプションを選択すると、学習中に 2 次元配列が表示されます。各ノードの強さが色で示されます。各ノードの強さが色で示されます。赤は、多くのレコードを勝ち取ったユニットを示します (強いユニット)。白は、ほとんどまたはまったくレコードを獲得しなかったユニットを示します (弱いユニット)。モデル作成にかかる時間が比較的短い場合、フィードバックが表示されない場合があります。この機能を選択すると、学習時間が長くなることがあります。学習時間を短縮するには、このオプションの選択を解除します。

停止条件 :「デフォルト」停止基準では、内部パラメーターに基づいて学習が停止されます。時間を停止基準に指定することもできます。「時間」にネットワークの学習時間を分ユニットで入力します。

ランダム シードの設定: ランダム・シードが設定されないと、ネットワークの重みを初期化する際に使用される乱数値のシーケンスが、ノードが実行されるたびに異なります。これは、ノード設定とデータ値がまったく同じでも、ノードが実行されるたびに異なるモデルが作成されるからです。このオプションを選択し、ランダム・シードを特定の値に設定すると、作成されたモデルを正確に再現することが可能になります。特定のランダム・シードからは常に同じシーケンスの乱数値が生成されるため、ノードを実行すると常に同じ生成モデルが作成されます。

注**:** データベースから読み込まれたレコードで「ランダム・シードの設定」オプションを使用する場合は、ノードを実行するたびに同じ結果になるように、サンプリングの前にソート・ノードが必要になることがあります。この理由は、ランダム シードがレコードの順序に依存しているためです。各レコードがリレーショナル・データベース内で同じ位置に留まる保証はありません。

注 : モデルに名義型 (セット型) フィールドを取り入れたいけれども、モデルの構築時にメモリー上の問題がある場合、またはモデルの構築に時間がかかりすぎるような場合は、値を減らすために大きなセット型フィールドに記録するか、またはラージ・セットの代わりに少ない値を持つ別のフィールドを使用することを検討してください。例えば、個別の製品の値が設定された *product_id* フィールドに問題がある場合は、モデルからこのフィールドを削除し、代わりに大まかな *product_category* フィールドを追加します。

最適化**:** 特定のニーズに応じて、モデルの構築時にパフォーマンスを向上させるために設計されたオプションを選択します。

- パフォーマンス向上のために処理過程のデータをディスクへ書き出さないようにアルゴリズムに指示する場合は、「速度」 を選択します。
- ある程度は速度が遅くなっても処理過程のデータをディスクへ書き出すようにアルゴリズムに指示するには、「メモリー」 を選択します。デフォルトでは、このオプションが選択されます。

  注**:** 分散モードで実行する場合、この設定は、`options.cfg` 内に指定された管理者オプションによってオーバーライドされることがあります。

クラスター・ラベルの結合 **:** 新しいモデルについてデフォルトで選択されますが、以前のバージョンの IBM SPSS Modeler から読み込まれたモデルについては選択を解除し、K-Means ノードとTwoStep ノードで作成された同じ種類のカテゴリー・スコア フィールドを 1 つ作成します。この文字列フィールドは、さまざまなモデル タイプのランク付け指標を計算する場合に、自動クラスタリング・ノードで使用されます。 詳しくは、トピック 79 ページの『自動クラスタリング・ノード』を参照してください。

Kohonen ノードの「エキスパート」オプション

Kohonen ネットワークをよく理解している場合は、エキスパート・オプションを使用して、学習過程を調整できます。エキスパート・オプションを利用するには、「エキスパート」タブで「モード」に 「エキスパート」 を設定してください。

幅および長さ **:** 2 次元出力マップのサイズ (幅と長さ) を、各次元の出力ユニット数で指定します。

学習率の減衰 **:** 学習率の減衰の「線型」または「指数」を選択します。学習率は、時間の経過とともに低下する重み付け因子です。ネットワークは、データの大きな特性のコード化から始め、次第により詳細なレベルへと焦点を当てていきます。

フェーズ **1** およびフェーズ **2 :** Kohonen ネットワークは 2 つのフェーズに分割されます。「フェーズ 1」は大まかな推定フェーズです。データの全体的なパターンの把握に使用されます。「フェーズ 2」は調整フェーズです。データの詳細な特性をモデル化するためのマップの調整に使用されます。各フェーズには 3 つのパラメーターがあります。

- 隣接 **:** 隣接の開始サイズ (半径) を設定します。これにより、学習中に勝ち取ったユニットとともに更新される「隣接」ユニットの数が決まります。フェーズ 1 の間、隣接サイズは フェーズ *1* 隣接 から始まり、(フェーズ 2隣接 + 1) まで低下します。フェーズ 2 の間、隣接サイズはフェーズ *2* 隣接から始まり、1.0 まで低下します。フェーズ *1* 隣接は、フェーズ *2* 隣接より大きくなければなりません。
- 初期 η **:** 学習率 η の開始値を設定します。フェーズ 1 の間、ηはフェーズ *1* 初期η から始まり、フェーズ *2* 初期η まで低下します。フェーズ 2 では、ηはフェーズ *2* 初期η から始まり、0まで低下します。フェーズ *1* 初期ηは、フェーズ *2* 初期η より大きくなければなりません。

- サイクル : 各学習フェーズのサイクル数を設定します。データに対して、各フェーズの実行を指定回数だけ繰り返します。

Kohonen モデル・ナゲット

Kohonen モデル・ナゲットには、学習済みの Kohenen ネットワークが取得したすべての情報と、Kohonen ネットワークのアーキテクチャーに関する情報が含まれます。

Kohonen モデル・ナゲットを含むストリームを実行すると、そのレコードに対して最も強く応答したユニットの Kohenen 出力グリッド内の *X* 座標と *Y* 座標を含む 2 つの新規フィールドが追加されます。新規フィールド名はモデル名から派生し、接頭辞の *$KX-* と *$KY-* が付けられます。例えば、モデル名が *Kohonen* の場合、新規フィールド名は *$KX-Kohonen* と *$KY-Kohonen* になります。

Kohonen ネットワークがコード化した内容をより詳しく理解するには、モデル・ナゲット・ブラウザーの「モデル」タブをクリックします。クラスター・ビューアーに、クラスター、フィールド、および重要度レベルがグラフィカルに表示されます。 詳しくは、トピック 264 ページの『クラスター・ビューアー -「モデル」タブ』を参照してください。

クラスターをグリッドとして視覚化する場合は、散布図ノードを使用して *$KX-* および *$KY-* フィールドをプロットすることにより、Kohonen ネットの結果を表示できます (各ユニットのレコードがすべて互いに重なり合ってプロットされることを防ぐために、散布図ノードの「**X** の拡散」および「**Y** の拡散」を選択してください)。散布図では、シンボル値フィールドをオーバーレイして、Kohonen ネットワークによってどのようにデータがクラスター化されたかを調べることもできます。

Kohonen ネットワークを詳しく調べるための効果的な手法として、他に、ルール算出を使用して、ネットワークによって検出されたクラスターを区別する特性を発見する方法があります 詳しくは、トピック 110 ページの『C5.0 ノード』を参照してください。

モデル・ブラウザー使用法に関する一般情報については、 43 ページの『モデル・ナゲットの参照』を参照してください。

Kohonen モデルの要約

Kohonen モデル・ナゲットの「要約」タブには、ネットワークのアーキテクチャーまたはトポロジーに関する情報が表示されています。2 次元 Kohonen 機能マップ (出力層) の長さと幅は、**$KX-** *model_name* および **$KY-** *model_name* として表示されます。入力層と出力層に対しては、各層のユニット数が表示されます。

K-Means ノード

K-Means ノードは、クラスター分析手法を提供しています。開始時にグループの性質がわからない場合に、このノードを使用してデータセットを異なるグループにクラスター化できます。他の IBM SPSS Modeler の学習方法とは異なり、K-Means モデルは対象フィールドを使用しません。このタイプの学習は、対象フィールドがないことから、教師なし学習と呼ばれます。K-Means では、結果が予測されるのではなく、一連の入力フィールドのパターンが明らかにされます。レコードは、1 つのグループまたはクラスター内のレコード同士がよく似た特性を持ち、異なるグループのレコードが互いに類似しないように分類されます。

K-Means では、データから派生した開始クラスター中心のセットが定義されます。その後、レコードの入力フィールド値を基に、各レコードが最も類似するクラスターに割り当てられます。ケースの割り当てが完了すると、クラスター中心が更新され、各クラスターに割り当てられた新しいレコードのセットが反映され

ます。その後、レコードを別のクラスターに再割り当てする必要があるかどうかが再確認されます。このレコード割り当てとクラスター反復の過程は、最大反復数に達するまで、またはある反復処理と次の反復処理間の変化が、指定された閾値を超えなくなるまで繰り返されます。

注 : 作成されたモデルは、学習データの順序にある程度依存します。データを並べ替えてモデルを再作成すると、異なる最終クラスター・モデルが作成されることがあります。

要件: K-Means モデルの学習には、役割が「入力」に設定された 1 つ以上のフィールドが必要です。出力、両方、またはなしが役割に設定されたフィールドは無視されます。

強度: 所属グループにデータがなくても K-Means モデルを作成することができます。たいていの場合 K-Means モデルを使用すると、大量のデータセットを最も高速にクラスター化できます。

K-Means ノードの「モデル」オプション

モデル名: ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

指定した数のクラスター: 生成するクラスタの数を指定します。デフォルトは 5 です。

距離フィールドの生成 : このオプションを選択すると、モデル・ナゲットに、各レコードの割り当て先クラスターの中心からの距離を表すフィールドが含まれます。

クラスター・ラベル : 生成された所属クラスター・フィールドの値の形式を指定します。所属クラスターは、文字列に指定したラベル接頭辞を付けて (例 : `"Cluster 1"`、`"Cluster 2"`など)、または数値で示すことができます。

注 : モデルに名義型 (セット型) フィールドを取り入れたいけれども、モデルの構築時にメモリー上の問題がある場合、またはモデルの構築に時間がかかりすぎるような場合は、値を減らすために大きなセット型フィールドに記録するか、またはラージ・セットの代わりに少ない値を持つ別のフィールドを使用することを検討してください。例えば、個別の製品の値が設定された *product_id* フィールドに問題がある場合は、モデルからこのフィールドを削除し、代わりに大まかな *product_category* フィールドを追加します。

最適化: 特定のニーズに応じて、モデルの構築時にパフォーマンスを向上させるために設計されたオプションを選択します。

- パフォーマンス向上のために処理過程のデータをディスクへ書き出さないようにアルゴリズムに指示する場合は、「速度」 を選択します。
- ある程度は速度が遅くなっても処理過程のデータをディスクへ書き出すようにアルゴリズムに指示するには、「メモリー」 を選択します。デフォルトでは、このオプションが選択されます。

  注: 分散モードで実行する場合、この設定は、`options.cfg` 内に指定された管理者オプションによってオーバーライドされることがあります。

K-means ノードの「エキスパート」オプション

K-means クラスター化をよく理解している場合は、エキスパート・オプションを使用して、学習過程を調整できます。エキスパート・オプションを利用するには、「エキスパート」タブで「モード」に 「エキスパート」 を設定してください。

停止条件 : モデルの学習に使用する停止基準を指定します。「デフォルト」の停止基準は、反復回数が 20 回に達した時点か、変化が 0.000001 未満になった時点です。どちらか早い時点で停止されます。独自の停止基準を指定するには、「ユーザー設定」を選択します。

- 「最大反復回数」。指定された回数だけ反復した後、モデルの学習を中止します。
- 収束基準 : このオプションを選択すると、1 回の反復処理におけるクラスター中心の最大変化が、指定されたレベル未満になった時点で、モデルの学習が停止されます。

ダミー変数の調整値 : セット型フィールドを数値型フィールドのグループとして記録するために使用する値を、0〜1.0 の間で指定します。デフォルト値は、0.5 の平方根 (約 0.707107) です。記録したフラグ型フィールドの適切な重みとなります。この値が 1.0 に近づくほど、セット型フィールドには数値フィールドより重い重みが付けられます。

K-Means モデル・ナゲット

K-Means モデル・ナゲットには、クラスター化モデルが取得したすべての情報と、学習データと推定プロセスに関する情報が含まれます。

K-Means モデル作成ノードを含むストリームを実行すると、そのレコードの所属クラスターと割り当てられたクラスターの中心からの距離を含む 2 つの新規フィールドが追加されます。新規フィールド名はモデル名から派生し、所属クラスターのフィールドには接頭辞の *$KM-*、クラスターの中心からの距離のフィールドには接頭辞の *$KMD-* が付けられます。例えば、モデルの名前が *Kmeans* の場合、新規フィールド名は *$KM-Kmeans* と *$KMD-Kmeans* になります。

K-Means モデルを詳しく調べるための効果的な手法として、ルール算出を使用して、モデルによって検出されたクラスターを区別する特性を発見する方法があります 詳しくは、トピック 110 ページの『C5.0 ノード』を参照してください。モデル・ナゲット・ブラウザーの「モデル」タブをクリックしてクラスター・ビューアーを表示し、クラスター、フィールド、および重要度レベルをグラフィカルに参照することもできます。詳しくは、トピック 264 ページの『クラスター・ビューアー -「モデル」タブ』を参照してください。

モデル・ブラウザー使用法に関する一般情報については、 43 ページの『モデル・ナゲットの参照』を参照してください。

K-Means モデルの要約

K-Means モデル・ナゲットの「要約」タブには、学習データ、推定過程、およびモデルが定義したクラスターなどに関する情報が表示されています。クラスター数や反復履歴などが表示されています。このモデル作成ノードに接続されている精度分析ノードを実行した場合、その精度分析情報もこのセクションに表示されます。

TwoStep クラスター・ノード

TwoStep クラスター・ノードでは、一種のクラスター分析が行われます。開始時にグループの性質がわからない場合に、このノードを使用してデータセットを異なるグループにクラスター化できます。Kohonen ノードや K-Means ノードのように、TwoStep クラスター・モデルも対象フィールドを使用しません。TwoStep クラスターでは、結果が予測されるのではなく、入力フィールドのセットのパターンが明らかにされます。レコードは、1 つのグループまたはクラスター内のレコード同士がよく似た特性を持ち、異なるグループのレコードが互いに類似しないように分類されます。

TwoStep クラスターは、2 段階のクラスター化方法です。最初のステップでは、データを 1 度通過させて、元の入力データを管理可能な一連のサブクラスターに圧縮します。2 番目のステップでは、階層クラスター化方法を使用して、データを再度通過させることなく、サブクラスターをより大きなクラスターに結合させていきます。階層クラスター化には、事前にクラスター数を選択する必要がないという利点があります。多くの階層クラスター化方法では、各レコードを開始クラスターとして開始され、結合を繰り返して大きなクラスターが生成されます。通常、この方法では、大量のデータを扱うとデータセットが破壊されることがあります。TwoStep では、事前にクラスター化を行うので、大きなデータセットでも高速に階層クラスター化を実行できます。

注**:** 作成されたモデルは、学習データの順序にある程度依存します。データを並べ替えてモデルを再作成すると、異なる最終クラスター・モデルが作成されることがあります。

要件**:** TwoStep クラスター・モデルの学習には、役割を入力 に設定した 1 つ以上のフィールドが必要です。対象、両方、またはなしが役割に設定されたフィールドは無視されます。TwoStep クラスターのアルゴリズムは、欠損値を扱いません。空白の入力フィールドがあるレコードは無視してモデルが作成されます。

利点**:** TwoStep クラスター・ノードでは、異なるフィールド・タイプが混ざっていてもかまわないため、大きなデータ・セットを効率的に処理できます。また、複数のクラスター解を検定して最適な解を選択できるため、最初に必要なクラスター数を指定する必要がありません。TwoStep クラスターでは、外れ値 (結果に悪影響を及ぼす可能性がある極端なケース) が除外されるように設定することができます。

重要**:**

IBM SPSS Modeler には、以下に示す 2 つの異なるバージョンの TwoStep クラスター・ノードがあります。

- **TwoStep** クラスター: IBM SPSS Modeler Server で稼働する従来のノードです。
- **TwoStep-AS** クラスターは、IBM SPSS Analytic Server に接続しているときに実行できます。

TwoStep クラスター・ノードの「モデル」オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

数値フィールドの標準化**:** デフォルトでは、TwoStep のすべての数値入力フィールドが、平均 0、分散 1 の尺度に標準化されます。数値型フィールドの元の尺度を保持する場合は、このオプションの選択を解除します。シンボル値フィールドは、このオプションの影響を受けません。

外れ値を除外**:** このオプションを選択した場合、実体的なクラスターに適合しないレコードは自動的に分析対象から除外されます。外れ値を除外することで、結果の歪曲が防止できます。

外れ値の検出は、クラスタリング前の段階で行われます。このオプションが選択された場合、他のサブクラスターに比べてレコード数が少ないサブクラスターは、外れ値候補とみなされ、それらのレコードを除外してサブクラスターのツリーが再構築されます。外れ値候補を含むとみなされるサブクラスターのサイズは、「パーセンテージ」 オプションで制御されます。これらの外れ値候補レコードが新しいサブクラスターのプロファイルのいずれかと十分に類似している場合、そのレコードを再構築されたサブクラスターに追加することができます。残りの結合できない外れ値候補は外れ値とみなされて「ノイズ」クラスターに追加され、階層クラスタリングからは除外されます。

外れ値処理を使用する TwoStep モデルでデータをスコアリング する場合、もっとも近い実体クラスターから一定の閾値以上の距離がある新しいケースは外れ値とみなされ、「ノイズ」クラスターに -1 という名前が割り当てられます。

クラスター ラベル**:** 生成された所属クラスター・フィールドの形式を指定します。所属クラスターは、文字列に指定したラベル接頭辞を付けて (例 : "Cluster 1"、"Cluster 2" など)、または数値で示すことができます。

クラスタ数を自動的に計算。TwoStep クラスターでは、多数のクラスター解が非常に高速に分析され、学習データに最適なクラスター数を選択することができます。試行する解の範囲を指定するには、「最大」クラスター数と「最小」クラスター数を設定します。 TwoStep では、2 段階の過程を経て最適なクラスター数が判断されます。最初の段階では、クラスターの追加による BIC (ベイズ情報量基準) の変化に基づいて、モデル内のクラスター数の上限が選択されます。2 番目の段階では、BIC 解の最小値よりも少ないクラスターを使用して、すべてのモデルにおけるクラスター間の最小距離の変化が検出されます。最終的なクラスター・モデルは、距離の変化のうち最大のものを使用して識別されます。

クラスタ数を指定。モデルに含めるクラスター数がわかっている場合は、このオプションを選択してクラスター数を入力します。

距離測度**:** このオプションにより、2 つのクラスター間の類似度を計算する方法を指定します。

- 対数尤度: この尤度測定により、変数の確率分布を求めます。連続変数は正規分布しているものと仮定し、カテゴリー変数は多項分布しているものと仮定します。すべての変数は独立しているものと仮定します。
- ユークリッド: ユークリッド測定は、2 つのクラスター間の「直線」距離です。この測定方法は、すべての変数が連続している場合にだけ使用できます。

クラスター化の基準: 自動クラスタリング・アルゴリズムで、クラスターの個数を判定する方法を指定します。ベイズの情報量基準 (BIC) または赤池情報量基準 (AIC) のどちらかを指定できます。

TwoStep クラスター・モデル・ナゲット

TwoStep クラスター・モデル・ナゲットには、クラスター化モデルが取得したすべての情報と、学習データと推定プロセスに関する情報が含まれます。

TwoStep クラスター・モデル・ナゲットを含むストリームを実行すると、ノードによって、そのレコードの所属クラスターを含む新規フィールドが追加されます。新規フィールド名はモデル名から派生し、接頭辞の *$T-* が付けられます。例えば、モデルの名前が *TwoStep* の場合、新規フィールド名は *$T-TwoStep* になります。

TwoStep モデルを詳しく調べるための効果的な手法として、ルール算出を使用して、モデルによって検出されたクラスターを区別する特性を発見する方法があります。詳しくは、トピック 110 ページの『C5.0 ノード』を参照してください。モデル・ナゲット・ブラウザーの「モデル」タブをクリックしてクラスター・ビューアーを表示し、クラスター、フィールド、および重要度レベルをグラフィカルに参照することもできます。詳しくは、トピック 264 ページの『クラスター・ビューアー -「モデル」タブ』を参照してください。

モデル・ブラウザー使用法に関する一般情報については、 43 ページの『モデル・ナゲットの参照』を参照してください。

TwoStep モデルの要約

TwoStep クラスター・モデル・ナゲットの「要約」タブで、学習データ、推定過程、および使用された構築設定についての情報と一緒に、発見されたクラスターの数が表示されます。

詳しくは、トピック 43 ページの『モデル・ナゲットの参照』を参照してください。

TwoStep-AS クラスター・ノード

IBM SPSS Modeler には、以下に示す 2 つの異なるバージョンの TwoStep クラスター・ノードがあります。

- **TwoStep** クラスター: IBM SPSS Modeler Server で稼働する従来のノードです。
- **TwoStep-AS** クラスターは、IBM SPSS Analytic Server に接続しているときに実行できます。

Twostep-AS クラスター分析

TwoStep クラスターは、通常ははっきりしない、データセット内での自然なグループ化 (またはクラスター) を明確にすることを目的として設計された探索ツールです。この手続きで使用されるアルゴリズムには、従来のクラスタリング手法とは異なる以下の優れた特徴があります。

- カテゴリー変数と連続型変数の処理。変数が独立変数であると仮定すると、カテゴリー変数と連続型変数の場合、多項分布と正規分布を結合した配置になると考えられます。
- クラスター数の自動選択。複数のクラスター解の間でモデル選択基準の値を比較することにより、この手続きは、最適なクラスター数を自動的に判定することができます。
- スケーラビリティー。レコードを要約するためのクラスター特性 (CF) ツリーを作成すると、TwoStep アルゴリズムは、より大規模なデータ・ファイルを分析することができるようになります。

例えば、通常、小売業者や消費者製品企業は、顧客の購買傾向、性別、年齢、収入レベルなどの属性を記述した情報に対してクラスタリング手法を適用します。これらの企業は、販売を増加させ、ブランド ロイヤリティーを構築するために、消費者グループごとにマーケティング戦略と商品開発戦略を調整します。

「フィールド」タブ

「フィールド」タブでは、分析で使用するフィールドを指定します。

定義済みの役割を使用。定義済みの役割が入力のフィールドがすべて選択されます。

ユーザー設定フィールドの割り当てを使用定義済みの役割の割り当てにかかわらず、フィールドの追加および削除を行います。任意の役割が割り当てられたフィールドを選択し、「予測変数 **(**入力**)**」リストへの追加、またはリストからの削除を行うことができます。

基本

クラスター数

自動的に判定
: この手続きは、指定した範囲内でクラスターの最適数を判定します。「最小値」は、1 より大きい値を指定する必要があります。これは、デフォルトのオプションです。

固定値を指定
: この手続きは、指定した数のクラスターを生成します。「クラスター数」は、1 より大きい値を指定する必要があります。

クラスタ化の基準

この選択項目では、自動クラスタリング・アルゴリズムでクラスター数を判定する方法を制御します。

ベイジアン情報基準 **(BIC)**
: -2 対数尤度に基づいてモデルを選択し、比較するための指標。値が小さいほどモデルが良好であることを示します。BIC もパラメータが過多のモデル (例えば、大量の入力がある複雑なモデル) にペナルティを科しますが、AIC よりも厳密にそれを行います。

赤池情報量基準 **(AIC)**
: -2 対数尤度に基づいてモデルを選択し、比較するための指標。値が小さいほどモデルが良好であることを示します。AIC は、パラメータが過多のモデル (例えば、大量の入力がある複雑なモデル) にペナルティを科します。

自動クラスタリング メソッド

「自動的に判定」を選択した場合は、クラスター数を自動的に判定するために使用するクラスター メソッドを以下より選択してください。

クラスタリング基準の設定の使用
: 情報の基準の収束は、2 つの現行クラスター解と最初のクラスター解に対応する情報の基準の比率です。「クラスタリング基準」グループで選択した基準が使用されます。

距離ジャンプ
: 距離ジャンプは、2 つの連続するクラスター解に対応する距離の比率です。

最大値
: 情報の基準収束メソッドと距離ジャンプ メソッドからの結果を結合し、2 番目のジャンプに対応するクラスターの数を生成します。

最小値
: 情報の基準収束メソッドと距離ジャンプ メソッドからの結果を結合し、1 番目のジャンプに対応するクラスターの数を生成します。

フィールドの重要度メソッド

「フィールドの重要度メソッド」は、クラスター解での機能 (フィールド) の重要度を判定します。出力には、フィールド全体の重要度と各クラスターの各機能フィールドの重要度に関する情報が含まれます。最小しきい値を満たさない機能は除外されます。

クラスタリング基準の設定の使用。
: これは、「クラスタリング基準」グループで選択した基準に基づくデフォルトのメソッドです。

効果サイズ
: フィールドの重要度は、有意確率の値ではなく、効果サイズに基づきます。

機能ツリー基準

クラスター機能ツリーの作成方法は、以下の設定で決定します。クラスター機能ツリーの作成とレコードの要約を行うと、TwoStep アルゴリズムで大規模なデータ・ファイルを分析できるようになります。つまり、TwoStep クラスターは、クラスター機能ツリーを使用してクラスターを作成し、多数のケースの処理を可能にします。

距離測度

このオプションにより、2 つのクラスター間の類似度を計算する方法を指定します。

対数尤度

この尤度測定により、フィールドの確率分布を求めます。連続フィールドは正規分布しているものと仮定し、カテゴリー・フィールドは多項分布しているものと仮定します。すべてのフィールドは独立しているものと仮定します。

ユークリッド

ユークリッド測定は、2 つのクラスター間の「直線」距離です。ユークリッド平方測定および Ward メソッドは、クラスター間の類似度を計算するために使用されます。この測定方法は、すべてのフィールドが連続している場合にだけ使用できます。

外れクラスタ

外れ値クラスターを含める

通常のクラスターからの外れ値であるケースのクラスターを含めます。このオプションを選択していない場合は、すべてのケースが通常のクラスターに含められます。

機能ツリーのリーフのケース数が次より少ない。

機能ツリーのリーフのケース数が、指定した値より少ない場合、リーフは外れ値と見なされます。この値には、1 より大きい整数を指定する必要があります。この値を変更する場合、値が大きいほど、外れ値クラスターの数が増える可能性があります。

上位外れ値の割合。

クラスター・モデルを作成する場合、外れ値は外れ値の強度でランク付けされます。上位外れ値の割合に含まれる外れ値の強度は、ケースを外れ値として分類するかどうかを判定するためのしきい値として使用されます。値が大きいほど、より多くのケースが外れ値として分類されます。この値には、1 から 100 までの範囲で指定する必要があります。

追加の設定

距離の変化の初期のしきい値

クラスター特性ツリーを成長させるために使用される初期のしきい値。リーフをツリーのリーフに挿入して生じた気密度がこのしきい値より小さい場合、リーフは分割されません。気密度がこのしきい値を超えると、リーフは分割されます。

リーフ ノードの枝の最大数

1 つのリーフ ノードが持つことができる子ノードの最大数。

リーフ ノード以外の枝の最大数

1 つのリーフ ノード以外のノードが持つことができる子ノードの最大数。

最大ツリー深度

クラスター・ツリーが持つことができるレベルの最大数。

尺度に対する重み付けの調整

連続型フィールドの重みを増やすことにより、カテゴリー型フィールドの影響を削減します。この値は、カテゴリー型フィールドの重みを削減するための分母を表します。そのため、たとえば、デフォルト値の 6 では、カテゴリー型フィールドの重みは 1/6 になります。

メモリ割り当て

クラスターのアルゴリズムで使用されるメモリーの最大量 (MB 単位)。この最大値を超えると、手続きはディスクを使用して、メモリー内に納まらない情報を保管します。

遅延分割
: クラスター特性ツリーの再作成を遅延させます。新規ケースの評価時に、クラスタリング・アルゴリズムにより、クラスター特性ツリーの再作成が複数回行われます。このオプションを選択すると、この操作の遅延とツリー再作成の回数削減が行われるため、パフォーマンスを向上させることができます。

標準化

クラスタリング・アルゴリズムは、標準化された連続型フィールドを処理します。デフォルトでは、すべての連続型フィールドが標準化されます。時間の節約やコンピューター処理の省略を行うために、既に標準化された連続型フィールドを「標準化しない」リストに移動させることができます。

特徴量選択

「特徴量選択」画面では、フィールドを除外するタイミングを決定するルールを設定できます。例えば、欠損値が多数あるときにこのフィールドを除外することが可能です。

フィールド除外のルール

欠損値の割合が次より大きい。
: 指定した値よりも欠損値の割合が大きいフィールドが分析から除外されます。この値には、ゼロより大きく 100 未満の正数を指定する必要があります。

カテゴリ型フィールドのカテゴリ数が次より大きい。
: 指定した数よりカテゴリー数が多いカテゴリー型フィールドが分析から除外されます。この値には、1 より大きい正整数を指定する必要があります。

単一値に向かう傾向のあるフィールド

: 連続型フィールドの変動係数が次より小さい。
: 指定した値よりも変動係数が小さい連続型フィールドが分析から除外されます。変動係数は、標準偏差と平均の比率です。値が小さいほど、値の変動が小さくなる傾向があります。値の範囲は 0 から 1 までです。

: カテゴリー型フィールドの **1** つのカテゴリーのケースの割合が次より大きい。
: 指定した値よりも 1 つのカテゴリーのケースの割合が大きいカテゴリー型フィールドが分析から除外されます。この値には、0 より大きく 100 未満の値を指定する必要があります。

適応特徴量選択

このオプションは、追加のデータ・パスを実行し、重要度が最も低いフィールドを検出して削除します。

モデル出力

モデル構築の集計

モデル指定
: モデル指定、最終モデルのクラスター数、最終モデルに含まれる入力 (フィールド) の要約。

レコード要約
: モデルに組み込まれたレコード (ケース) とモデルから除外されたレコード (ケース) の数および割合。

除外された入力
: 最終モデルに組み込まれなかったフィールドがある場合、そのフィールドが除外された理由。

評価

モデル品質
: 各クラスターの適合度と重要度、および全体的なモデル適合度のテーブル。

変数の重要度の棒グラフ
: すべてのクラスターに渡る変数 (フィールド) の重要度の棒グラフ。グラフの棒が長い変数 (フィールド) は、棒が短いフィールドよりも重要度が高くなります。これらの棒は、重要度の降順でソートもされます (先頭の棒が最も重要度が高い)。

変数の重要度のワード クラウド
: すべてのクラスターに渡る変数 (フィールド) の重要度のワード クラウド。テキストが大きい変数 (フィールド) は、テキストが小さい変数 (フィールド) よりも重要度が高くなります。

外れ値クラスター
: 外れ値を含めないことを選択した場合、以下のオプションは無効になります。

  対話式のテーブルと図表
  : 外れ値の強度、および通常のクラスターに対する外れ値クラスターの相対的な類似度のテーブルおよび図表。テーブル内で選択した行により、図表に表示される外れ値クラスターの情報は異なります。

  ピボット・テーブル
  : 外れ値の強度、および通常のクラスターに対する外れ値クラスターの相対的な類似度を示すテーブル。このテーブルには、インタラクティブ表示と同じ情報が含まれます。このテーブルでは、ピボットおよび編集を行うためのすべての標準機能がサポートされています。

  最大数
  : 出力に表示する外れ値の最大数。外れ値クラスターの数が 20 を超える場合は、代わりにピボット・テーブルが表示されます。

解釈

クラスター間フィールドの重要度プロファイル

  対話式のテーブルと図表
  : クラスター解で使用される各入力 (フィールド) のフィールドの重要度およびクラスター中心のテーブルと図表。テーブル内で選択した行により、表示される図表は異なります。カテゴリー型フィールドの場合は、棒グラフが表示されます。連続型フィールドの場合は、平均と標準偏差の図表が表示されます。

  ピボット・テーブル
  : 各入力 (フィールド) のフィールドの重要度およびクラスター中心のテーブル。このテーブルには、インタラクティブ表示と同じ情報が含まれます。このテーブルでは、ピボットおよび編集を行うためのすべての標準機能がサポートされています。

クラスター内フィールドの重要度
: クラスターごとの、各入力 (フィールド) のフィールドの重要度およびクラスター中心。クラスターごとに個別のテーブルがあります。

クラスター距離
: クラスター間の距離を表示するパネル グラフ。クラスターごとに個別のパネルがあります。

クラスター ラベル

テキスト
: 各クラスターのラベルは、「接頭辞」に指定した値に連続番号を続けたものです。

クラスター数
: 各クラスターのラベルは、連続番号です。

モデル オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

TwoStep-AS クラスター・モデル・ナゲット

TwoStep-AS モデル・ナゲットは、出力ビューアーの「モデル」タブにモデルの詳細を表示します。ビューアーの使用方法について詳しくは、「Modeler ユーザーズ・ガイド」(ModelerUsersGuide.pdf) の『出力の処理』を参照してください。

TwoStep-AS クラスター・モデル・ナゲットには、クラスタリング・モデルが取得したすべての情報と、学習データおよび推定プロセスに関する情報が含まれます。

TwoStep-AS クラスター・モデル・ナゲットを含むストリームを実行すると、ノードにより、そのレコードの所属クラスターを含む新規フィールドが追加されます。新規フィールド名はモデル名から派生し、接頭辞の ***$AS-*** が付けられます。例えば、モデルの名前が TwoStep の場合、新規フィールドの名前は ***$AS-TwoStep*** になります。

TwoStep-AS モデルを詳しく調べるための効果的な手法として、ルール算出を使用して、モデルによって検出されたクラスターを区別する特性を発見する方法があります。詳しくは、トピック 110 ページの『C5.0 ノード』を参照してください。

モデル・ブラウザー使用法に関する一般情報については、 43 ページの『モデル・ナゲットの参照』を参照してください。

TwoStep-AS クラスター・モデル・ナゲットの設定

「設定」タブには、TwoStep-AS モデル・ナゲットに関する追加のオプションが用意されています。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- ネイティブの **SQL** に変更してスコアリング**:** これを選択すると、データベース内でモデルをスコアリングするためのネイティブ SQL が生成されます。

注: このオプションの方が短時間で結果を得ることができますが、モデルの複雑度が増大すると、ネイティブ SQL のサイズと複雑度も、それに応じて増大します。

- データベースの外部でスコアリング: このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

K-Means-AS ノード

K-Means は、最も一般的に使用されるクラスタリング アルゴリズムの 1 つです。このアルゴリズムは、データ ポイントをクラスタリングして、事前定義された数のクラスタを作成します。[1] SPSS Modeler の K-Means-AS ノードは Spark で実装されています。

K-Means アルゴリズムについて詳しくは、https://spark.apache.org/docs/2.2.0/ml-clustering.html を参照してください。

K-Means-AS ノードでは、カテゴリ変数の場合にワン ホット エンコーディングが自動的に実行されることに留意してください。

[1] "Clustering." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

K-Means-AS ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

定義済みの役割を使用: このオプションを選択すると、上流のデータ型ノードからのフィールド情報がこのノードで使用されます。これは、デフォルトで選択されます。

カスタム・フィールド割り当ての使用: 入力フィールドを手動で割り当てる場合は、このオプションを選択し、1 つ以上の入力フィールドを選択します。このオプションの使用は、データ型ノードのフィールドの役割を「入力」に設定するのと似ています。

K-Means-AS ノードの作成オプション

K-Means-AS ノードの作成オプションを指定するには、「作成オプション」タブを使用します。このタブには、モデル作成のための通常オプション、クラスタ中心の初期化のための初期化オプション、および計算反復とランダム シードのための詳細オプションが含まれています。詳しくは、JavaDoc for K-Means on SparkML を参照してください。[1]

通常

「モデル名」。特定のクラスタへのスコアリングの後に生成されるフィールドの名前。「自動」(デフォルト) を選択するか、「カスタム」を選択してから名前を入力します。

クラスター数: 生成するクラスタの数を指定します。デフォルト値は **5**、最小値は **2** です。

初期化

初期化モード: クラスタ中心の初期化の方法を指定します。**K-Means|| ** がデフォルトです。これらの 2 つの方法について詳しくは、ScalableK-Means++ を参照してください。[2]

初期化ステップ: **K-Means|| ** 初期化モードが選択されている場合は、初期化ステップの数を指定します。**2** がデフォルトです。

詳細

詳細設定: 詳細オプションを以下のように設定する場合は、このオプションを選択します。

最大反復: クラスタ中心を検索するときに実行する最大反復数を指定します。**20** がデフォルトです。

許容度: 反復アルゴリズムの収束許容度を指定します。**1.0E-4** がデフォルトです。

ランダム シードの設定: 乱数発生ルーチンによって使用されるシードを生成するには、このオプションを選択し、「生成」をクリックします。

表示

グラフの表示: 出力にグラフを含める場合は、このオプションを選択します。

次の表に、SPSS Modeler の K-Means-AS ノードの設定と K-Means Spark パラメータとの間の関係を示します。

表 *13.* ノードのプロパティと *Spark* パラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**K-Means SparkML** パラメータ
入力フィールド	features	
クラスター数	clustersNum	k
初期化モード	initMode	initMode
初期化ステップ	initSteps	initSteps
最大反復	maxIter	maxIter
許容度	toleration	tol
ランダム シード	randomSeed	seed

[1] "Class KMeans." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

[2] Bahmani, Moseley, et al. "Scalable K-Means++." Feb 28, 2012. http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf.

クラスター・ビューアー

通常、クラスター・モデルは、調査された変数に基づく類似レコードのグループ (またはクラスター) の検索に使用されます。同じグループのメンバー間の類似度は高く、異なるグループのメンバー間の類似度は低くなります。この結果を使用して、明らかではなかった関連度を特定することができます。例えば、顧客の嗜好、収入レベル、購買習慣のクラスター分析を通して、特定のマーケティング・キャンペーンに回答する確率が高い顧客のタイプを特定できる場合があります。

クラスター表示での結果を解釈する方法は、次の 2 とおりあります。

- クラスターを調べて、そのクラスターに固有の特性を判別します。*1* つのクラスターに、高収入な借り手がすべて含まれているかどうか。そのクラスターに、他のクラスターよりも多くのレコードが含まれているかどうか。
- クラスター間でフィールドを調べて、値がクラスター間でどのように分布しているかを判別します。クラスター内の所属は、学歴により決定されているか。信用度の高さにより、クラスターごとの所属が区別されているか。

クラスター・ビューアーのメイン・ビューおよび各種リンク・ビューを使用して、これらの質問に答える上で役立つ洞察力を得ることができます。

次のクラスター・モデル・ナゲットを IBM SPSS Modeler で生成できます。

- Kohonen ネット・モデル・ナゲット
- K-Means モデル・ナゲット
- TwoStep クラスター・モデル・ナゲット

クラスター・モデル・ナゲットに関する情報を表示するには、モデル・ノードを右クリックして、コンテキスト・メニューから「参照」(ストリーム内のノードの場合は「編集」) を選択します。または、自動クラスタリング・モデル作成ノードを使用している場合は、自動クラスタリング・モデル・ナゲット内の必要なクラスター・ナゲットをダブルクリックします。詳しくは、トピック 79 ページの『自動クラスタリング・ノード』を参照してください。

クラスター・ビューアー -「モデル」タブ

クラスター・モデルの「モデル」タブには、クラスター間のフィールドの要約統計および分布がグラフィカルに表示されます。これはクラスター・ビューアーと呼ばれます。

注: 13 より前のバージョンの IBM SPSS Modeler で作成されたモデルでは、「モデル」タブは使用できません。

クラスター・ビューアーは 2 つのパネルで構成されており、左側にメイン・ビュー、右側にリンク・ビューまたは補助ビューがあります。メイン・ビューには、以下の 2 つがあります。

- モデルの要約 (デフォルト)。詳しくは、トピック『「モデルの要約」ビュー』を参照してください。
- クラスター。詳しくは、トピック 265 ページの『「クラスター」ビュー』を参照してください。

リンク/補助ビューには、以下の 4 つがあります。

- 予測値の重要度。詳しくは、トピック 266 ページの『「クラスター予測値の重要度」ビュー』を参照してください。
- クラスター・サイズ (デフォルト)。詳しくは、トピック 266 ページの『「クラスター・サイズ」ビュー』を参照してください。
- セルの分布。詳しくは、トピック 267 ページの『「セルの分布」ビュー』を参照してください。
- クラスターの比較。詳しくは、トピック 267 ページの『「クラスターの比較」ビュー』を参照してください。

「モデルの要約」ビュー

「モデルの要約」ビューには、陰影によって結果 (悪い、普通、良い) が示される、クラスターの結合および分離のシルエット測定など、クラスター・モデルのスナップショット、つまり要約が表示されます。このスナップショットを使用すると、品質が悪いかどうかを素早く確認できます。品質が悪い場合は、より良い結果を生成するために、モデル作成ノードに戻ってクラスター・モデルの設定を修正する決断を下すことができます。

結果 (悪い、普通、良い) は、クラスター構造の解釈に関する Kaufman と Rousseeuw (1990) の研究に基づいています。「モデルの要約」ビューでは、良い結果は Kaufman と Rousseeuw の評価をクラスター構造の合理的または強力な証拠として反映するデータに相当し、普通の結果は弱い証拠の評価を反映するデータに相当し、悪い結果は重要な証拠のない評価を反映するデータに相当します。

すべてのレコードに対するシルエット測定平均は (B-A) / max(A,B) となります。A はレコードのクラスター中心へのレコードの距離、B はレコードが属さない最近隣のクラスター中心へのレコードの距離です。シルエット係数 1 は、すべてのケースが自身のクラスター中心の真上にあることを意味します。値 –1 は、すべてのケースが他のクラスターのクラスター中心にあることを意味します。平均の 0 の値は、ケースが自身のクラスター中心と、その他の最近隣のクラスター中心から等距離にあることを意味します。

要約には、次の情報について示すテーブルも含まれます。

- アルゴリズム: 使用されたクラスタリング・アルゴリズム (例えば「TwoStep」)。
- 入力フィールド: 入力または予測とも呼ばれる、フィールドの数。
- クラスター: 解のクラスター数。

「クラスター」ビュー

「クラスター」ビューには、各クラスターのクラスター名、サイズ、プロファイルが含まれた、クラスターとフィールドのグリッドがあります。

グリッドの列には、次の情報が含まれています。

- クラスター: アルゴリズムによって作成されたクラスター番号。
- ラベル: 各クラスターに適用されるラベル (デフォルトではブランクです)。セル内をダブルクリックし、クラスターの内容を説明するラベル (例えば「高級車購入者」) を入力します。
- 説明: クラスターの内容の説明 (デフォルトではブランクです)。セル内をダブルクリックし、クラスターの説明 (例えば「55 歳以上、専門職、収入 $100,000 以上」) を入力します。
- サイズ: クラスター・サンプル全体のパーセントとしての各クラスターのサイズ。グリッド内の各サイズのセルには、クラスター内のサイズのパーセントを示す垂直バー、数値形式によるサイズのパーセント、クラスター・ケース度数が表示されます。
- フィールド: 個別の入力または予測。デフォルトでは全体の重要度でソートされています。サイズが等しい列が複数ある場合、それらはクラスター番号の昇順でソートされて表示されます。

  フィールド全体の重要度は、セルの背景の陰影の色で示されます。最も重要なフィールドは濃く、最も重要でないフィールドは陰影なしとなります。テーブルの上のガイドは、各フィールドのセルの色に関連する重要度を示します。

セル上にマウスを移動すると、フィールドの完全名/ラベルとセルの重要度の値が表示されます。ビューおよびフィールドのタイプによっては、より詳細な情報が表示されます。「クラスター中心」ビューの場合、この情報には「平均: 4.32」など、セル統計量やセル値が含まれます。カテゴリー・フィールドの場合、セルは最も頻度の高い (最頻) カテゴリーの名前とそのパーセントを示します。

「クラスター」ビュー内では、以下のさまざまな方法を選択して、クラスター情報を表示できます。

- クラスターとフィールドの入れ替え。詳しくは、トピック『クラスターとフィールドの入れ替え』を参照してください。
- フィールドのソート。詳しくは、トピック 266 ページの『フィールドのソート』を参照してください。
- クラスターのソート。詳しくは、トピック 266 ページの『クラスターのソート』を参照してください。
- セル内容の選択。詳しくは、トピック 266 ページの『セル内容』を参照してください。

クラスターとフィールドの入れ替え: デフォルトでは、クラスターは列として、フィールドは行として表示されます。この表示を逆にするには、「フィールドのソート基準」ボタンの左にある「クラスターとフィールドを入れ替え」ボタンをクリックします。例えば、表示されるクラスターが多い場合にこれを行うと、データを表示するために必要となる横方向のスクロール量を減らすことができます。

フィールドのソート: 「フィールドのソート基準」ボタンを使用すると、どのようにフィールド・セルが表示されるかを選択できます。

- 全体の重要度: これは、デフォルトのソート順序です。フィールドは全体の重要度の降順にソートされ、ソート順序はクラスター間で同じです。同じ重要度の値を持つフィールドがある場合、それらのフィールドは、フィールド名の昇順でソートされてリストされます。
- クラスター内重要度: フィールドは、各クラスターのフィールドの重要度に応じてソートされます。同じ重要度の値を持つフィールドがある場合、それらのフィールドは、フィールド名の昇順でソートされてリストされます。このオプションを選択すると、通常、ソート順序はクラスターによって異なります。
- 名前: フィールドは、名前のアルファベット順にソートされます。
- データ順: フィールドは、データ・セット内のフィールドの順序でソートされます。

クラスターのソート: デフォルトでは、クラスターはサイズの降順でソートされています。「クラスターのソート基準」ボタンを使用すると、名前のアルファベット順にクラスターをソートできます。または固有のラベルを作成してある場合は、代わりにラベルの英数字順にソートできます。

同じラベルを持つフィールドは、クラスター名でソートされます。クラスターがラベルでソートされている場合にクラスターのラベルを編集すると、ソート順序は自動的に更新されます。

セル内容: 「セル」ボタンを使用すると、フィールドおよび評価フィールドのセル内容の表示を変更できます。

- クラスター中心: デフォルトでは、セルには、フィールド名/ラベルと各クラスター/フィールドの組み合わせの中心傾向が表示されます。連続型フィールドの場合は平均値が表示され、カテゴリー・フィールドの場合は最頻値 (最も頻繁に発生するカテゴリー) がカテゴリー・パーセントとともに表示されます。
- 絶対分布: フィールド名/ラベルと各クラスター内のフィールドの絶対分布が表示されます。カテゴリー・フィールドの場合、データ値が昇順に並んでいるカテゴリーを重ね合わせた棒グラフが表示されます。連続型フィールドの場合、各クラスターに対して同じエンドポイントと区間を使用する平滑密度プロットが表示されます。

  濃い赤はクラスター分布を示し、薄い赤は全体のデータを表します。
- 相対分布: フィールド名/ラベルと相対分布がセルに表示されます。一般的に、相対分布が代わりに表示されるという点を除いて、絶対分布の表示と類似しています。

  濃い赤はクラスター分布を示し、薄い赤は全体のデータを表します。
- 基本ビュー: クラスターが多いと、スクロールせずにすべての詳細を確認するのは難しい場合があります。スクロールの量を減らすために、このビューを選択して、よりコンパクトなバージョンのテーブルに表示を変更します。

「クラスター予測値の重要度」ビュー

「予測値の重要度」ビューには、モデルの推定における各フィールドの相対重要度が表示されます。

「クラスター・サイズ」ビュー

「クラスター・サイズ」ビューには、各クラスターが含まれた円グラフが表示されます。各クラスターのサイズのパーセントが各スライスに表示されます。各スライス上にマウスを移動すると、そのスライスに度数が表示されます。

グラフの下のテーブルには、以下のサイズ情報がリストされます。

- 最小クラスターのサイズ (度数と全体の割合)
- 最大クラスターのサイズ (度数と全体の割合)
- 最大クラスターの最小クラスターに対するサイズの比率

「セルの分布」ビュー

「セルの分布」ビューには、「クラスター」メイン・パネルのテーブルで選択したフィールド・セルのデータの分布に関する、拡張された、より詳細なプロットが示されます。

「クラスターの比較」ビュー

「クラスターの比較」ビューは、グリッド・スタイルのレイアウトで構成され、フィールドは行に、選択したクラスターは列に表示されます。このビューは、クラスターを構成する要素をより良く理解する上で役立ちます。また、全体のデータと比較するだけでなく、クラスター同士で比較して、クラスター間の差分を表示することもできます。

表示するクラスターを選択するには、「クラスター」メイン・パネルでクラスター列の一番上をクリックします。Ctrl キーまたは Shift キーを押しながらクリックすることで、比較対象として複数のクラスターを選択または選択解除できます。

注: 表示するクラスターは、5 個まで選択できます。

クラスターは選択した順に表示されます。一方でフィールドの順序は、「フィールドのソート基準」で決まります。「クラスター内重要度」を選択した場合、フィールドは常に全体の重要度でソートされます。

背景のプロットには、各フィールドの全体の分布が表示されます。

- カテゴリー・フィールドはドット・プロットとして表示されます。ドットのサイズは、フィールドごとの各クラスターの最も頻度の高い (最頻) カテゴリーを示します。
- 連続型フィールドは箱ひげ図として表示され、全体の中央値と 4 分位範囲を示します。

これらの背景ビューに、選択したクラスターの箱ひげ図が重ね合わせて表示されます。

- 連続型フィールドの場合、四角形のポイント・マーカーと水平線は、それぞれ各クラスターの中央値と 4 分位範囲を示します。
- 各クラスターは異なる色で表され、ビューの最上位に表示されます。

クラスター・ビューアーのナビゲート

クラスター・ビューアーはインタラクティブ表示です。以下を行うことができます。

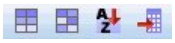
- フィールドまたはクラスターを選択して詳細を表示する。
- クラスターを比較して関心のある項目を選択する。
- 表示を変更する。
- 軸を入れ替える。
- 「生成」メニューを使用して、フィールド生成ノード、フィルター・ノード、条件抽出ノードを生成する。

ツールバーの使用

ツールバー・オプションを使用して、左右のパネルに表示される情報を制御します。ツールバー・コントロールを使用して、表示方向 (上から下、左から右、右から左) を変更できます。また、ビューアーをデフォルト設定にリセットし、ダイアログ・ボックスを開いて、メイン・パネルの「クラスター」ビューの内容を指定することもできます。

「フィールドのソート基準」、「クラスターのソート基準」、「セル」、「表示」の各オプションは、メイン・パネルで「クラスター」ビューを選択した場合にのみ使用できます。詳しくは、トピック 265 ページの『「クラスター」ビュー』を参照してください。

表 14. ツールバーのアイコン

アイコン	トピック
	『クラスターとフィールドの入れ替え』を参照
	『フィールドのソート基準』を参照
	『クラスターのソート基準』を参照
	『セル』を参照

クラスター・モデルからのノードの生成

「生成」メニューを使用すると、クラスター・モデルに基づいて新しいノードを作成できます。このオプションは、生成されたモデルの「モデル」タブから使用でき、このオプションを使用して、現在の表示または選択 (表示されるすべてのクラスターまたは選択したすべてのクラスター) に基づいてノードを生成できます。例えば、1 つのフィールドを選択し、フィルター・ノードを生成して、その他すべての (表示されない) フィールドを破棄します。生成されたノードは、領域に未接続の状態で配置されます。また、モデル・ナゲットのコピーをモデル・パレットに生成できます。実行前に必ずノードを接続して、必要な編集を行ってください。

- モデル作成ノードの生成: ストリーム領域にモデル作成ノードを作成します。これは、例えばストリームでこれらのモデル設定を使用する必要があるが、モデルの生成に使用するモデル作成ノードがない場合に役立ちます。
- パレットのモデル: モデル・パレットにナゲットを作成します。これは、同僚から、モデル自体ではなくモデルを含むストリームが送信されてきた場合に役立ちます。
- フィルター・ノード: 新しいフィルター・ノードを作成して、クラスター・モデルで使用されていないフィールド、または現在のクラスター・ビューアー表示に表示されないフィールド、あるいはその両方をフィルタリングします。このクラスター・ノードの上流にデータ型ノードがある場合、役割が「対象」のフィールドはすべて、生成されたフィルター・ノードによって破棄されます。
- フィルター・ノード (選択から): 新しいフィルター・ノードを作成して、クラスター・ビューアーでの選択内容に基づいてフィールドをフィルタリングします。Ctrl キーを押しながらクリックする方法を使用して、複数のフィールドを選択します。クラスター・ビューアーで選択されたフィールドは下流で破棄されますが、実行前にフィルター・ノードを編集することにより、この動作を変更できます。

- 条件抽出ノード: 新しい条件抽出ノードを作成して、現在のクラスター・ビューアー表示に表示されるクラスターの所属に基づいてレコードを選択します。選択条件は自動的に生成されます。
- 条件抽出ノード **(**選択から**)**: 新しい条件抽出ノードを作成して、クラスター・ビューアーで選択されたクラスターの所属に基づいてレコードを選択します。Ctrl キーを押しながらクリックする方法を使用して、複数のクラスターを選択します。
- フィールド作成ノード: 新しいフィールド作成ノードを作成して、クラスター・ビューアーに表示されるすべてのクラスターの所属に基づいて *True* または *False* の値をレコードに割り当てるフラグ・フィールドを派生させます。派生条件は自動的に生成されます。
- フィールド作成ノード **(**選択から**)**: 新しいフィールド作成ノードを作成して、クラスター・ビューアーで選択されたクラスターの所属に基づいてフラグ・フィールドを派生させます。Ctrl キーを押しながらクリックする方法を使用して、複数のクラスターを選択します。

「生成」メニューからは、ノードを生成するほかにグラフを作成することもできます。 詳しくは、トピック『クラスター・モデルからのグラフの生成』を参照してください。

「クラスター」ビュー表示の制御

メイン・パネルの「クラスター」ビューの表示内容を制御するには、「表示」ボタンをクリックします。これにより、「表示」ダイアログが開きます。

フィールド: デフォルトで選択されています。すべての入力フィールドを非表示にするには、チェック・ボックスを選択解除します。

評価フィールド: 表示する評価フィールド (クラスター・モデルの作成には使用されないが、クラスターの評価のためにモデル・ビューアーに送信されるフィールド) を選択します。デフォルトでは表示される評価フィールドはありません。注: 評価フィールドは、複数の値が含まれた文字列でなければなりません。使用可能な評価フィールドがない場合、このチェック・ボックスは選択できません。

クラスターの説明: デフォルトで選択されています。すべてのクラスターの説明のセルを非表示にするには、チェック・ボックスを選択解除します。

クラスター・サイズ: デフォルトで選択されています。すべてのクラスター・サイズのセルを非表示にするには、チェック・ボックスを選択解除します。

カテゴリーの最大数: カテゴリー・フィールドのグラフに表示するカテゴリーの最大数を指定します。デフォルトは 20 です。

クラスター・モデルからのグラフの生成

クラスター・モデルは多くの情報を提供します。ただし、ビジネス・ユーザーにとっては、必ずしも使用しやすい形式ではありません。ビジネス・レポート、プレゼンテーションなどに用意に組み込むことができる方法でデータを提供するために、選択したデータのグラフを作成できます。例えば、クラスター・ビューアーから選択したクラスターのグラフを生成できます。つまり、そのクラスターのケースのグラフだけを生成します。

注 : モデル・ナゲットをストリーム内のその他のノードに接続する場合にのみ、クラスター・ビューアーからグラフを生成できます。

グラフの生成

1. クラスター・ビューアーを含むモデル・ナゲットを開きます。

2. 「モデル」タブの「表示」ドロップダウン・リストから「クラスター」を選択します。
3. メイン・ビューで、グラフを作成するクラスターを選択します。
4. 「ノードの生成」メニューで 「グラフ **(**選択項目から**)**」 を選択します。グラフボードの「基本」タブが表示されます。

   注 : この方法でグラフボードを表示した場合、「基本」タブと「詳細」タブのみを使用できます。
5. 「基本」タブまたは「詳細」タブいずれかの設定を使用し、グラフに表示する詳細を指定します。
6. 「OK」をクリックしてグラフを生成します。

グラフの見出しは、選択されたモデル タイプおよびクラスターを示します。

第 12 章 アソシエーション・ルール

アソシエーション・ルールは、特定の結果 (特定の製品の購入など) と条件セット (複数の他の製品の購入など) を関連付けます。例えば、次のルール

```
beer <= cannedveg & frozenmeal (173, 17.0%, 0.84)
```

は、缶詰野菜と冷凍食品を同時に購入するときに、ビールがよく購入されることを示しています。このルールは信頼度 84% で、データの 17 %、つまり 173 個のレコードにあてはまります。アソシエーション・ルールのアルゴリズムは、ユーザーが Web グラフ・ノードなどの視覚化手法を使用して手動で見つけていた連関を、自動的に見つけ出します。

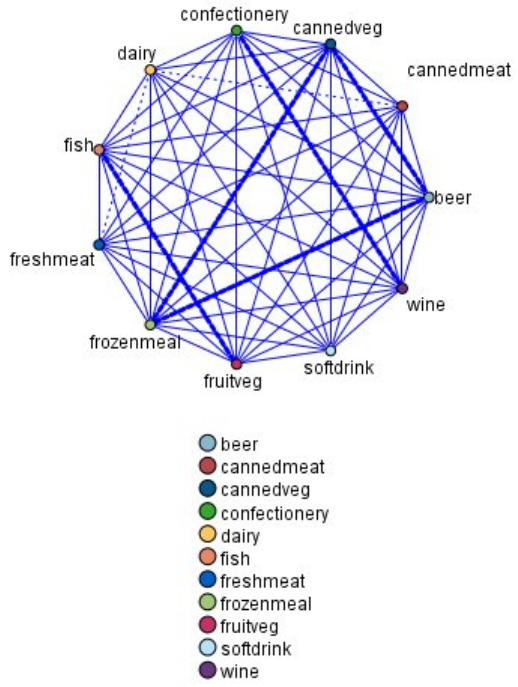


図 *45.* 買い物かごの品物の連関を示す *Web* グラフ・ノード

アソシエーション・ルールのアルゴリズムは、任意の 属性の間にアソシエーションを成立させることができるという点で、より一般的なディシジョン・ツリーのアルゴリズム (C5.0 や C&R Trees など) より勝っています。ディシジョン・ツリーのアルゴリズムは、一つの結果にいたるルールを構築するのに対し、アソシエーション・ルールのアルゴリズムは、それぞれが異なる結果にいたる多数のルールを見つけようとします。

アソシエーション・ルール・アルゴリズムは、パターンの検索範囲が非常に大きくなる可能性があり、そのためディシジョン・ツリーのアルゴリズムを実行するよりもはるかに時間がかかるという欠点があります。アソシエーション・ルール・アルゴリズムでは、ルール検索に生成と検定の手法を使用します。最初に簡単なルールが生成され、データセットに対して妥当性が検証されます。有効なルールは保存され、その後でさまざまな制約に従って、すべてのルールが特殊化されます。特殊化とは、条件をルールに追加する処理のことです。次に、データに対して新しいルールの妥当性が検証され、この繰り返しによって、最善または最も

関心の高いルールが保存されます。通常、ユーザーは、ルールで許可する推定数を制限し、情報理論に基づく各種の手法や効果的なインデックス作成方法を使用して、広大になる可能性がある検索範囲を縮小します。

処理の最後に、最善のルールのテーブルが表示されます。ディシジョン・ツリーとは異なり、このアソシエーション・ルールは、標準モデル (ディシジョン・ツリーやニューラル・ネットワークなど) のように、直接予測に使用することはできません。このルールには、可能性のある結果が多数存在するからです。アソシエーション・ルールを分類ルール・セットに変換するには、別のレベルの変換が必要です。そのため、アソシエーション・ルール・アルゴリズムで生成されたアソシエーション・ルールは、未精製モデルと呼ばれます。ユーザーは、これらの未精製モデルを参照できますが、未精製モデルから分類モデルを生成するように操作しない限り、これらのモデルを分類モデルとして明示的に使用することはできません。この操作は、「ノードの生成」メニュー・オプションを使用して、ブラウザーから実行できます。

次の 2 つのアソシエーション・ルール・アルゴリズムがサポートされています。

Apriori ノードで、データからルール・セットを抽出し、情報内容が最も充実したルールを引き出します。Apriori には、5 種類のルール選択方法があり、高度なインデックス作成方法を使用して、大きなデータ・セットが効率的に処理されます。大きな問題の場合は、一般に、Apriori の方が高速に学習できます。保持できるルール数に特に制限はありません。また、最大 32 の前提条件を持つルールを処理できます。Apriori では、入力フィールドと出力フィールドのすべてがカテゴリーであることが必要ですが、この種類のデータに合わせて最適化されているので、よりよいパフォーマンスを実現します。

シーケンス・ノードで、シーケンシャルな、または時間経過が伴うデータ内のアソシエーション・ルールを検出します。予測可能な順序で起こる傾向にあるアイテム・セットのリストを、シーケンスと呼びます。例えば、顧客がひげそりとアフター・シェーブ ローションを購入した場合、その顧客は次の購入時にシェービング クリームを購入する可能性があります。シーケンス・ノードは CARMA アソシエーション・ルール・アルゴリズムに基づいているため、効率的な 2 段階通過法でシーケンスが検出されます。

テーブル形式データとトランザクション形式・データ

アソシエーション・ルール・モデルで使用されるデータは、以下に説明するように、トランザクション形式でもテーブル形式でもかまいません。これらは一般的な説明であり、特定の要件は、各モデル タイプのドキュメンテーションで説明されているとおりに多様です。モデルのスコアリング時に、スコアリングされるデータは、モデルを構築するために使用されたデータの形式と同一である必要があります。テーブル形式データを使用して構築されたモデルは、テーブル形式のデータだけをスコアリングするのに使用できます。トランザクション形式のデータを使用して構築されたモデルは、トランザクション形式のデータだけをスコアリングできます。

トランザクション形式の形式

トランザクション形式のデータには、各トランザクションまたは項目に対応する独立したレコードがあります。例えば、顧客が複数の買い物をした場合、それぞれが顧客 ID にリンクされた項目に関連付けられた、独立したレコードになります。ペーパーロール形式とも呼ばれます。

顧客	購入品
1	ジャム
2	牛乳

顧客	購入品
3	ジャム
3	bread (パン)
4	ジャム
4	bread (パン)
4	牛乳

Apriori、CARMA、およびシーケンスの各ノードではすべて、トランザクション形式のデータを使用できます。

テーブル形式のデータ

テーブル形式のデータ (バスケットまたは真理値表データとも呼ばれる) には、フラグで区切られて表現された項目があります。各フラグ型フィールドで、特定の項目の有無が表現されます。各レコードで、関連付けられている項目の完全セットが表現されます。フラグ型フィールドは、カテゴリーまたは数値とすることができます。ただし、ある種のモデルでは、さらに特定の要件があります。

顧客	ジャム	パン	牛乳
1	T	F	F
2	F	F	T
3	T	T	F
4	T	T	T

Apriori ノード、CARMA ノード、GSAR ノード、およびシーケンス・ノードでは、いずれもテーブル形式のデータを使用することができます。

Apriori ノード

Apriori ノードも、 データ中のアソシエーション・ルールを発見します。Apriori には、ルール選択方法が 5 つあります。高度なインデックス作成方法を使用して、大きなデータ・セットが効率的に処理されます。

要件: Apriori ルール・セットを作成するには、1 つ以上の入力 フィールドと 1 つ以上の対象 フィールドが必要です。入力フィールドおよび出力フィールド (役割が入力、対象、または両方のフィールド) はシンボル値でなければなりません。役割が「なし」のフィールドは無視されます。フィールド・タイプは、ノードを実行する前に完全にインスタンス化する必要があります。データはテーブル形式またはトランザクション形式が可能です。 詳しくは、トピック 272 ページの『テーブル形式データとトランザクション形式・データ』を参照してください。

利点: 大きな問題の場合、通常は Apriori の方が学習速度が速くなります。. 保持できるルール数に特に制限はありません。また、最大 32 の前提条件を持つルールを処理できます。Apriori には 5 種類の学習方法があるので、データ・マイニング手法をより柔軟に問題に適合させることができます。

Apriori ノードの「モデル」オプション

モデル名: ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

最小前提条件サポート**:** ルールをルールセットに保持する際のサポート (範囲) 基準を指定できます。サポート **(**範囲**)**は、前提条件 （if 文） が真 (true) の学習データ中のレコードの割合を表しています(このサポートの定義は、CARMA およびシーケンス・ノードで使われているものとは異なります詳しくは、トピック 290 ページの『シーケンス・ノードの「モデル」オプション』を参照してください。) データの非常に小さいサブセットに適用するルールを生成する場合は、この設定値を上げます。

注**:** Apriori のサポートの定義は、前提条件を持つレコードの数に基づきます。このことは、範囲の定義がルール (つまり先行条件と結果の両方) 中の全てのアイテムを持つレコードの数に基づく CARMA およびシーケンス・アルゴリズムとは異なります。アソシエーション・モデルの結果は (前提条件) サポートとルール・サポート (範囲) の測定値の両方を示します。

最小ルール確信度**:** 確信度の基準を指定できます。確信度は、ルールの前提条件が真のレコードの中で、結果も真 (true) のレコードの割合です。つまり、正しいルールをベースにした予測の割合です。（削除）ルール数が多すぎる場合は、設定値を増やしてください。ルールが少なすぎる場合 (またはない場合) は、設定値を減らしてください。

注**:** 必要に応じ、値を強調表示して、独自の値を入力できます。確信度を 1.0 未満に下げると、プロセスに多くの空きメモリーが必要になるだけでなく、ルールの作成に極端に長い時間がかかる場合があります。

最大前提条件数**:** 任意のルールに対する前提条件の最大数を指定できます。これにより、ルールの複雑さを制限します。ルールが複雑すぎる場合や詳細すぎる場合は、この設定を下げてみてください。この設定は、学習時間にも大きく影響します。ルールセットの学習に時間がかかる場合は、設定を下げてみてください。

フラグは真の値のみ**:** このオプションをテーブル形式 (真理値表) のデータで選択すると、結果のルールには真 (true) の値だけが表示されます。これにより、ルールが理解しやすくなります。このオプションは、トランザクション形式のデータには適用されません。 詳しくは、トピック 272 ページの『テーブル形式データとトランザクション形式・データ』を参照してください。

注**:** CARMA モデル構築ノードでは、フィールド・タイプがフラグである場合に、モデルの構築時に空のレコードが無視されますが、Apriori モデル構築ノードでは、空のレコードも処理の対象となります。空のレコードとは、モデル構築に使用されるすべてのフィールドの値が false であるレコードです。

最適化**:** 特定のニーズに応じて、モデルの構築時にパフォーマンスを向上させるために設計されたオプションを選択します。

- パフォーマンス向上のために処理過程のデータをディスクへ書き出さないようにアルゴリズムに指示する場合は、「速度」 を選択します。
- ある程度は速度が遅くなっても処理過程のデータをディスクへ書き出すようにアルゴリズムに指示するには、「メモリー」 を選択します。デフォルトでは、このオプションが選択されます。

  注**:** 分散モードで実行する場合、この設定は、*options.cfg* 内で指定された管理者オプションによってオーバーライドされることがあります。詳しくは、「*IBM SPSS Modeler Server Administrator's Guide*」を参照してください。

Apriori ノードのエキスパート・オプション

Apriori の操作をよく理解している場合は、次のエキスパート・オプションを使用して、算出過程を調整できます。エキスパート・オプションを利用するには、「エキスパート」タブで「モード」に 「エキスパート」 を設定してください。

評価測定 **:** Apriori には、ルール候補を評価するために 5 種類の方法が用意されています。

- ルール確信度 **:** このオプションはデフォルトの設定で、ルールの確信度 (または精度) でルールが評価されます。この方法では、「モデル」タブの「最小ルール確信度」 オプションがあるため、不要な「評価測定下限値」 は無効になっています。 詳しくは、トピック 273 ページの『Apriori ノードの「モデル」オプション』を参照してください。
- 確信度との差異 **:** (事前確信度との差の絶対値とも呼ばれます。)この評価測定は、ルールの確信度とその前の確信度の差異の絶対値です。このオプションを使用すると、結果が均等に分布しないような偏りがなくなります。これにより「明らかな」ルールが保持されるのを防止できます。例えば、顧客の 80% が最も人気のある製品を購入する場合がこれに当たります。ルールによりその人気製品の購入が 85% Ì 精度で予測されたとしても、精度的にはかなり優れているように思えますが、新しい知識をもたらしてくれる訳ではありません。評価測定下限値を、ルールを保持する確信度の最小差に設定します。
- 確信度の比 **:** (確信係数と **1** の差異とも呼ばれます。)この評価測定は、ルールの確信度と前の確信度の比 (比が 1 より大きい場合は、その逆数) を 1 から減算したものになります。確信度の差異のように、この方法は不均衡な分布が強調されます。この値は、稀なイベントを予測するルールを見つける場合に特に適しています。 例えば、患者の 1% にしか発生しない稀な症状があるとします。この症状を 10% の精度で予測できるルールは、絶対的な尺度ではそれほど高い精度ではありません。しかし、このケースでは無作為の推量に比べて格段に優れているといえます。評価測定下限値を、ルールを保持する差異に設定します。
- 情報の差 **:** (事前確信度との情報の差とも呼ばれます。)この測定は、情報の対応測定値に基づきます。特定の結果の確率を論理値 (ビット) と見なす場合、情報の対応ではそのビットの何割が前提条件に基づいて決定されるかが示されます。情報の差は、前提条件が与えられた場合と、結果の以前の確信度のみが与えられた場合の情報の対応の差です。この方法の重要な特徴は、特定レベルの確信度で、より多くのレコードをカバーするルールが優先されるように、範囲を考慮していることです。評価測定の下限を、ルールを保持する情報の差に設定します。

  注: この測定値の尺度は、他の尺度よりも抽象的なので、適切なルールセットを取得するためにはいろいろな下限値を試す必要があることもあります。
- カイ **2** 乗値の正規化 **:** (カイ **2** 乗値の正規化の測定とも呼ばれます。)この測定は、前提条件と結果間の連関を示す統計指標です。測定値は、0〜1 の値となるように正規化されます。この測定は、情報の差の測定よりもさらに強く範囲に依存しています。評価測定の下限を、ルールを保持する情報の差に設定します。

  注: 情報の差の測定値と同様に、この測定値の尺度は他の尺度より抽象的なため、適切なルールセットを取得するためにはいろいろな下限値を試す必要があることもあります。

前提条件を持たないルールを許可。結果 (アイテムまたはアイテムのセット) のみを含むルールを許可するときに選択します。これは、共通アイテムまたはアイテムのセットを決定するために調査する場合に役立ちます。例えば、cannedveg は、缶詰野菜 の購入がデータ中に一般的に発生することを示す、前提条件のない単一アイテム・ルールです。場合によっては、最も確率の高い予測操作のみに注目する場合、このようなルールを含めることができます。このオプションは、デフォルトではオフになっています。表記方法により、前提条件サポートのないルールの前提条件サポートは 100% として表示され、ルール範囲は確信度と同じになります。

CARMA ノード

CARMA ノードは、アソシエーション・ルール検出アルゴリズムを使用して、データ内のアソシエーション・ルールを検出します。アソシエーション・ルールは、次の形式のステートメントです。

if *antecedent(s)* **then** *consequent(s)*

例えば、Web 顧客がワイヤレス・カードおよびハイエンド ワイヤレス・ルータを購入する場合、ワイヤレス音楽サーバーを提案すれば、その顧客が購入する可能性も高いものになります。CARMA モデルは、入力または対象フィールドを指定しなくても、データからルールのセットを抽出します。つまり、生成したルールは広範囲に利用できるということです。例えば、このノードが生成したルールは、この休暇シーズンに販売促進する項目が結果となる、商品またはサービス (前提条件) のリストを調べるのに利用できます。IBM SPSS Modeler を使用して、どの顧客が前提条件商品を購入したかを判断し、結果商品を販売促進するマーケティング・キャンペーンを構築できます。

要件: Apriori とは異なり、CARMA ノードでは「入力」フィールドや「対象」フィールドは必要ありません。これはアルゴリズムが作用する上で非常に重要で、「両方」 に設定されているすべてのフィールドを持つ Apriori モデルを構築することと同じです。構築後にモデルをフィルタリングすることによって、どの項目が前提条件または結果としてのみ現れるか制御できます。例えば、モデル・ブラウザーを使用して、この休暇シーズンに販売促進する項目を結果とする、商品またはサービス (前提条件) のリストを調べるのに利用できます。

CARMA ルールセットを作成するには、ID フィールドと 1 つ以上の内容フィールドを指定します。ID フィールドの役割や測定の尺度はどれでもかまいません。役割が「なし」のフィールドは無視されます。フィールド・タイプは、ノードを実行する前に完全にインスタンス化する必要があります。Apriori のように、データはテーブル形式またはトランザクション形式が可能です。 詳しくは、トピック 272 ページの『テーブル形式データとトランザクション形式・データ』を参照してください。

利点: CARMA ノードは CARMA アソシエーション・ルール・アルゴリズムに基づいています。 Apriori とは対照的に、CARMA ノードは、前提条件サポートではなく、ルール・サポート (前提条件と結果の両方のサポート) の構築の設定ができます。CARMA は複数の結果を持つルールも許可します。Apriori のように、CARMA ノードによって生成されたモデルをデータ・ストリームに挿入して、予測を行なうことができます。 詳しくは、トピック 38 ページの『モデル・ナゲット』を参照してください。

CARMA ノードのフィールド・オプション

CARMA ノードを実行する前に、CARMA ノードの「フィールド」タブで、入力フィールドを指定する必要があります。モデル作成ノードのほとんどが同じ「フィールド」タブの設定ですが、CARMA ノードにはいくつかの固有のオプションがあります。次にすべてのオプションについて解説します。

データ型ノードの設定を使用**:** このオプションを選択すると、上流のデータ型ノードからのフィールド情報がこのノードで使用されます。これがデフォルトです。

ユーザー設定を使用**:** このオプションを選択すると、上流のデータ型ノードからのフィールド情報ではなく、ここで指定したフィールド情報がこのノードで使用されます。このオプションを選択した後で、読み取るデータがトランザクション形式かテーブル形式かに応じて、次のフィールドを指定します。

トランザクション形式を使用 **:** データがテーブル形式とトランザクション形式かに応じて、このオプションはダイアログ・ボックスの残りのフィールドの設定が変わります。トランザクション形式のデータで複数のフィールドを使用している場合、あるレコードのフィールドで言及されているアイテムはすべて、単一のタイム・スタンプを使った単一のトランザクションで検出されたものとみなされます。 詳しくは、トピック 272 ページの『テーブル形式データとトランザクション形式・データ』を参照してください。

テーブル形式データ

「トランザクション形式を使用」 が選択されていない場合、次のフィールドが表示されます。

- 入力**:** 1 つ以上の入力フィールドを選択します。これは、データ型ノードのフィールドの役割を 「入力」 に設定するのと似ています。

- データ区分**:** このフィールドでは、モデル構築の学習、テスト、および検証の各ステージ用に、データを独立したサブセット (サンプル) に分割するフィールドを指定できます。1 組のサンプルをモデルの生成に使用し、別のサンプルで生成したモデルをテストすることにより、そのモデルが、このデータに似た性質を持つより大きなデータセットにどの程度適用できるかについての良い目安を得ることができます。データ型ノードまたはデータ区分ノードを使用して複数のデータ区分フィールドが定義されている場合、データ区分を使用する各モデル作成ノードの「フィールド」タブで単一のデータ区分フィールドを選択する必要があります (1 つのデータ区分だけが存在している場合、データ分割を有効にすると、そのデータ区分が必ず自動的に使用されます)。また、選択したデータ区分を分析に適用するには、そのノードの「モデル・オプション」タブでデータ区分が有効になっている必要があります (このオプションの選択を解除すると、フィールド設定を変更することなくデータ区分を無効にすることができます)。

トランザクション形式のデータ

「トランザクション形式を使用」 が選択されている場合、次のフィールドが表示されます。

- **ID:** トランザクション形式データの場合は、リストから ID フィールドを選択します。ID フィールドとして使用できるのは、数値またはシンボル値のフィールドです。選択したフィールドでは、一意の値がそれぞれ、ある分析ユニットを示している必要があります。例えば、マーケット・バスケット分析なら、各 ID が 1 人の顧客を表します。Web ログ分析なら、各 ID が 1 台のコンピューター (IP アドレス) あるいは 1 人のユーザー (ログイン・データ) を表します。
- 連続する **ID :** (Apriori ノードおよび CARMA ノードのみ) データ・ストリーム中で同じ ID を持つすべてのレコードが一緒に表示されるようにデータをソートしている場合、このオプションを選択すると処理を高速化することができます。データがあらかじめソートされていない場合 (またはわからない場合) は、このオプションは選択しないでください。この場合、ノードが自動的にデータをソートします。

  注**:** データがソートされていない場合にこのオプションを選択すると、モデルで意味のない結果しか得られない可能性があります。
- 内容。モデルの内容フィールドを指定します。これらのフィールドには、アソシエーション・モデリングで関心の対象となる項目が含まれています。複数のフラグ・フィールド (データがテーブル形式の場合) または単一の名義型フィールド (データがトランザクション形式の場合) を指定できます。

CARMA ノードの「モデル」オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

最小ルール・サポート **(%) :** サポート基準を指定します。ルール・サポート **(**範囲**)** はルール全体を含む学習データ中の ID の割合を参照します。(このサポート範囲の定義は、Apriori ノードで使われている前提条件サポートとは違うことに注意してください。)より一般的なルールに焦点を当てたいときは、設定値を大きくします。

最小ルール確信度 **(%) :** ルールをルールセットに保持する際の確信度基準を指定できます。確信度は、(ルールで予測が行われたすべての ID の中で) 正しい予測が行われた ID の割合を表しています。これは、学習データに基づいて、ルール全体を満たした ID の数を、前提条件を満たした ID の数で割って算出されます。（削除）ルールの数が多すぎる場合や意味のないルールが含まれている場合は、設定値を増やしてください。ルールの数が少なすぎる場合は、設定値を下げます。

注**:** 必要に応じ、値を強調表示して、独自の値を入力できます。確信度を 1.0 未満に下げると、プロセスに多くの空きメモリーが必要になるだけでなく、ルールの作成に極端に長い時間がかかる場合があります。

最大ルール・サイズ **:**ルール内の**アイテム・セット** (**アイテム**ではなく) の最大数を設定します (同じものは 1 つとして数えます)。興味の対象となるルールが比較的短い場合は、設定値を小さくしてルール・セットの作成をスピードアップさせることができます。

注**:** CARMA モデル構築ノードでは、フィールド・タイプがフラグである場合に、モデルの構築時に空のレコードが無視されますが、Apriori モデル構築ノードでは、空のレコードも処理の対象となります。空のレコードとは、モデル構築に使用されるすべてのフィールドの値が false であるレコードです。

CARMA ノードの「エキスパート」オプション

CARMA ノードの操作をよく理解している場合は、次のエキスパート・オプションを使用して、モデル構築処理を調整できます。エキスパート・オプションを利用するには、「エキスパート」タブでモード を「エキスパート」 に設定してください。

複数の結果を持つルールを除外**:** 「双頭」の結果 (2 つの項目を持つ結果) を除外する場合に選択します。例えば、ルール `bread & cheese & fish -> wine&fruit` には双頭の結果、`wine&fruit` が含まれます。デフォルトでは、このようなルールが含まれます。

剪定値の設定 **:** 使用される CARMA アルゴリズムは、処理中に潜在的アイテム・セットのリストからあまり頻繁に出現しないアイテム・セットを定期的に除外 (剪定) し、メモリーを節約します。剪定の頻度を調整するこのオプションを選択すると、指定された値によって剪定の頻度が調整されます。値を小さくすると、アルゴリズムのメモリー必要容量が減少します (一方で、学習時間が長くなる可能性があります)。値を大きくすると学習時間が短くなります (一方で、メモリー必要容量が増加する可能性があります)。デフォルト値は 500 です。

可変サポート **:** 不規則に表示されるため頻繁に出現するように見えるが実はあまり頻繁には出現しないというようなアイテム・セットを削除し、効率を改善する場合に選択します。これは、サポートを高レベルから開始し、徐々に「モデル」タブで指定したレベルまで下げることにより実現できます。推定トランザクション数に値を入力して、サポート・レベルが低下する速度を設定します。

前提条件を持たないルールを許可。結果 (アイテムまたはアイテムのセット) のみを含むルールを許可するときに選択します。これは、共通アイテムまたはアイテムのセットを決定するために調査する場合に役立ちます。例えば、`cannedveg` は、缶詰野菜 の購入がデータ中に一般的に発生することを示す、前提条件のない単一アイテム・ルールです。場合によっては、最も確率の高い予測操作のみに注目する場合、このようなルールを含めることができます。このオプションは、デフォルトでは選択されていません。

アソシエーション・ルールのモデル・ナゲット

アソシエーション・ルールのモデル・ナゲットは、次に示すアソシエーション・ルール・モデル作成ノードの 1 つによって発見されたルールを表します。

- Apriori
- CARMA

モデル・ナゲットには、モデル構築中にデータから抽出されたルールに関する情報が含まれます。

注**:** トランザクション形式データを ID でソートしない場合は、アソシエーション・ルールのナゲット・スコアが不正確になる場合があります。

結果の表示

ダイアログ・ボックスのタブをクリックして、アソシエーション・モデル （Apriori、CARMA） やシーケンス・モデルによって生成されたルールを表示できます。モデル・ナゲットにはルールについての情報が表示され、新しいノード生成やモデルのスコアリング前に、フィルタリングやソートのためのオプションが提供されます。

モデルのスコアリング

調整済みモデル・ナゲット （Apriori、CARMA、Sequence） は、ストリームに追加され、スコアリングに使用されることもあります。詳しくは、トピック 49 ページの『ストリーム内でのモデル・ナゲットの使用』を参照してください。スコアリングに使用したモデル・ナゲットには、追加の設定タブがあり、それぞれのダイアログ・ボックスがあります。詳しくは、トピック 283 ページの『アソシエーション・ルールのモデル・ナゲットの設定』を参照してください。

未調整のモデル・ナゲットは、そのままではスコアリングに使用できません。代わりに、ルールセットを生成し、そのルールセットを使用してスコアリングを行います。 詳しくは、トピック 284 ページの『アソシエーション・モデル・ナゲットからルールセットを生成する』を参照してください。

アソシエーション・ルールのモデル・ナゲットの詳細

生アソシエーション・ルールのモデル・ナゲットの「モデル」タブには、テーブルにアルゴリズムから抽出されたルールが表示されています。テーブル中の各行は、ルールを表しています。最初の列は結論 (ルールの「then」部分) を、その次の列は先行条件 (ルールの「if」部分) を表しています。それに続く列には、確信度、サポート、リフトのようなルール情報が含まれています。

アソシエーション・ルールは、多くの場合、次の表の形式で表示されます。

表 *15.* アソシエーション ルールの例

結果	前提条件
Drug = drugY	Sex = F BP = HIGH

例のルールは、「性別 = *F* で、血圧 = 高なら、薬品は *drugY*」、あるいは「性別 = *F* で、血圧 = 高のレコードについては、薬品は *drugY*」というフレーズに解釈されます。ダイアログ・ボックスのツールバーを使用して、確信度、サポート、インスタンスなどの付加情報を表示できます。

「ソート」メニュー **:** ツールバーの「ソート」メニュー・ボタンで、ルールのソートを制御します。ソート順 (昇順または降順) は、ソート方向ボタン（上向きまたは下向き矢印）を使用して変更できます。

ルールのソートは次によって行います。

- サポート
- 信頼度
- ルール サポート
- 結果
- 評価
- リフト
- デプロイアビリティー

メニューの表示/非表示 : メニューの表示/非表示 （基準項目表示のツールバー・ボタン） は、ルール表示のオプションを制御します。

図 *46.* ボタンの表示/非表示

次の表示オプションを利用できます。

- ルール **ID** は、モデル作成中に割り当てられたルール ID を表示します。ルールID によって、どのルールが提供された予測に適用されているかを特定できます。ルール ID によって、展開性、商品情報、前提条件のような付加ルール情報を、後から結合させることもできます。
- インスタンスは、ルールが適用される特有な ID について、その数に関する情報を表示します。つまり、そのレコードの前提条件は真 (true) です。例えば、`bread -> cheese` というルールが与えられた場合、パン という前提条件を含む学習データ内のレコード数は、インスタンスと呼ばれます。
- サポートは、前提条件サポート、すなわち、学習データを基準にして前提条件が真 (true) である ID の比率を表示します。例えば、学習データの 50% がパンを購入していれば、`bread -> cheese` というルールは、50% の前提条件サポートとなります。注 : ここで定義されたサポートは、インスタンスと同じですが、パーセント表示です。
- 確信度は、前提条件サポートに対するルール・サポートの比率を表示します。これは、指定した前提条件を持つ ID の一部で、結論も真 (true) となるものを示しています。例えば、学習データの 50% がパン (前提条件サポートです) を含むだけでなく、20% がパンとチーズの両方 (ルール・サポートです) も含んでいる場合、`bread -> cheese` というルールの確信度は、`Rule Support / Antecedent Support` で計算され、この場合は、40% となります。
- ルール サポートは、ルール全体、前提条件、結論が真 (true) となる ID の比率を表示します。例えば、学習データの 20% がパンとチーズの両方を含む場合、`bread -> cheese` というルールのルール・サポートは、20% となります。
- 「評価」は、エキスパート・アソシエーション・ルール基準 (確信度との差異、確信度の比、情報の差、またはカイ 2 乗値の正規化) のいずれかを選択した場合に含まれます。これらのエキスパート基準指標は、ユーザーによって設定された「評価測定下限値」の数値と比較されます (また、エキスパート基準ルールが選択された場合にのみ適用されます)。「評価」統計量は、各エキスパート・アソシエーション・ルール基準に対して以下の意味を持ちます。
  - 確信度との差異: 事後確信度 - 事前確信度
  - 確信度の比: (事後確信度 - 事前確信度)/事後確信度
  - 情報の差: 情報利得指標
  - カイ 2 乗値の正規化: カイ 2 乗値の正規化統計量

  これらの統計量はそれぞれ、ユーザーによって設定された「評価測定下限値」の数値と比較され、統計量がこの数値を超過した場合にルールが選択されます。
- リフトは、結果が得られる事前確率に対するルールの確信度の比率を表示します。例えば、全人口の 10% がパンを購入する場合、20% の確信度でパンを購入するかどうかを予測するルールは、20/10 = 2 のリフトを持つことになります。11% の確信度でパンを購入する場合であれば、リフトは 1 に近くなります。このことは、前提条件を持つことで、結論が得られる確率に大きな違いが生じないということを意味します。一般に、ルールのリフトが 1 に近い場合より、1 から離れた場合の方が、より興味深い結果が得られます。
- デプロイアビリティーは、前提条件を満足しつつ結論を満足しない学習データの割合を示す尺度です。製品購入については、全顧客ベースで見て、前提条件を所有し （あるいはすでに購入し） かつ未だ結

論を購入していない人の割合を基本的に意味しています。展開性の統計は ((Antecedent Support in # of Records - Rule Support in # of Records) / Number of Records) * 100 として定義されます。前提条件のサポート は前提条件が true であるレコードの数、ルール・サポート は前提条件と結果がいずれも true であるレコードの数を意味します。

「フィルター」ボタン : メニューにある「フィルター」ボタン (じょうごの形をしたアイコン) を押すと、ダイアログ・ボックスのボタンが展開され、パネルには有効なルール・フィルターが表示されます。フィルターは、「モデル」タブに表示されるルール数を減らすために使用します。

図 *47.* 「フィルター」ボタン

フィルターを作成するには、展開したパネルの右に表示される「フィルター」アイコンをクリックします。クリックすると、ルールの表示に関する制約を指定するための別個のダイアログ・ボックスが開きます。注意を要するのは、「フィルター」ボタンが「ノードの生成」メニューとともによく使用されることで、これにより、最初にルールにフィルターをかけ、次にルールのサブセットを含むモデルを生成します。詳しくは、以下の 『ルールに適用するフィルターを指定する』を参照してください。

「ルールを検索」ボタン :「ルールを検索」ボタン (双眼鏡アイコン) を使用して、指定されたルール ID に表示されるルールを検索できます。隣接するディスプレイ ボックスは、有効数のうち、現在表示されているルールの数を表示しています。モデルによって割り当てられるルール ID は、その時点で、見つかった順番となり、スコアリング中にデータにデータに追加されます。

図 *48.* 「ルールを検索」ボタン

ルール ID を並べ替えるためには次のようにします。

1. IBM SPSS Modeler を使用してルール ID を並べ替えるには、まず、確信度やリフトといった希望する基準に従って、ルール表示テーブルをソートします。
2. 次に、「ノードの生成」メニューのオプションを使用して、フィルタリングされたモデルを生成します。
3. フィルタリングされたモデルのダイアログ・ボックスで、「開始番号を指定してルールに連続した番号を付ける」を選択し、開始番号を指定します。

詳しくは、 285 ページの『フィルタリングされたモデルの生成』を参照してください。

ルールに適用するフィルターを指定する

デフォルトでは、Apriori、CARMA、Sequence といったルール・アルゴリズムによって、大量で煩雑なルールが生成されます。ルールのスコアリングを見るときの明確性を改善したり、ルールのスコアリングを効率化するためには、フィルタリング・ルールを工夫し、結果や着目した前提条件がより強調されて表示されるようにする必要があります。ルール・ブラウザーの「モデル」タブにあるフィルタリング設定を使い、フィルター適性を指定するためのダイアログ・ボックスを開くことができます。

結果: 「フィルターを有効化」 を選択して、特定の結果を含めたり除外したりして決定した、フィルタリング ルールのオプションを有効にします。「いずれかを含む」 を選択し、指定した結果の少なくとも 1 つがルールに含まれるフィルターを作成します。または、「除外」 を選択し、指定した結果を除外するフ

ィルターを作成します。リスト・ボックスの右にあるピッカー・アイコンを使用して、結果を選択できます。そうすると、生成されたルールに含まれる全ての結果が一覧表示されたダイアログ・ボックスが開きます。

注：結果は、複数のアイテムを含んでいることがあります。フィルターは、指定したアイテムのいずれかが、結果に含まれているかだけを確認します。

前提条件**:**「フィルターを有効化」 を選択して、指定された前提条件の包含や除外によるフィルタリング・ルールのオプションを有効にします。リスト・ボックスの右にあるピッカー・アイコンを使用して、アイテムを選択できます。ダイアログ・ボックスが開き、生成されたルールに含まれるすべての前提条件が一覧表示されます。

- 「すべてを含む」 を選択し、指定した前提条件のすべてがルールに含まれる中間所得者対象のフィルターとして設定します。
- 「いずれかを含む」 を選択し、指定した前提条件の少なくとも 1 つを含むルールのためのフィルターを作成します。
- 「除外」 を選択し、指定した前提条件を含むルールが除外されるフィルターを作成します。

確信度**:**「フィルターを有効化」 を選択し、ルールの確信度レベルに基づくフィルタリング ルールのオプションを有効にします。「最小値」 と 「最大値」 を設定し、確信度の範囲を特定できます。生成されたモデルを参照する場合、確信度がパーセント表示されます。出力をスコアリングする場合、確信度は 0 と 1 の間の数字で表現されます。

前提条件サポート**:**「フィルターを有効化」 を選択し、ルールの前提条件サポートのレベルに基づくフィルタリング ルールのオプションを有効にします。前提条件サポートは、同じ前提条件を現在選択されているルールとして含む学習データの比率を表示します。その比率は、人気の指標と似ています。「最小値」 と 「最大値」 を設定し、サポートのレベルに基づくルールをフィルタリングするための範囲を指定できます。

リフト**:**「フィルターを有効化」 を選択し、ルールのリフト測定に基づくフィルタリング ルールのオプションを有効にします。注：リフトのフィルタリングは、本製品のリリース 8.5 以降で生成したアソシエーション・モデルまたはそれ以前のモデルでリフト測定を含むものに対してのみ有効です。シーケンス・モデルでは、このオプションは利用できません。

「**OK**」 をクリックして、このダイアログ・ボックスで有効にしたすべてのフィルターを適用します。

ルールのグラフを生成する

アソシエーション・ノードは多くの情報を提供します。ただし、その情報はビジネス・ユーザーが容易にアクセスできる形式であるとは限りません。ビジネス・レポート、プレゼンテーションなどに用意に組み込むことができる方法でデータを提供するために、選択したデータのグラフを作成できます。「モデル」タブから、選択したルールのグラフを生成できるため、そのルールのケースのグラフのみ作成します。

1. 「モデル」タブで、関心のあるルールを選択します。
2. 「生成」メニューの 「グラフ **(**選択項目から**)**」 を選択します。グラフボードの「基本」タブが表示されます。

   注：この方法でグラフボードを表示した場合、「基本」タブと「詳細」タブのみを使用できます。
3. 「基本」タブまたは「詳細」タブいずれかの設定を使用し、グラフに表示する詳細を指定します。
4. 「OK」をクリックしてグラフを生成します。

グラフの見出しは選択されたルールおよび前提条件の詳細を識別します。

アソシエーション・ルールのモデル・ナゲットの設定

この「設定」タブを使用して、アソシエーション・モデル （Apriori、CARMA） のスコアリング・オプションを設定します。このタブが利用可能になるのは、モデル・ナゲットがスコアリングを目的としてストリームに追加された後です。

注: 未精製モデルをスコアリングすることはできないため、未精製モデルを参照するためのダイアログ・ボックスには「設定」タブがありません。「未精製」モデルをスコアリングするには、最初にルール・セットを生成する必要があります。詳しくは、トピック 284 ページの『アソシエーション・モデル・ナゲットからルールセットを生成する』を参照してください。

最大予測数: バスケット・アイテムの各セットに含める最大予測数を指定します。このオプションは、以下に示すルール基準とともに使用され、「最上位」の予測を行います。ここで、最上位 というのは、以下で設定される確信度、サポート、リフトなどについて、最も高いレベルであることを示しています。

ルール基準: ルールの強さを決定するために使用される尺度を選択します。アイテムセットに最上位の予測を返すために、ここで選択した基準の強さによって、ルールがソートされます。使用可能な基準を以下のリストに示します。

- 確信度
- サポート
- ルール・サポート（サポート * 確信度）
- リフト
- デプロイアビリティー

予測の繰り返しを許可: スコアリング時に同じ結果の複数のルールを処理対象とする場合に選択します。例えば、このオプションを選択すると次のルールのスコアリングができるようになります。

```
bread & cheese -> wine
cheese & fruit -> wine
```

スコアリング時に予測の繰り返しを除外するには、このオプションをオフにします。

注: 複数の結果を持つルール (パン & チーズ & フルーツ -> ワイン & パテ) は、すべての結果 (ワイン & パテ) があらかじめ予測されている場合のみ、反復の予測と見なされます。

一致しないバスケット アイテムを無視: アイテム・セット内の追加アイテムの存在を無視する場合に選択します。例えば、[tent & sleeping bag & kettle] を含むバスケットにこのオプションが選択された場合、 バスケットに追加のアイテム (kettle) がある場合でもルール tent & sleeping bag -> gas_stove が適用されます。

状況によっては、余計なアイテムを除外する方が良いこともあります。例えば、テント、寝袋、やかんを購入した人が既にガスストーブを所有していることも考えられますが、やかんの存在で表記されます。つまり、ガスストーブは最良の予測ではありません。このような場合、「一致しないバスケット アイテムを無視」 の選択を解除して、ルールの前提条件がバスケットの中身と完全に一致するようにします。デフォルトでは、一致しないアイテムは無視されます。

予測がバスケットにないことを検査: 結果がバスケットの中に入っていないことを確認します。例えば、スコアリングの目的が、家庭で使う家具製品を推奨することであれば、ダイニング・テーブルを既に含むバスケットが別のものを購入するケースはほとんどありません。このような場合、このオプションを使用してください。一方、製品が腐りやすかったり、使い捨てのものである場合 （チーズ、粉ミルク、ティッシュペ

ーパーなど）、バスケットにすでに結論が入っているルールは、価値があります。後者の場合、最も便利なオプションは下にある 「バスケットに予測があるかどうかを検査しない」 です。

予測がバスケットにあることを検査**:** 結果もバスケットの中に入っていることを確認する場合に、このオプションを選択します。このアプローチは、既存の顧客やトランザクションに対する洞察を得ようとする場合に役立ちます。例えば、最上位のリフトを持つルールを識別したり、さらにどの顧客がそのルールに適合するのかを調べたい場合があります。

バスケットに予測があるかどうかを検査しない**:** バスケットの中に結果があってもなくても、スコアリング時にすべてのルールを含める場合に、このオプションを選択します。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

アソシエーション・ルールのモデル・ナゲットの要約

アソシエーション・ルールのモデル・ナゲットの「要約」タブには、発見ルール数、およびルールの範囲、リフト、確信度および展開性の最小値と最大値が表示されます。

アソシエーション・モデル・ナゲットからルールセットを生成する

Apriori や CARMA などのアソシエーション・モデル・ナゲットは、データを直接スコアリングするために使われます。また、ルール セットと呼ばれるルールのサブセットを最初に生成する方法もあります。ルール・セットは、スコアリングで直接使用できない未精製モデルを処理する場合に特に便利です。詳しくは、トピック 53 ページの『未精製モデル』を参照してください。

ルール・セットを生成するには、モデル・ナゲット・ブラウザーにある「生成」メニューから 「ルール セット」 を選択します。ルールをルール・セットに変換する場合は、次のオプションを指定できます。

ルール セット名**:** 新規に生成されるルール・セット・ノードの名前を指定できます。

ノードの生成先**:** 新しく生成されるルール・セット・ノードの場所を制御します。「キャンバス」、「**GM** パレット」、または 「両方」 を選択します。

対象フィールド**:** 生成されたルール・セット・ノードで使用される出力フィールドを決定します。リストから 1 つの出力フィールドを選択します。

最低のサポート**:** 生成されたルール・セット内で保持されるルールの最低のサポートを指定します。指定した値より小さい範囲を持つルールは新規ルール・セット内に表示されません。

最低確信度**:** 生成されたルール・セット内で保持されるルールの最低確信度を指定します。指定した値より小さい確信度を持つルールは新規ルール・セット内に表示されません。

デフォルト値**:** 該当するルールがない、得点計算されたレコードに割り当てられる、対象フィールドのデフォルト値を指定できます。

フィルタリングされたモデルの生成

Apriori、CARMA、Sequence ルール・セット・ノードなどのアソシエーション・モデル・ナゲットから、フィルタリングされたモデルを生成するには、モデル・ナゲット・ブラウザーにある「ノードの生成」メニューから、「除外されたモデル」を選択します。これにより、ブラウザーに現在表示されているルールだけを含むサブセット・モデルが生成されます。注 : 未調整モデルのフィルター処理されたモデルは生成できません。

フィルタリングのためのルールとして次のようなオプションを指定できます。

新規モデルの名前**:** 新規に生成されるフィルタリングされたモデル・ノードの名前を指定できます。

ノードの生成先**:** 新しく生成されるフィルタリングされたモデル・ノードの場所を制御します。「キャンバス」、「**GM** パレット」、または 「両方」 を選択します。

ルールの番号付け**:** フィルタリングされたモデルに含まれるルールのサブセットにおいて、ルール ID にどのように番号を付けるかを指定します。

- 元のルール **ID** 番号を保持**:** ルールに対する元のナンバリングを維持する場合に選択します。デフォルトでは、アルゴリズムによって検出された順番に対応する ID がルールに割り振られます。その順番は、採用されたアルゴリズムによって変わります。
- 開始番号を指定してルールに連続した番号を付ける**:** フィルタリングされたルールのための新しいルール ID を付けるときに選択します。「モデル」タブのルール・ブラウザー・テーブルに表示されたソート順に基づいて、ここで指定した番号から始まる新しい ID が割り振られます。右の矢印を使用して、ID の最初の番号を指定できます。

スコアリング・アソシエーション・ルール

新しいデータをアソシエーション・ルールのモデル・ナゲットに流して生成されたスコアは、個別のフィールドに返ります。予測を表す *P*、確信を表す *C* そしてルール ID を表す *I* と、各予測に対して 3 つの新しいフィールドが追加されます。これら出力フィールドの構成は入力データがトランザクション形式かテーブル形式かによって異なります。これらの形式の概要は、 272 ページの『テーブル形式データとトランザクション形式・データ』を参照してください。

例えば、次の 3 つのルールに基づいて予測を生成するモデルを使用して、バスケット・データをスコアリングするとします。

```
Rule_15  bread&wine -> meat (confidence 54%)
Rule_22  cheese -> fruit (confidence 43%)
Rule_5  bread&cheese -> frozveg (confidence 24%)
```

テーブル形式データ **:** テーブル形式データの場合、3 つの予測 （デフォルトでは 3 つ） は、1 つのレコードで返されます。

表 16. テーブル形式のスコア

ID	パン	ワイン	チーズ	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	肉	0.54	15	フルーツ	0.43	22	冷凍野菜	.24	5

トランザクション形式のデータ **:** トランザクション形式のデータには、各予測について個別のレコードが生成されます。予測はそのまま個別の列に追加されますが、スコアは計算結果のまま返されます。このことは、次の出力例に示すように、レコードが不完全な予測を伴うことを意味します。2 つ目と 3 つ目の予測（P2 と P3）は、最初のレコードで、関連付けられた確信度とルール ID を伴って空白になっています。しかし、スコアが返される場合、最後のレコードは 3 つの予測をすべて含んでいます。

表 17. トランザクション形式内のスコア

ID	項目	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	bread (パン)	肉	0.54	14	$null$	$null$	$null$	$null$	$null$	$null$
Fred	チーズ	肉	0.54	14	フルーツ	0.43	22	$null$	$null$	$null$
Fred	wine (ワイン)	肉	0.54	14	フルーツ	0.43	22	冷凍野菜	0.24	5

報告や展開のために完全な予測だけが必要な場合、条件抽出ノードを使用して完全なレコードを選びます。

注 : これらの例で使ったフィールドの名前は、わかりやすくするために省略されています。実際に使用する場合は、アソシエーション・モデルの結果フィールドは次の表に示すように名前が付けられます。

表 18. アソシエーション・モデルの結果フィールドの名前

新規フィールド	フィールドの名前の例
予測	*$A-TRANSACTION_NUMBER-1*
確信度 （あるいは他の基準）	*$AC-TRANSACTION_NUMBER-1*
ルール ID	*$A-Rule_ID-1*

複数の結論がある場合のルール

CARMA アルゴリズムの場合は、複数の結論を持つルールがあっても構いません。例えば、

```
bread -> wine&cheese
```

こうした「双頭」ルールをスコアリングすると、次のテーブルに示す形で予測が返されます。

表 19. 複数の結論を持つ予測を含む結果のスコアリング

ID	パン	ワイン	チーズ	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	肉&野菜	0.54	16	フルーツ	0.43	22	冷凍野菜	.24	5

展開の前にそのようなスコアを分割する必要があることもあります。複数の結論を持つ予測を分割するには、CLEM ストリング関数を使用してフィールドを解析する必要があります。

アソシエーション・モデルを展開する

アソシエーション・モデルをスコアリングする場合、予測と確信度は個別の列に出力されます（ *P* は予測、*C* は確信度、*I* はルール ID を表しています）。これは、入力データはテーブル形式であるのかトランザクション形式であるのか、という場合にあたります。 詳しくは、トピック 285 ページの『スコアリング・アソシエーション・ルール』を参照してください。

展開用にスコアを準備する際、使用するアプリケーションによっては、列ではなく行の予測を伴った形式に出力データを移行する必要があるかもしれません（行あたり 1 つの予測で、これは「ペーパー・ロール」形式としても知られます）。

テーブル形式のスコアの行列入れ替え

以下の手順で示すように、IBM SPSS Modeler のステップの組み合わせを使用して、テーブル形式のスコアを列から行に移行できます。

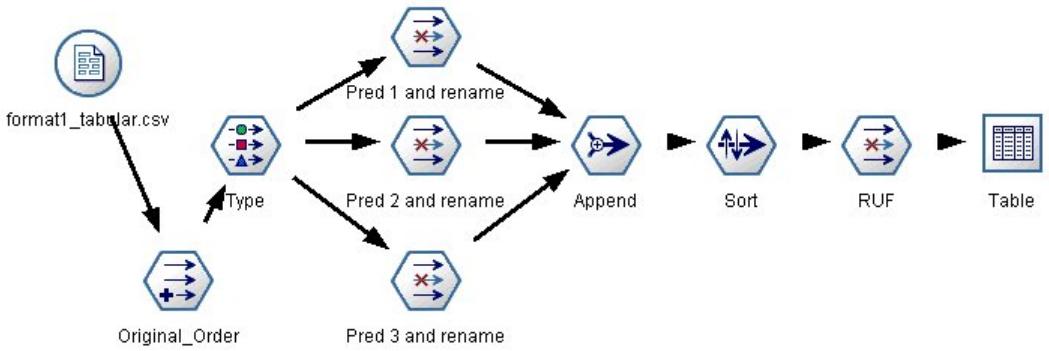


図 *49.* テーブル形式のデータをペーパー・ロール形式に移行させるためのストリームの例

1. フィールド作成ノードにある `@INDEX` 関数を使い、予測の現在の順番を確認し、この指標を新しいフィールド、例えば *Original_order* に保存します。
2. データ型ノードを追加して、すべてのフィールドがインスタンス化されていることを確認します。
3. デフォルトの予測、確信度、および ID フィールド (*P1*、*C1*、*I1*) の名前を *Pred*、*Crit*、*Rule_ID* といった共通フィールド変更するために、フィルター・ノードを使います。これらは後でレコードの追加のために使います。生成した各予測に対して、1 つのフィルター・ノードが必要です。

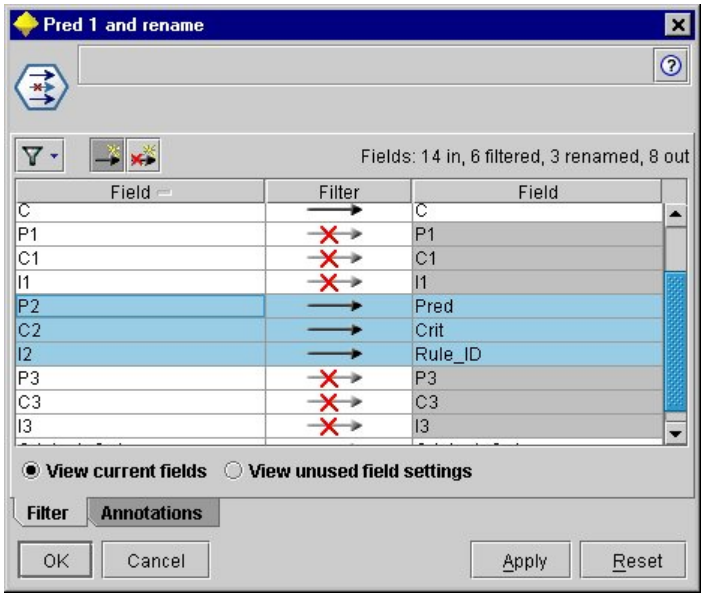


図 *50.* 予測 *2* のフィールドの名前を変更しながら、予測 *1* と *3* のフィールドのフィルタリングを行う。

4. レコード追加ノードを使用して、共有している *Pred*、*Crit*、*Rule_ID* の値を追加する。
5. *Original_order* フィールドについては昇順で、*Crit* フィールドについては降順でレコードをソートするために、ソート・ノードを接続します。Crit フィールドは、確信度、リフト、サポートといった基準による予測のソートに使用されるフィールドです。
6. 別のフィルター・ノードを使用して、*Original_order* フィールドを出力からフィルタリングします。

この時点では、データ展開のための準備はできています。

トランザクション形式スコアの移行

このプロセスは、トランザクション形式スコアの移行に似ています。例えば、次に示したストリームでは、展開に使えるように、スコアは各行に単一の予測を伴った形式に移行されます。

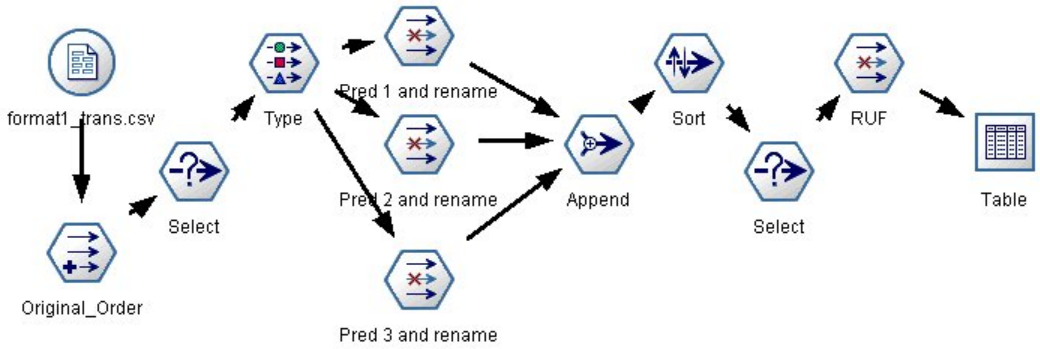


図 *51.* トランザクション・データをペーパー・ロール形式に移行させるためのストリームの例

2 つの条件抽出ノードを追加すると、そのプロセスは上で説明したテーブル形式のためのものと全く同じです。

- 最初の条件抽出ノードを使用して、ルール ID を隣接するレコードと比較し、特有なレコードまたは未定義のレコードだけをインクルードします。この条件抽出ノードは、CLEM 式を使用してレコードを選択します。`ID /= @OFFSET(ID,-1) or @OFFSET(ID,-1) = undef`

- 2 番目の条件抽出ノードを使用して、無関係なルールや、Rule_ID がヌル値のルールを破棄します。この条件抽出ノードは、次の CLEM 式を使用してレコードを破棄します。not(@NULL(Rule_ID))

展開のためのスコア移行についての詳細は、 テクニカル・サポートまでお問い合わせください。

シーケンス・ノード

シーケンス・ノードは、**パン > チーズ** 形式の中に、シーケンシャルな、または時間経過が伴うデータ中のパターンを検出します。シーケンスの要素はアイテム・セットと呼ばれます。これは、1 つのトランザクションを構成します。例えばある人が店でパンと牛乳を購入し、数日後に同じ店でチーズを購入した場合、この人の購買活動は 2 つのアイテム・セットで表すことができます。パンと牛乳を含んだセットと、チーズを含んだセットです。予測可能な順序で起こる傾向にある項目セットのリストを、シーケンスと呼びます。シーケンス・ノードでは、頻繁に生じるシーケンスが検出され、予測を行うための生成されたモデル・ノードが作成されます。

要件: シーケンス・ルール・セットを作成するには、ID フィールドを指定する必要があります。必要に応じて、時間フィールドと 1 つ以上の内容フィールドを指定することができます。これらの設定は、モデル作成ノードの「フィールド」タブで行わないと、上流のデータ型ノードから読むことができないことに注意してください。ID フィールドの役割や測定の尺度はどれでもかまいません。時間フィールドを指定する場合、役割はどれでもかまいませんが、ストレージは数値、日付、時間、またはタイムスタンプでなければなりません。時間フィールドを指定しなかった場合、シーケンス・ノードでは暗示的にタイム・スタンプが使用されます。実際には、行番号が時間値となります。内容フィールドには測定の尺度でも役割でもかまいませんが、すべての内容フィールドは同じ種類でなければなりません。数値の場合は、整数の範囲でなければなりません (実数ではない)。

利点: シーケンス・ノードは CARMA アソシエーション・ルール・アルゴリズムに基づいており、効率的な 2 段階通過法を使用してシーケンスを検出します。さらに、シーケンス・ノードで作成される生成されたモデル・ノードは、データ・ストリームに挿入して予測を行うことができます。生成されたモデル・ノードでは、特定シーケンスの検出とカウント、および特定シーケンスをもとにした予測を行うためのスーパーノードも作成できます。

シーケンス・ノードの「フィールド」オプション

シーケンス・ノードを実行する前に、シーケンス・ノードの「フィールド」タブで、ID と内容フィールドを指定する必要があります。時間フィールドを使用したい場合は、それもここで指定する必要があります。

ID フィールド: リストから ID フィールドを選択します。ID フィールドとして使用できるのは、数値またはシンボル値のフィールドです。選択したフィールドでは、一意の値がそれぞれ、ある分析ユニットを示している必要があります。例えば、マーケット・バスケット分析なら、各 ID が 1 人の顧客を表します。Web ログ分析なら、各 ID が 1 台のコンピューター (IP アドレス) あるいは 1 人のユーザー (ログイン・データ) を表します。

- **連続する ID :** データ・ストリーム中で同じ ID を持つすべてのレコードが一緒に表示されるようにデータをソートしている場合、このオプションを選択すると処理を高速化することができます。データがあらかじめソートされていない場合 (またはわからない場合) は、このオプションは選択しないでください。この場合、シーケンス・ノードが自動的にデータをソートします。

注 : データがソートされていないのにこのオプションを選択すると、シーケンス・モデルで不正な結果しか得られません。

時間フィールド **:** データ中のフィールドを使用してイベント時間を示す場合、「時間フィールドを使用」を選択して、使用するフィールドを指定します。時間フィールドは、数値、日付、時間、またはタイムスタンプでなければなりません。時間フィールドを指定しなかった場合、データ・ソースから順番にレコードが取得されたものとみなされ、レコード番号が時間値として使用されます (第 1 レコードの時間が "1"、第 2 レコードの時間が"2" など)。

内容フィールド**:** モデルの内容フィールドを指定します。これらのフィールドには、シーケンス・モデル作成の対象となるイベントが含まれています。

シーケンス・ノードで扱えるデータは、テーブル形式またはトランザクション形式のいずれかの形式です。トランザクション形式のデータで複数のフィールドを使用している場合、あるレコードのフィールドで言及されているアイテムはすべて、単一のタイム・スタンプを使った単一のトランザクションで検出されたものとみなされます。 詳しくは、トピック 272 ページの『テーブル形式データとトランザクション形式・データ』を参照してください。

データ区分**:** このフィールドでは、モデル構築の学習、テスト、および検証の各ステージ用に、データを独立したサブセット (サンプル) に分割するフィールドを指定できます。1 組のサンプルをモデルの生成に使用し、別のサンプルで生成したモデルをテストすることにより、そのモデルが、このデータに似た性質を持つより大きなデータセットにどの程度適用できるかについての良い目安を得ることができます。データ型ノードまたはデータ区分ノードを使用して複数のデータ区分フィールドが定義されている場合、データ区分を使用する各モデル作成ノードの「フィールド」タブで単一のデータ区分フィールドを選択する必要があります (1 つのデータ区分だけが存在している場合、データ分割を有効にすると、そのデータ区分が必ず自動的に使用されます)。また、選択したデータ区分を分析に適用するには、そのノードの「モデル・オプション」タブでデータ区分が有効になっている必要があります (このオプションの選択を解除すると、フィールド設定を変更することなくデータ区分を無効にすることができます)。

シーケンス・ノードの「モデル」オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

最小ルール サポート **(%)** サポート基準を指定できます。ルール・サポートはシーケンス全体を含む学習データ中の ID の割合を参照します。より一般的なシーケンスに焦点を当てたいときは、設定値を大きくします。

最小ルール確信度 **(%)** シーケンスをシーケンス セット内に保持する確信度基準を指定できます。確信度は、ルールで予測が行われたすべての ID の中で、正しい予測が行われた ID の割合を表しています。これは、学習データに基づいて、見つかったシーケンス全体を前提条件が見つかった ID 数で除算した ID 数として算出されます。指定した基準より確信度の低いシーケンスが破棄されます。シーケンスの数が多すぎる場合や意味のないシーケンスが含まれている場合は、この設定値を上げてみてください。シーケンスの数が少なすぎる場合は、設定値を下げます。

注**:** 必要に応じ、値を強調表示して、独自の値を入力できます。確信度を 1.0 未満に下げると、プロセスに多くの空きメモリーが必要になるだけでなく、ルールの作成に極端に長い時間がかかる場合があります。

最大シーケンス サイズ シーケンス内の異なる項目の最大数を設定できます。興味の対象となるシーケンスが比較的短い場合は、設定値を小さくしてシーケンス・セットの作成をスピードアップさせることができます。

ストリームに追加する予測 結果としてできる生成されたモデル ノードによって、予測をいくつストリームに追加するかを指定します。詳しくは、 292 ページの『シーケンス・モデル・ナゲット』を参照してください。

シーケンス・ノードの「エキスパート」オプション

シーケンス・ノードの操作をよく理解している場合は、次のエキスパート・オプションを使用して、モデリング処理を調整できます。エキスパート・オプションを利用するには、「エキスパート」タブで「モード」に 「エキスパート」 を設定してください。

最大長の設定 : このオプションを選択すると、長さ (第 1 アイテム・セットから最終アイテム・セットまでの時間) が指定された値以下であるようなシーケンスだけがレポートされます。時間フィールドを指定していない場合、長さは生データ内の行 (レコード) で表されます。使われている「時間」フィールドが、時間、日付、またはタイムスタンプ・フィールドの場合、長さは秒で表されます。数値型フィールドの場合、長さはそのフィールドと同じ単位で表されます。

剪定値の設定 : シーケンス・ノードで使用される CARMA アルゴリズムは、処理中に潜在的アイテム・セットのリストからあまり頻繁に出現しないアイテム・セットを定期的に削除 (剪定) し、メモリーを節約します。剪定の頻度を調整するには、このオプションを選択します。指定された値によって、剪定の頻度が判断されます。値を小さくすると、アルゴリズムのメモリー必要容量が減少します (一方で、学習時間が長くなる可能性があります)。値を大きくすると学習時間が短くなります (一方で、メモリー必要容量が増加する可能性があります)。

メモリー中の最大シーケンス数の設定 : このオプションを選択すると、モデリングの際、CARMA アルゴリズムによって、メモリーに保存する候補シーケンスの数が指定された数に制限されます。シーケンス・モデルを作成するのに IBM SPSS Modeler が多くのメモリーが使用している場合、このオプションを選択します。ここで指定する最大シーケンス数は、モデルが作成される際に内部で追跡される候補シーケンスの数のことです。最終モデルで予測されるシーケンスの数よりもずっと大きい数である必要があります。

アイテム・セット間の隔たりを制限 : アイテム・セット間の時間の隔たりを制限することができます。このオプションを選択すると、最小の隔たりの指定値より隔たりの小さいアイテム・セット、および最大の隔たりの指定値より隔たりの大きいアイテム・セットはシーケンスに含められなくなります。これを利用して、長い間隔を含んだシーケンスや短時間で終わってしまうシーケンスを除外することができます。

注 : 使われている「時間」フィールドが、時間、日付、またはタイムスタンプ・フィールドの場合、時間間隔は秒で表されます。数値型フィールドの場合、時間間隔は時間フィールドと同じ単位で表されます。

例えば、次のようなトランザクションのリストについて考えてみます。

表 *20.* トランザクションのサンプル・リスト

ID	時間	内容
1001	1	りんご
1001	2	bread (パン)
1001	5	チーズ
1001	6	ドレッシング

これらのデータをもとに、最小の隔たりを 2 に設定してモデルを作成すると、以下のようなシーケンスができます。

```
apples -> cheese

apples -> dressing

bread -> cheese

bread -> dressing
```

apples と breadの隔たりは、最小の隔たりより小さいため、apples -> bread というシーケンスはレポートされません。同様に、次の代替データについて考えてみます。

表 21. トランザクションのサンプル・リスト

ID	時間	内容
1001	1	りんご
1001	2	bread (パン)
1001	5	チーズ
1001	20	ドレッシング

最大の隔たりが 10 に設定されている場合、**ドレッシング**を含んだシーケンスはレポートされません。なぜなら、**チーズ**と**ドレッシング**の隔たりが大きすぎて、それらを同じシーケンスの一部とみなすことができないからです。

シーケンス・モデル・ナゲット

シーケンス・モデル・ナゲットは、シーケンス・ノードによって検出された特定の出力フィールドで見つかったシーケンスを表し、予測生成のためにストリームに追加されます。

シーケンス・ノードを含むストリームを実行すると、シーケンス・ノードによって、予測と各予測に関連付けられた確信度の値から成る 1 対のフィールドが、シーケンス・モデルからデータに追加されます。デフォルトでは、上位 3 つの予測からなる 3 対のフィールド (および対応する確信度値) が追加されます。モデル・ナゲットをストリームに追加した後の「設定」タブ上だけでなく、モデル作成時にシーケンス・ノードのモデル・オプションを設定することによっても、生成される予測の数を変更できます。 詳しくは、トピック 295 ページの『シーケンス・モデル・ナゲットの設定』を参照してください。

新規フィールド名はモデル名から派生します。予測フィールドの名前は *$S-sequence-n* (*n* は *n* 番目の予測を示す)、確信度フィールドの名前は*$SC-sequence-n* です。連続する複数のシーケンス・ルール・ノードを含むストリームでは、新しいフィールド名の接頭辞にはそれぞれのノードを区別するための数字が含まれます。ストリーム内の最初のシーケンス・セット・ノードは通常の名前を使用します。2 番目のノードは *$S1-* と *$SC1-* で始まる名前、3 番目のノードは *$S2-* と *$SC2-* で始まる名前というように名前が付けられていきます。予測は、確信度の順に表示されます。したがって、*$S-sequence-1* には最も確信度の高い予測が含まれ、*$S-sequence-2* には次に確信度の高い予測が含まれます。使用可能な予測数が要求された予測数より少ないレコードの場合、残りの予測には値 $null$ が含まれます。例えば、特定のレコードに対して 2 つの予想しか行えなかった場合、*$S-sequence-3* および *$SC-sequence-3* の値は $null$ となります。

各レコードに対して、これまでに現在の ID に対して処理された一連のトランザクションとモデル中のルールが比較されます (現在のレコード、および同じ ID とより以前のタイムスタンプを持つ前のレコードを含む)。この一連のトランザクションに適用される、確信度の値が最も高い *k* ルールを使用して、レコードの *k* 予測が生成されます。*k* は、モデルをストリームに追加した後に「設定」タブで指定された予測の数

です (複数のルールがトランザクション・セットに対して同じ結果を予測した場合、もっとも確信度が高いルールだけが使用されます)。詳しくは、トピック 295 ページの『シーケンス・モデル・ナゲットの設定』を参照してください。

アソシエーション・ルール・モデルのほかのタイプと同様に、データの形式は、シーケンス・モデルの作成時に使用した形式と一致している必要があります。例えば、テーブル形式データを使用して作成されたモデルは、テーブル形式データのスコアリングだけに使用できます。 詳しくは、トピック 285 ページの『スコアリング・アソシエーション・ルール』を参照してください。

注 : ストリームで生成されたシーケンス・セット・ノードを使用してデータの得点計算を行う場合、計算では、モデルの作成で選択した許容度と隔たりの設定は無視されます。

シーケンス・ルールからの予測

ノードはレコードを時間に依存する方法で処理します (モデル構築にタイムスタンプ・フィールドが使われていない場合は、順序に依存する方法で)。レコードは、ID フィールドとタイムスタンプ・フィールド (ある場合) でソートされていなければなりません。しかし、予測はこれらの追加先レコードのタイム・スタンプには関連付けられません。これらは現在の ID に対して、現在のレコードまでのトランザクション履歴に基づいて、将来のある時点で 最も可能性の高いアイテムを表しているだけです。

各レコードに対する予測が必ずしもレコードのトランザクションに依存するわけではないことに注意してください。現在のレコードのトランザクションが特定のルールの要因とならない場合、ルールは現在の ID の以前のトランザクションに基づいて選択されます。つまり、現在のレコードにより有益な予測情報がシーケンスに追加されない場合は、この ID に対する前回の有益な予測が現在のレコードで使用されます。

例えば、あるシーケンス・モデルに次のルールがある場合に、

```
Jam -> Bread (0.66)
```

これに次のレコードを渡す場合を考えてみましょう。

表 22. 例のレコード

ID	購入品	予測
001	ジャム	bread (パン)
001	牛乳	bread (パン)

最初のレコードは、期待通りに予測「パン」を生成しています。2 番目のレコードの場合、ジャム の後にミルク が続くルールはないため、ミルク のトランザクションにより有益な情報は追加されません。そのため、ルール Jam -> Bread のルールが引き続き適用され、予測はパン になります。

ノードの生成

「ノードの生成」メニューでは、シーケンス・モデルに基づいて新しいスーパーノードを作成することができます。

- ルール スーパーノード: 得点計算されたデータ中のシーケンスの出現回数を検出、カウントできる、スーパーノードを作成します。ルールが選択されていない場合、このオプションは無効になります。 詳しくは、トピック 296 ページの『シーケンス・モデル・ナゲットからルール・スーパーノードを作成』を参照してください。
- パレットのモデル: モデルをモデル・パレットに戻します。同僚が、モデル自体ではなくモデルを含むストリームを送信した場合に役立ちます。

シーケンス・モデル・ナゲットの詳細

シーケンス・モデル・ナゲットの「モデル」タブでは、アルゴリズムによって抽出されたルールが表示されます。テーブルの各行は、最初の列の前提条件（ルールの「if」部分）と、2 番目の列の結果（ルールの「then」部分）を伴って、1 つのルールを表します。

各ルールは次の書式で表示されます。

表 *23.* ルールの書式

前提条件	結果
ビールと缶詰野菜	ビール
魚 魚	魚

最初の例のルールでは、同じトランザクションに「ビール」と「缶詰野菜」がある *ID* には、それ以降、「ビール」が出現する可能性が高い と解釈されます。2 番目の例のルールでは、あるトランザクションに「魚」があり、別のトランザクションにも「魚」がある *ID* には、それ以降、「魚」が出現する可能性が高い と解釈できます。最初のルールでは、ビール と缶詰野菜 が同時に購買され、2 番目のルールでは、魚 が 2 つの個別のトランザクションで購買されたことに注目してください。

「ソート」メニュー **:** ツールバーの「ソート」メニュー・ボタンで、ルールのソートを制御します。ソート順 (昇順または降順) は、ソート方向ボタン（上向きまたは下向き矢印）を使用して変更できます。

ルールのソートは次によって行います。

- サポート %
- 確信度 %
- ルール・サポート %
- 結果
- 最初の前提条件
- 最後の前提条件
- アイテム数（前提条件）

例えば、次のテーブルはアイテム数の降順にソートされます。前提条件セットで多数項目を持つルールは、より少ない項目を持つルールに優先します。

表 *24.* アイテム数をソート項目にするルール

前提条件	結果
ビールと缶詰野菜と冷凍食品	冷凍食品
ビールと缶詰野菜	ビール
魚 魚	魚
清涼飲料	清涼飲料

基準項目の表示/非表示**:** 基準項目の表示/非表示 （グリッド・アイコン） は、ルール表示のオプションを制御します。次の表示オプションを利用できます。

- インスタンスは、フル・シーケンス で前提条件と結果の両方が表示される特有な ID について、その数に関する情報を表示します。このことは、前提条件だけ が適用される ID の数に、インスタンスの数が

参照される関連モデルとは異なることに注目してください。) 例えば、bread -> cheese というルールが与えられた場合、パン とチーズ を含む学習データ内の ID の数は、インスタンスと呼ばれます。

- 範囲は、前提条件が真 (true) である学習データ中の ID の割合を表示します。例えば、学習データの 50% が前提条件パン を含めば、bread -> cheese というルールのサポートは 50% となります。(上記のように、アソシエーション・モデルの場合と異なって、サポートはインスタンスの数を基準にしていません。)
- 確信度は、ルールで予測が行われたすべての ID の中で、正しい予測が行われた ID の割合を表しています。これは、学習データに基づいて、見つかったシーケンス全体を前提条件が見つかった ID 数で除算した ID 数として算出されます。例えば、学習データの 50% が cannedveg (前提条件サポートです) を含むだけでなく、20% が cannedveg と frozenmeal の両方 (ルール・サポートです) も含んでいる場合、cannedveg -> frozenmeal というルールの確信度は、Rule Support / Antecedent Support で計算され、この場合は、40% となります。
- シーケンス・モデルの ルール・サポートはインスタンスを基準にしていて、ルール全体、前提条件、結論が真 (true) となる学習レコードの比率を表示します。例えば、学習データの 20% がパン とチーズの両方を含む場合、bread -> cheese というルールのルール・サポートは、20% となります。

比率は、総トランザクション数ではなく、有効なトランザクション数 (最低 1 つの観測されている項目または真 (true) の値があるトランザクション) に基づいていることに注意してください。不正なトランザクション、つまり項目または真 (true) の値がないトランザクションは、これらの計算から除外されています。

「フィルター」ボタン : メニューにある「フィルター」ボタン (じょうごの形をしたアイコン) を押すと、ダイアログ・ボックスのボタンが展開され、パネルには有効なルール・フィルターが表示されます。フィルターは、「モデル」タブに表示されるルール数を減らすために使用します。

図 *52.* 「フィルター」ボタン

フィルターを作成するには、展開したパネルの右に表示される「フィルター」アイコンをクリックします。クリックすると、ルールの表示に関する制約を指定するための別個のダイアログ・ボックスが開きます。注意を要するのは、「フィルター」ボタンが「ノードの生成」メニューとともによく使用されることで、これにより、最初にルールにフィルターをかけ、次にルールのサブセットを含むモデルを生成します。詳しくは、以下の 281 ページの『ルールに適用するフィルターを指定する』を参照してください。

シーケンス・モデル・ナゲットの設定

シーケンス・モデル・ナゲットの「設定」タブでは、モデルのスコアリング・オプションが表示されます。このタブが利用可能になるのは、モデルがスコアリングのストリーム キャンバスに追加された後です。

最大予測数 : バスケット・アイテムの各セットに含まれる最大予測数を指定します。このトランザクションのセットに適用される最高確信度の値を持つルールは、指定限度までレコードの予測を生成するために使用されます。

シーケンス・モデル・ナゲットの要約

シーケンス・ルールのモデル・ナゲットの「要約」タブには、発見ルール数、およびルールの範囲と確信度の最小値と最大値が表示されます。このモデル作成ノードに接続されている精度分析ノードを実行した場合、その精度分析情報もこのセクションに表示されます。

詳しくは、トピック 43 ページの『モデル・ナゲットの参照』を参照してください。

シーケンス・モデル・ナゲットからルール・スーパーノードを作成

シーケンス・ルールに基づくルール・スーパーノードを作成する手順は、次のとおりです。

1. シーケンス・ルールのモデル・ナゲットの「モデル」タブで、テーブルの行をクリックして希望するルールを選択します。
2. ルール・ブラウザー・メニューから次の項目を選択します。

   「生成」 > 「ルール スーパーノード」

重要 : 生成されたスーパーノードを使用するには、データをスーパーノードに渡す前に、ID フィールド順に (ある場合は時間フィールド順にも) ソートしておく必要があります。ソートされていないと、スーパーノードは正しくデータ中のシーケンスを検出できません。

ルール・スーパーノードを生成するには、次のオプションを指定する必要があります。

検出**:** スーパーノードに渡されたデータに、どのように一致が定義されているかを指定します。

- 前提条件のみ**:** スーパーノードは、結果が見つかったかどうかにかかわらず、同じ ID を持つレコード・セット内に、選択したルールの先行条件が見つかった場合に、一致と判断します。これには、元のシーケンス・モデル作成ノードのタイムスタンプの許容度やアイテム隔たり制限の設定は考慮されていないことに注意してください。ストリーム中の最後の先行条件アイテム・セットが検出されたら (また、すべての先行条件が正しい順序で見つかっていたら)、現在の ID を持つ以降のすべてのレコードには、下で選択された要約が含まれます。
- シーケンス全体**:** スーパーノードは、同じ ID を持つレコード・セット内に、正しい順序で選択したルールの先行条件および結果が見つかった場合に、一致と判断します。これには、元のシーケンス・モデル作成ノードからのタイムスタンプ許容度またはアイテム隔たり制限の設定は、考慮されていません。ストリーム中に結果が検出された場合 (また、すべての先行条件も正しい順序で見つかった場合)、現在のレコードと、現在の ID を持つ以降すべてのレコードには、下で選択された要約が含まれています。

表示: ルール・スーパーノードの出力に、データにどのように一致要約が追加されるかを指定します。

- 最初の結果値**:** データに追加された値は、最初に発生した一致に基づいて予測された結果値です。値は、*rule_n_consequent* という名前の新規フィールドとして追加されます。*n* はルール番号です。この番号は、ストリーム内のルール・スーパーノードの作成順序に基づいています。
- 最初の真の値**:** データに追加された値は、ID に対して 1 つ以上の一致があったら真、一致がない場合は偽になります。値は、*rule_n_flag* という名前の新規フィールドとして追加されます。
- 出現回数**:** データに追加された値は、ID の一致数になります。値は、*rule_n_count* という名前の新規フィールドとして追加されます。
- ルール番号**:** 追加された値は、選択されたルールのルール番号になります。ルール番号は、スーパーノードがストリームに追加された順番に基づいて割り当てられます。例えば、最初のルール・スーパーノードは *rule 1*、2 番目のルール・スーパーノードは *rule 2*、のように番号が付けられていきます。このオプションは、ストリーム中に複数のルール・スーパーノードを配置する場合に役立ちます。値は、*rule_n_number* という名前の新規フィールドとして追加されます。
- 確信度を含める**:** このオプション選択した場合、データ・ストリームに、選択した要約のほかにルール確信度値も追加されます。値は、*rule_n_confidence* という名前の新規フィールドとして追加されます。

アソシエーション・ルール・ノード

アソシエーション・ルールは、次の形式のステートメントです。

例えば、顧客がひげそりとアフター・シェーブ ローションを購入した場合、その顧客は 80 % の確信度でシェービング クリームを購入します。アソシエーション・ルール・ノードは、データからルール・セットを抽出し、情報内容が最も充実したルールを引き出します。アソシエーション・ルール・ノードは、Apriori ノードに非常に似ていますが、重要な違いがいくつかあります。

- アソシエーション・ルール・ノードはトランザクション形式のデータを処理できません。
- アソシエーション・ルール・ノードは、ストレージ・タイプが「一覧」、測定レベルが「集計棒グラフ」であるデータを処理できます。
- アソシエーション・ルール・ノードは、IBM SPSS Analytic Server とともに使用できます。これは、スケーラビリティーが得られ、ビッグデータの処理や高速な並列処理の利用が可能になることを意味しています。
- アソシエーション・ルール・ノードは、追加の設定 (生成するルールの数を制限する機能など) を備えているため、処理速度を向上させることが可能です。
- モデル・ナゲットからの出力は、出力ビューアーに表示されます。

注: アソシエーション・ルール・ノードは、IBM SPSS Collaboration and Deployment Services 内のモデル評価ステップまたは Champion Challenger ステップをサポートしません。

注: フィールドのデータ型がフラグ型である場合、アソシエーション・ルール・ノードは、モデルを構築するときに空のレコードを無視します。空のレコードとは、モデル構築に使用されるすべてのフィールドの値が false であるレコードです。

アソシエーション ルールの実用的な使用例を示すストリーム (geospatial_association.str) およびこれにより参照されるデータ ファイル InsuranceData.sav、CountyData.sav、および ChicagoAreaCounties.shp が、IBM SPSS Modeler インストール済み環境の Demos ディレクトリにあります。Demos ディレクトリーは、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。geospatial_association.str ファイルは streams ディレクトリにあります。

アソシエーション ルール - フィールド オプション

「フィールド」タブで、上流のノード (以前のデータ型ノードなど) で既に定義されている、フィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

定義済みの役割を使用
: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。入力の役割を割り当てられたフィールドは、条件と見なされます。対象の役割を割り当てられたフィールドは、予測と見なされます。入力および対象として使用されるフィールドは、両方の役割を割り当てられていると見なされます。

ユーザー設定フィールドの割り当てを使用
: この画面で対象、予測、およびその他の役割を手動で割り当てる場合、このオプションを選択します。

フィールド
: 「ユーザー設定フィールドの割り当てを使用」を選択した場合は、矢印ボタンを使用して、このリストから画面右側のボックスに手動で項目を割り当てます。アイコンは、各フィールドの有効な測定の尺度を示します。

両方(条件または予測)
: このリストに追加されたフィールドは、モデルが生成したルールにおいて、条件または予測のいず

れかの役割を取得します。これは、ルールに応じて異なります。したがって、フィールドは、ルールによっては、条件になる場合も予測になる場合もあります。

予測のみ
: このリストに追加されたフィールドは、ルールの予測 (「結果」とも呼ばれる) としてのみ表示できます。このリストに表示された場合、そのフィールドはどのルールでも使用されるのではなく、使用される場合は、予測にしかならないことを意味しているにすぎません。

条件のみ
: このリストに追加されたフィールドは、ルールの条件 (「前提条件」とも呼ばれる) としてのみ表示できます。このリストに表示された場合、そのフィールドはどのルールでも使用されるのではなく、使用される場合は、条件にしかならないことを意味しているにすぎません。

アソシエーション ルール - ルール作成

ルールごとの項目数

各ルールで使用可能な項目または値の数を指定するには、以下のオプションを使用します。

注: 以下の 2 つのフィールドを組み合わせた合計が 10 を超えることはできません。

条件の最大数
: 単一のルールに含めることができる条件の最大数を選択します。

予測の最大数
: 単一のルールに含めることができる予測の最大数を選択します。

ルールの構築

構築するルールの数とタイプを指定するには、以下のオプションを使用します。

ルール最大値
: モデルのルールを構築する際に使用を検討できるルールの最大数を指定します。

上位 N 件のルール基準
: 上位 N 件のルールを確立するために使用する基準を選択します。ここで N は、「ルール最大値」フィールドに入力された値です。以下の基準から選択できます。

  - 確信度
  - ルール サポート
  - 条件サポート
  - リフト
  - デプロイアビリティー

フラグは真の値のみ
: データがテーブル形式の場合、このオプションを選択すると、結果のルールには、フラグ型フィールドの真の値 (true) のみが含まれるようになります。真の値が選択されることで、ルールの理解が容易になります。このオプションは、トランザクション形式のデータには適用されません。 詳しくは、 272 ページの『テーブル形式データとトランザクション形式・データ』を参照してください。

ルール基準

「ルール基準の有効化」を選択した場合は、以下のオプションを使用して、モデルでの使用を検討するためにルールが満たしている必要のある最小強度を選択できます。

- 確信度: モデルが生成するルールの確信度レベルの最小パーセント値を指定します。モデルが生成したルールのレベルがこの値よりも小さい場合、そのルールは破棄されます。
- ルール サポート: モデルが生成するルールのルール サポート レベルの最小パーセント値を指定します。モデルが生成したルールのレベルがこの値よりも小さい場合、そのルールは破棄されます。
- 条件サポート: モデルが生成するルールの条件サポート レベルの最小パーセント値を指定します。モデルが生成したルールのレベルが、指定した値よりも小さい場合、そのルールは破棄されます。
- リフト: モデルが生成したルールに許可される最小リフト値を指定します。モデルが生成したルールの値が、指定した値よりも小さい場合、そのルールは破棄されます。

ルールの除外

2 つ以上のフィールド間のアソシエーションが既知、または自明の場合があります。その場合は、フィールドが互いに予測するルールを除外できます。両方の値を含むルールを除外することにより、無関係の入力が削減され、有用な結果が検出される可能性が高まります。

フィールド
: ルール構築時に一緒に使用しない関連フィールドを選択します。例えば、車のメーカーと車のモデル、学年と生徒の年齢が関連フィールドである場合があります。モデルがルールを作成したときに、ルールのいずれかの側 (条件または予測) で選択されたフィールドの少なくとも一つがルールに含まれている場合、そのルールは破棄されます。

アソシエーション ルール - 変換

データ分割

連続型 (数値範囲) フィールドの分割方法を指定するには、以下のオプションを使用します。

ビン数
: 自動分割されるように設定されたすべての連続型フィールドが、指定したビン数に分割されます (各ビンの領域は等しい)。2 から 10 までの範囲で任意の数値を選択できます。

リスト フィールド

リストの最大長
: リスト フィールドの長さが不明の場合、モデルに含める項目の数を制限するには、リストの最大長を入力します。1 から 100 までの範囲で任意の数値を選択できます。入力した数値よりもリストが長い場合、モデルは、依然としてフィールドを使用しますが、含められる値は、この数値までに限られます。フィールド内の余分な値はすべて無視されます。

アソシエーション ルール - 出力

モデルの構築時に生成される出力を制御するには、このペインのオプションを使用します。

ルール テーブル

選択した基準ごとに (指定した数値に基づいて) 最適な数のルールを表示するテーブル タイプを 1 つ以上作成するには、以下のオプションを使用します。

確信度
: 確信度は、ルール サポートと条件サポートの比です。リストされた条件値を持つ項目のうち、予測された結果値を持つ項目の割合です。出力に含められる確信度に基づく N 個の最適なアソシエーション・ルールを含むテーブルを作成します (ここで N は、「表示するルール **(Rules to display)**」の値です)。

ルール サポート
: ルール全体、条件、予測が真 (true) となる項目の比率です。データセットのすべての項目において、ルールにより正確に説明および予測される割合です。この指標により、全体的なルールの重要度が得られます。出力に含められるルール サポートに基づく N 個の最適なアソシエーション・ルールを含むテーブルを作成します (ここで N は、「表示するルール **(Rules to display)**」の値です)。

リフト
: 予測が得られる事前確率とルールの確信度の比率。ルールの確信度値と、母集団で結果値が発生する割合の比率です。この比率により、ルールが見込みよりどの程度改善されるかを示す指標が得られます。出力に含められるリフトに基づく N 個の最適なアソシエーション・ルールを含むテーブルを作成します (ここで N は、「表示するルール **(Rules to display)**」の値です)。

条件サポート
: 条件が真 (true) となる項目の比率です。出力に含められる前提条件サポートに基づく N 個の最適なアソシエーション・ルールを含むテーブルを作成します (ここで N は、「表示するルール **(Rules to display)**」の値です)。

デプロイアビリティー
: 条件を満たしているが予測を満たしていない学習データのパーセンテージの指標。この指標では、ルールがあてはまらなくなる頻度が示されます。これは事実上、確信度の反対です。出力に含められるデプロイアビリティーに基づく N 個の最適なアソシエーション・ルールを含むテーブルを作成します (ここで N は、「表示するルール **(Rules to display)**」の値です)。

表示する規則
: テーブルに表示する規則の最大数を設定します。

モデル情報テーブル

出力に含めるモデル テーブルを選択するには、以下のオプションを 1 つ以上使用します。

- フィールド変換
- レコード要約
- ルールの統計
- 最頻出値
- 最頻出フィールド

ルールのソート可能なワード クラウド

ルール出力を表示するワード クラウドを作成するには、以下のオプションを使用します。ワードは、その重要度を示すためにさまざまな文字サイズで表示されます (文字サイズが大きいほど重要度が高くなります)。

ソート可能なワード クラウドの作成。
: ソート可能なワード クラウドを出力で作成するには、このボックスにチェック マークを付けます。

デフォルトのソート
: ワード クラウドを最初に作成するときに使用するソート タイプを選択します。ワード クラウドは対話式です。モデル ビューアーで基準を変更すると、別のルールおよびソートを参照できます。以下のソート オプションから選択できます。

  - 信頼度。
  - ルール サポート

- リフト
- 条件サポート。
- デプロイアビリティー

表示するルールの最大数
: ワード クラウドに表示するルールの数を設定します。選択できる最大数は 20 です。

アソシエーション ルール - モデル オプション

アソシエーション ルール モデルのスコアリング・オプションを指定するには、このタブの設定を使用します。

モデル名: 対象フィールド (該当するフィールドが指定されていない場合は、モデル タイプ) に自動的に基づいてモデル名を生成するか、またはカスタム名を指定することができます。

最大予測数: スコア結果に組み込む最大予測数を指定します。このオプションは、「ルール基準」の項目とともに使用され、「最上位」の予測を生成します。ここで、「最上位」というのは、確信度、サポート、リフトなどについて、最も高いレベルであることを示しています。

ルール基準 ルールの強さを決定するために、測定を選択します。アイテムセットに最上位の予測を返すために、ここで選択した基準の強さによって、ルールがソートされます。5 つの異なる基準から選択できます。

- 確信度: 確信度とは、ルール サポートと条件サポートの比率です。リストされた条件値を持つ項目のうち、予測された結果値を持つ項目の割合です。
- 条件サポート: 条件が真 (true) となる項目の比率です。
- ルール サポート: ルール全体、条件、予測が真 (true) となる項目の比率です。「条件サポート」の値と「確信度」の値を乗算して計算します。
- リフト: ルール確信度と予測が得られる事前確率の比率。
- デプロイアビリティ: 条件を満たすが、予測を満たさない学習データのパーセンテージの指標。

予測の繰り返しを許可: スコアリング時に同じ予測を持つ複数のルールを含めるには、このチェック・ボックスを選択します。例えば、このオプションを選択すると、以下のルールをスコアリングできます。

```
bread & cheese -> wine
cheese & fruit -> wine
```

注: 複数の予測を持つルール (**パン & チーズ & フルーツ -> ワイン & パテ**) は、すべての予測 (**ワイン & パテ**) があらかじめ予測されている場合にのみ、予測の繰り返しと見なされます。

入力に予測が存在しない場合のみルールをスコアリング: 入力に予測も存在しないようにするには、このオプションを選択します。スコアリングの目的が家具製品の提案である場合を例に挙げます。ここでは、入力に既にダイニング・テーブルが入っている状態でもう 1 つ購入する可能性は低いと考えられます。そのような場合に、このオプションを選択します。ただし、製品が生鮮食料品や使い捨て製品の場合 (チーズ、粉ミルク、ティッシュ・ペーパーなど) は、結果をあらかじめ入力に表示するルールが有用となることがあります。後者の場合、最も便利なオプションは、「すべてのルールをスコアリング」です。

入力に予測が存在する場合のみルールをスコアリング: 入力に予測も存在するようにするには、このオプションを選択します。このアプローチは、既存の顧客やトランザクションに対する洞察を得ようとする場合に役立ちます。例えば、最上位のリフトを持つルールを識別したり、さらにどの顧客がそのルールに適合するのかを調べたい場合があります。

すべてのルールをスコアリング: 予測の有無にかかわらず、スコアリング時にすべてのルールを含めるには、このオプションを選択します。

アソシエーション ルールのモデル ナゲット

モデル・ナゲットには、モデル構築中にデータから抽出されたルールに関する情報が含まれます。

結果の表示

アソシエーション ルール モデルにより生成されたルールを参照するには、ダイアログ ボックスの「モデル」タブを使用します。モデル・ナゲットを参照することにより、新規ノードの生成、またはモデルのスコアリングを行う前に、ルールに関する情報を確認できます。

モデルのスコアリング

調整済みモデル・ナゲットは、ストリームに追加されて、スコアリングに使用されることもあります。詳しくは、トピック 49 ページの『ストリーム内でのモデル・ナゲットの使用』を参照してください。スコアリングに使用したモデル・ナゲットには、追加の設定タブがあり、それぞれのダイアログ・ボックスがあります。詳しくは、トピック『アソシエーション ルールのモデル ナゲットの設定』を参照してください。

アソシエーション ルールのモデル ナゲットの詳細

アソシエーション ルールのモデル ナゲットは、出力ビューアの「モデル」タブにモデルの詳細を表示します。ビューアーの使用方法について詳しくは、「Modeler ユーザーズ・ガイド」(ModelerUsersGuide.pdf) の『出力の処理』を参照してください。

GSAR モデリング操作で、次の表に示すように接頭辞 $A が付いた多数の新規フィールドを作成します。

表 25. アソシエーション ルール モデリング操作により作成される新規フィールド

フィールド名	説明
$A-<prediction>#	このフィールドには、スコアリングされたレコードのモデルからの予測が含まれます。 <prediction> は、モデルの予測役割に入っているフィールドの名前です。# は、出力ルールの数値シーケンスです (例えば、3 つのルールを含めるようにスコアが設定されている場合、数値シーケンスは 1 から 3 になります)。
$AC-<prediction>#	このフィールドには、予測の確信度が含まれます。 <prediction> は、モデルの予測役割に入っているフィールドの名前です。# は、出力ルールの数値シーケンスです (例えば、3 つのルールを含めるようにスコアが設定されている場合、数値シーケンスは 1 から 3 になります)。
$A-Rule_ID#	この列には、スコアリングされたデータセット内の各レコードについて予測されたルールの ID が含まれます。 # は、出力ルールの数値シーケンスです (例えば、3 つのルールを含めるようにスコアが設定されている場合、数値シーケンスは 1から 3 になります)。

アソシエーション ルールのモデル ナゲットの設定

アソシエーション ルール モデル・ナゲットの「設定」タブでは、モデルのスコアリング オプションが表示されます。このタブが利用可能になるのは、モデルがスコアリングのストリーム キャンバスに追加された後です。

最大予測数: 各アイテム・セットに含める最大予測数を指定します。このトランザクションのセットに適用される最高確信度の値を持つルールは、指定限度までレコードの予測を生成するために使用されます。このオプションは、「最上位」の予測を行うために「ルール基準」オプションとともに使用します。ここで、最上位 というのは、確信度、サポート、リフトなどについて、最も高いレベルであることを示しています。

ルール基準 ルールの強さを決定するために、測定を選択します。アイテムセットに最上位の予測を返すために、ここで選択した基準の強さによって、ルールがソートされます。以下の基準から選択できます。

- 確信度
- ルール サポート
- リフト
- 条件サポート
- デプロイアビリティー

予測の繰り返しを許可: スコアリング時に同じ結果を持つ複数のルールを含めるには、このチェック・ボックスを選択します。例えば、このオプションを選択すると、以下のルールをスコアリングできます。

```
bread & cheese -> wine
cheese & fruit -> wine
```

スコアリング時に予測の繰り返しを除外するには、このチェック・ボックスをクリアします。

注: 複数の結果を持つルール (**パン & チーズ & フルーツ -> ワイン & パテ**) は、すべての結果 (**ワイン & パテ**) があらかじめ予測されている場合のみ、反復の予測と見なされます。

入力に予測が存在しない場合のみルールをスコアリング: 入力に結果も存在しないようにする場合に選択します。スコアリングの目的が家具製品の提案である場合を例に挙げます。ここでは、入力に既にダイニング・テーブルが入っている状態でもう 1 つ購入する可能性は低いと考えられます。そのような場合に、このオプションを選択します。逆に、製品が生鮮食料品や使い捨て製品の場合 (チーズ、粉ミルク、ティッシュ・ペーパーなど) は、結果をあらかじめ入力に表示するルールが有用となることがあります。後者の場合、最も便利なオプションは、「すべてのルールをスコアリング」です。

入力に予測が存在する場合のみルールをスコアリング: このオプションは、入力に結果も存在するようにする場合に選択します。このアプローチは、既存の顧客やトランザクションに対する洞察を得ようとする場合に役立ちます。例えば、最上位のリフトを持つルールを識別したり、さらにどの顧客がそのルールに適合するのかを調べたい場合があります。

すべてのルールをスコアリング 入力の結果の有無にかかわらず、スコアリング時にすべてのルールを含めるには、このオプションを選択します。

第 13 章 時系列モデル

なぜ予測できるのでしょうか?

予測とは、時間の経過に伴う 1 つ以上の系列の数値を予測することです。例えば、製造や流通に対して資源を割り当てるために、一連の製品やサービスに対する需要の見込みを予測したい場合があります。意思決定は実行するのに時間を要するため、多くの企画プロセスにおいて予測は不可欠なツールとなっています。

時系列のモデル作成方法では、まったく同じでないにしても過去に起こったことは繰り返され、過去に起こったことを検証することでその精度は近づき、将来に対するよりよい意思決定ができると仮定しています。例えば、翌年の売り上げを予測するために、おそらく今年の売り上げに目を通すことから始まり、さかのぼって近年構築された傾向やパターンがあればそれを検証するでしょう。しかし、パターンを計測することは困難です。例えばもし、数年にわたって売り上げが成長した場合、それは季節性のサイクルなのでしょうか?それとも長期傾向の始まりなのでしょうか?

統計的モデル作成の技術を使用し、過去のデータにあるパターンを分析し、これらのパターンから推定して、系列の将来値が収まると思われる範囲を決定することができます。結果は、意思決定の基礎となるより正確な予測です。

時系列データ

時系列は、例えば日々の在庫価格や週間売上データなど、一定の間隔で行われた測定値を順番に収集したものです。測定値は対象となるもので、各系列は一般的に次のように区分されます。

- 従属変数: 予測したい系列です。
- 予測値: 広告予算を使用して売り上げを予測するなど、ターゲットの説明を支援する系列です。予測値は、ARIMA モデルでのみ使用できます。
- イベント: 販売促進など、繰り返し発生する予測可能な出来事を説明するために使用する、特別な予測値の系列です。
- 干渉: 停電や従業員のストライキなど、1 度だけ発生した過去の出来事を説明するのに使用される、特別な予測値の系列です。

間隔は、どのような時間の単位でも表すことができますが、すべての測定において一定の間隔である必要があります。さらに、測定のないいかなる間隔も、欠損値に設定される必要があります。そのため、測定 (欠損値がある場合を含めて) がある間隔数は、データの履歴スパンの時間の長さを定義します。

時系列の特徴

系列の過去の動きを検証することで、パターンを識別し、より正確な予測を行うことができます。作図する場合、多くの時系列が、次の特徴のうち 1 つ以上示します。

- 傾向
- 季節性および非季節性サイクル
- パルスおよびステップ
- 外れ値

傾向

傾向は、時間の経過に伴って増加または減少する系列または流れのレベルにおいて、徐々に増加もしくは減少することです。

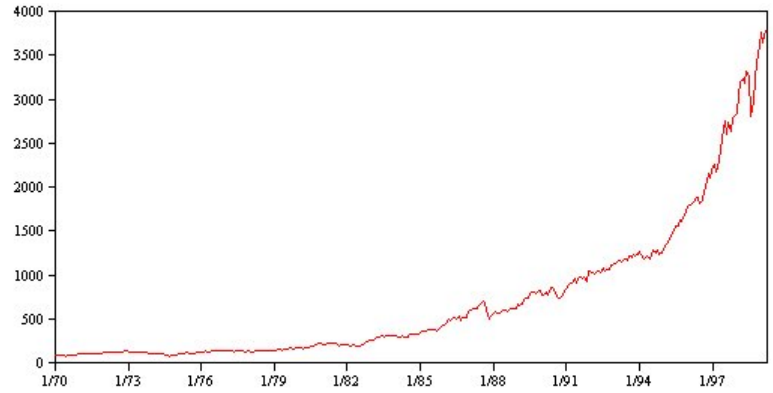


図 *53.* 傾向

傾向は、ローカルまたはグローバルのどちらか一方ですが、単一の系列は両方のタイプを示すことがあります。これまでは、株式市場の一連の作図は、グローバルな増加傾向を示しています。ローカルの減少傾向は不景気の時に現れ、またローカルの増加傾向は好景気のときに現れます。

また傾向は、線形または非線形のものがあります。線形傾向は、系列のレベルに対し正または負の相加的増分で、元金に対する単利の影響と比較できます。非線形傾向は、以前の一連の数値に対し比例する増分が含まれ、相乗的であることが多く見られます。

グローバルな線形傾向は、指数平滑方および ARIMA モデルによって適合し、的確な予測となります。ARIMA モデルの構築の際、傾向を示す系列は、一般的に傾向の影響を削除する差異となります。

季節性サイクル

季節性サイクルは、系列の値において繰り返し発生し予測可能なパターンです。

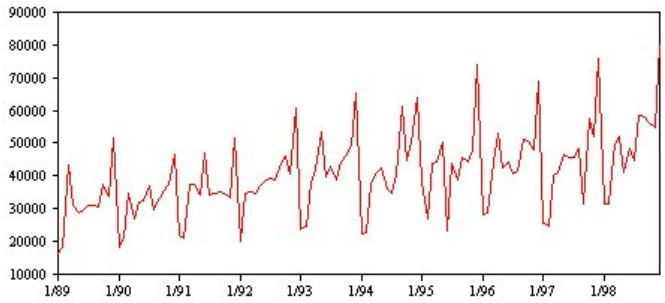


図 *54.* 季節性サイクル

季節性サイクルは、系列の間隔と関連しています。例えば、月間データは一般的に四半期および年間のサイクルで循環します。月間の系列は、第 1 四半期で低い値を示す有意な四半期のサイクル、または毎年 12 月にピークとなる年間サイクルを示す場合があります。季節ごとのサイクルを示す系列は、季節性と呼ばれています。

季節性パターンは、より正確な適合や予測を得るのに役に立ちます。また、指数平滑法や季節性を取得する ARIMA モデルが含まれています。

非季節性サイクル

非季節性サイクルは、系列の値において繰り返し発生し予測可能なパターンです。

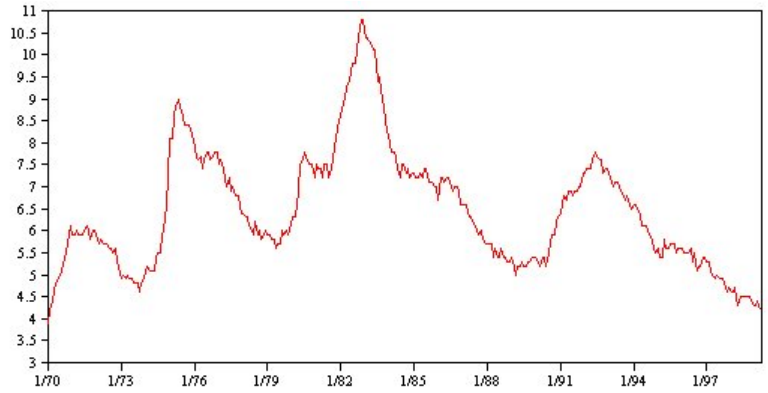


図 *55.* 非季節性サイクル

失業率など、景気による作用を明確に示す系列もありますが、サイクルの周期性は時期に応じて異なるため、高いもしくは低い値が発生する場合、予測することが困難になります。予測可能なサイクルの系列もありますが、グレゴリオ暦に正確に一致せず、 1 年以上のサイクルはありません。例えば、潮汐は太陰暦に従い、オリンピック関連の海外旅行や国際取引は 4 年おきに増加し、またグレゴリア暦の日付が年ごとに変わる宗教上の休日が数多くあります。

非季節性サイクルのパターンは、モデル作成を行うには難しく、一般的には予測時の不確定要素が増大します。例えば株式市場では、予測の影響を無視した系列の例を数多く提供します。それでも、非季節性のパターンが存在する場合には、説明される必要があります。多くの場合、合理的に履歴データを適合するモデルを指定することで、予測時の不確定要素を最小化するチャンスを最大限に得ることができます。

パルスおよびステップ

多くの系列で、レベルの突然の変更を経験します。一般的に 2 つのタイプがあります。

- 系列レベルにおける、突然の一時的な 移行、またはパルス
- 系列レベルにおける、突然の永久の 移行、またはステップ

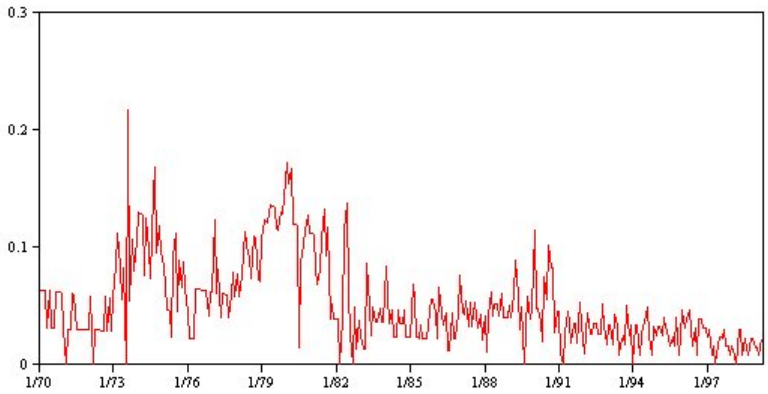


図 *56.* パルスが含まれる時系列

ステップまたはパルスが観察された場合、納得のいく説明を見つけることが重要です。時系列モデルは、突然ではなく徐々に表れた変化を説明するために設計されています。結果として、モデルはパルスを過小評価

し、ステップによって損なわれる傾向にあり、そのため低品質のモデルを適合し、不確実な予測を行うことになります。(突然のレベル変更を示す季節性の例もありますが、レベルはある季節から次の季節まで一定です。)

混乱を説明することができる場合、干渉またはイベントを使用してモデル作成をすることができます。例えば、1973 年の 8 月中、石油輸出国機構 (OPEC) が行った石油貿易禁止措置によって物価上昇率が劇的に変化し、次月には通常のレベルに回復しました。石油貿易禁止の月に対してポイント干渉を指定することで、モデルの適合を向上させ、間接的に予測を向上させることができます。例えば、小売店で通常の売り上げより非常に高い売り上げを記録した場合、その日は全商品 50 % 引きで販売しました。50% オフの販売を繰り返し行われるイベントとして指定することによって、モデルの適合を向上させ、将来販促活動を繰り返し行う効果を見積もることができます。

外れ値

説明できない時系列のレベルへの移行は、外れ値として参照されます。これらの観察は系列とは一貫しておらず、分析に劇的な影響を与えることができ、以後時系列モデルの予測能力に影響を与えます。

次の図では、時系列で一般的に発生する外れ値の種類をいくつか紹介します。青線は、外れ値の含まない系列を表します。赤線は、系列が外れ値を含む場合に起こりうるパターンを表示しています。これらの外れ値は、系列の平均レベルにのみ影響を与えるため、決定的なものとして分類されます。

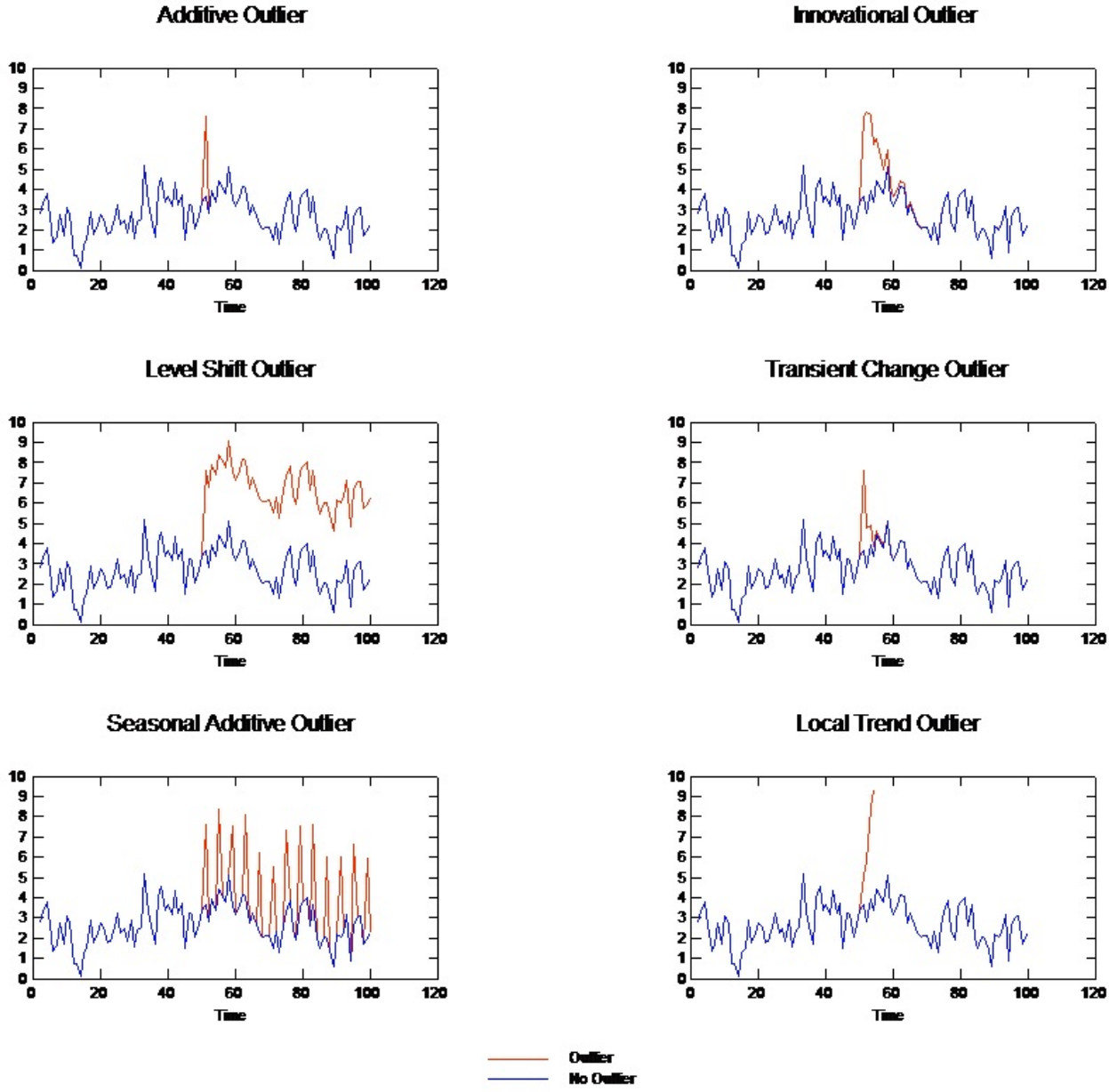


図 *57.* 外れ値のタイプ

- 相加的外れ値**:** 相加的外れ値は、単一の観察に対して発生する非常に大きいまたは小さい値として表示されます。後続の観察は、相加的外れ値の影響を受けません。連続する相加的外れ値は、一般的に相加的外れ値パッチと呼ばれます。
- 技術革新的外れ値**:** 技術革新的外れ値は、後続の観察にも続く初期の影響によって特徴付けられています。外れ値の影響は、時間が進むにつれて拡大する場合があります。
- レベル・シフト外れ値**:** レベル・シフトの場合、外れ値の後に表示されるすべての外れ値は、新しいレベルに移行します。相加的外れ値とは対照的に、レベル・シフト外れ値は多くの観察に影響を与え、また影響は永久に続きます。
- 過渡変化外れ値**:** 過渡変化外れ値は、レベル・シフト外れ値と類似していますが、後続の観察に対する外れ値の影響は急激に減少します。その結果、系列は通常のレベルに戻ります。

- 周期的付加外れ値: 季節性付加外れ値は、一定の間隔で繰り返し発生する非常に大きい、または小さい値として表示されます。
- 局所トレンド外れ値: ローカル トレンド外れ値は、最初の外れ値に続く外れ値のパターンによって引き起こされる系列に、一般的な傾向を得ることができます。

時系列の外れ値検出では、時系列に存在する外れ値の場所、種類、絶対値を決定します。Tsay 教授は、決定的外れ値を識別するために平均レベルの変更を検出する反復手順を提案しました (1988 年)。このプロセスでは、外れ値を導入するもうひとつのモデルに伝達される外れ値はないと仮定する時系列モデルを比較します。モデル間の相違点によって、指定されたポイントを外れ値として扱う影響を推定します。

自己相関および偏自己相関関数

自己相関および偏自己相関は、現在と過去の時系列値の関連性の測定で、どの過去の時系列値が将来値の予測に最も役立つかを示します。この知識を使用すると、ARIMA モデルにおける処理の順番を決定することができます。具体的には次のとおりです。

- 自己相関関数 **(ACF)**ラグ *k*の場合、これは間隔が離れた *k* の時系列値間における相関関係です。
- 偏自己相関関数 **(PACF)**ラグ *k*の場合、これは間隔が離れた *k* の時系列値間における相関関係で、間の値を説明します。

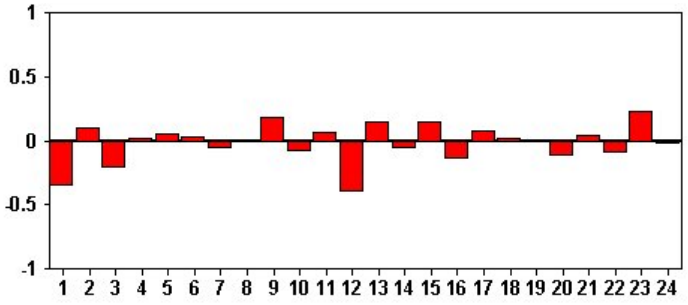


図 *58.* 時系列の *ACF* プロット

ACF プロットの *X* 座標は、自己相関が計算されたラグを表示し、*Y* 座標は相関関係の値を表示します (-1 と 1 の間)。例えば、ACF プロット内のラグ 1 の急な山形は、各時系列と先行値の間の強い相関関係を示し、ラグ 2 の山形は、以前に 2 点が発生した各値の間などの強い相関関係を示します。

- 正の相関は、大きな現在値が指定のラグの大きな値に対応することを示し、負の相関は大きな現在値は指定のラグの小さい値に対応することを示します。
- 相関の絶対値は、関連性の強さを示す測定値で、絶対値が大きいほど強い関連性を示します。

系列の変換

変換は、モデルを見積もる前に系列を固定するのに役立ちます。これは特に ARIMA モデルに関して重要で、モデルの見積もりを行う前に系列を固定する必要があります。グローバル レベル(平均値)およびレベル(分散)からの平均偏差が系列を通して一定である場合、系列は固定化されています。

重要な系列の多くは固定化されていませんが、自然対数、差異、もしくは季節性差異などの変換を適用することによって系列が固定化されている限り、ARIMA は効果的です。

差異固定変換。 時間の経過に伴って差異が変化する系列は、自然対数または平方根変換を使用して固定化することができます。これらは、関数変換とも呼ばれます。

- 自然対数: 自然対数は系列値に適用されます。

- 平方根: 平方根関数は系列値に適用されます。

自然対数および平方根変換は、負の値を持つ系列に使用することはできません。

レベル固定変換**:** ACF の値が徐々に減少することは、各系列値が以前の値と強い相関関係を持っていることを示します。系列値の変化を分析することによって、安定したレベルを取得します。

- 単純な差分**:** 系列の各値と以前の値との差異は、当然、系列内で最も古い値を除いて計算されます。つまり差異の系列には、元の系列より 1 つ小さい値が含まれます。
- 季節差分: 各値と以前の季節性値と間の差異が計算される点を除いては、シンプルな差異と同様です。

シンプルまたは季節的な差異のどちらかが、対数または平方根変換と同時に使用される場合、差異固定変換が必ず最初に適用されます。シンプルおよび季節性差異が両方使用される場合、シンプルな差異または季節性差異のどちらが最初に適用されても、結果として生じる系列値は同じです。

予測値の系列

予測値には、予測される系列の行動の説明を支援する関連データが含まれます。例えば、Webもしくはカタログ ベースの小売店は、送付するカタログ数、開設する電話回線数、会社のWeb ページのヒット数に基づいて売り上げを予測することができます。

いかなる系列も、予測する将来に拡張し欠損値なくデータを完成する予測値として使用できます。

予測値をモデルに追加する場合は、注意して使用してください。多くの予測値を追加すると、モデルの推定に必要な時間を拡大することができます。予測値を追加すると、モデルが改善され過去のデータに一致するようになりますが、必ずしもモデルがより正確な予測ができるというわけでなく、追加された複雑性がトラブルに値するということはありません。目標は、正確な予測のできる単純なモデルを指定することが理想です。

原則として、予測値の数は、 15 に分けられた標本サイズ(最大 15 ケースあたり 1 つの予測値)より小さいことが推奨されます。

欠損データを伴う予測値**:** 不完全または欠損データを含む予測値は、予測に使用することはできません。これは、過去のデータおよび将来の数値に適用します。モデルの推定スパンを設定してモデル推定時に最も古いデータを除外することによって、制限を回避することができる場合があります。

時空間予測モデル作成ノード

時空間予測 (STP) は、潜在的な用途が多数あり、ビルや施設のエネルギー管理、機械点検エンジニアのパフォーマンスの分析と予測、公共輸送機関の計画などに応用されます。多くの場合、これらの用途では、エネルギー使用量などの測定値が空間および時間に渡って取得されます。これらの測定値の記録に関する問題としては、どの因子が将来の観測に影響を与えるかや、どうすれば必要な変更を達成したりシステムをよりよく管理できるかなどがあります。これらの問題に対処するには、統計手法を駆使してさまざまなロケーションの将来の値を予測し、調整可能な因子を明示的にモデル化して what-if 分析を実行します。

STP 分析では、ロケーション・データを含むデータ、予測用の入力フィールド (予測値)、時間フィールド、および対象フィールドが使用されます。各ロケーションには、それぞれの測定時の各予測値を表すデータの行が多数あります。データを分析すると、そのデータを使用して、分析で使用される形状データ内の任意のロケーションの対象値を予測できます。また、将来のポイント・イン・タイム入力データがいつ既知になるのかを予測することもできます。

注: STP ノードは、IBM SPSS Collaboration and Deployment Services 内のモデル評価ステップまたは Champion Challenger ステップをサポートしません。

STP の実用的な使用例 (stp_server_demo.str) を表示し、データ・ファイル room_data.csv および score_data.csv を参照するストリームが、IBM SPSS Modeler インストール済み環境の Demos ディレクトリーにあります。Demos ディレクトリーは、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。stp_server_demo.str ファイルは、streams ディレクトリーにあります。

時空間予測 - フィールド・オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

定義済みの役割を使用
: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象および予測のみ) を使用します。

ユーザー設定フィールドの割り当てを使用
: この画面で対象、予測、およびその他の役割を手動で割り当てるには、このオプションを選択します。

フィールド
: データ内の選択可能なフィールドをすべて表示します。このリストの項目を画面右側の各種ボックスに手動で割り当てるには、矢印ボタンを使用します。アイコンは、各フィールドの有効な測定の尺度を示します。

  注: STP が正常に機能するには、場所当たり、時間区分当たり 1 レコードが必要です。したがって、これらは必須フィールドです。

  「フィールド」ペインの下部で、「すべて」ボタンをクリックすると、測定の尺度にかかわらず、すべてのフィールドが選択され、測定の尺度のボタンを個々にクリックすると、その測定の尺度のフィールドがすべて選択されます。

対象
: 1 つのフィールドを予測の対象として選択します。

  注: 測定の尺度が連続型であるフィールドのみを選択できます。

場所
: モデルで使用する場所タイプを選択します。

  注: 測定の尺度が地理空間であるフィールドのみを選択できます。

場所ラベル
: 多くの場合、形状データには、層の機能名を示すフィールドが含まれています。例えば、州や国の名前です。このフィールドを使用して名前またはラベルを場所に関連付けるには、カテゴリー型フィールドを選択し、出力内の選択済みの「場所」フィールドにラベルを付けます。

時間フィールド
: 予測で使用する時間フィールドを選択します。

  注: 選択できるフィールドは、測定の尺度が連続型、ストレージ・タイプが時間、日付、タイムスタンプ、または整数のフィールドのみです。

予測値 (入力)
: 1 つ以上のフィールドを予測の入力として選択します。

注**:** 測定の尺度が連続型であるフィールドのみを選択できます。

時空間予測 - 時間区分

「時間区分」ペインでは、時間区分を設定するオプションと、必要に応じて長期間の集計を設定するオプションを選択できます。

STP モデルを構築する前に、時間フィールドをインデックスに変換するためにデータを準備する必要があります。変換を可能にするには、時間フィールドで記録間に一定の区分が必要です。この情報がデータにまだ含まれていない場合は、モデル作成ノードを使用する前に、このペインのオプションを使用して、この区分を設定します。

時間区分: データ セットを変換する時間区分を選択します。使用可能なオプションは、「フィールド」タブでモデルの「時間フィールド」として選択されたフィールドのストレージ・タイプによって決まります。

- 期間: 整数型の時間フィールドでのみ使用できます。これは、他の使用可能などの区分にも一致しない区分の系列 (各測定間の区分は均一) です。
- 年: 「日付」時間フィールドまたは「タイムスタンプ」時間フィールドにのみ使用可能です。
- 四半期: 「日付」時間フィールドまたは「タイムスタンプ」時間フィールドにのみ使用可能です。このオプションを選択した場合は、第 1 四半期の「開始月」を選択するようプロンプトが出されます。
- 月: 「日付」時間フィールドまたは「タイムスタンプ」時間フィールドにのみ使用可能です。
- 週: 「日付」時間フィールドまたは「タイムスタンプ」時間フィールドにのみ使用可能です。
- 日: 「日付」時間フィールドまたは「タイムスタンプ」時間フィールドにのみ使用可能です。
- 時間: 「時刻」時間フィールドまたは「タイムスタンプ」時間フィールドにのみ使用可能です。
- 分: 「時刻」時間フィールドまたは「タイムスタンプ」時間フィールドにのみ使用可能です。
- 秒: 「時刻」時間フィールドまたは「タイムスタンプ」時間フィールドにのみ使用可能です。

「時間区分」を選択した場合は、さらにフィールドを入力するようプロンプトが出されます。使用可能なフィールドは、時間区分とストレージ・タイプの両方によって決まります。表示される可能性のあるフィールドを以下のリストに示します。

- 週あたりの日数
- **1** 日の時間数
- 週の開始: 週の初日
- **1** 日の開始: 新しい 1 日の開始と見なす時刻。
- 間隔値: 1、2、3、4、5、6、10、12、15、20、または 30 のいずれかのオプションを選択できます。
- 開始月: 会計年度の開始月。
- 開始期間: 「期間」を使用する場合、開始期間を選択します。

指定した時間区分設定に一致するデータ: データに正しい時間区分情報が既に含まれており、変換の必要がない場合は、このチェック ボックスを選択します。このボックスを選択すると、「集計」領域のフィールドが使用不可になります。

集計

「指定した時間区分設定に一致するデータ」チェック・ボックスをクリアした場合にのみ使用可能です。指定した区分に一致するフィールドを集計するオプションを指定します。例えば、週と月が混在するデータが

あるとします。月区分を均一にするために週の値を「ロール・アップ」して集計することができます。さまざまなフィールド・タイプの集計に使用するデフォルト設定を選択し、特定のフィールドに必要なカスタム設定を作成します。

- 連続型: 個々に指定されていないすべての連続型フィールドに適用するデフォルトの集計方法を設定します。以下に示す複数の方法から選択できます。
  - 合計
  - 平均値
  - 最小値
  - 最大値
  - 中央値
  - 最初の 4 分位
  - 3 番目の 4 分位

指定されたフィールドのカスタム設定: 特定の集計関数を個々のフィールドに適用するには、このテーブルでフィールドを選択してから、集計方法を選択します。

- フィールド: 「フィールドの追加」ボタンを使用して「フィールドの選択」ダイアログ・ボックスを表示し、必要なフィールドを選択します。選択したフィールドがこの列に表示されます。
- 集計関数: ドロップダウン・リストから集計関数を選択すると、指定した時間区分にフィールドが変換されます。

時空間予測 - 基本的な作成オプション

基本的なモデル作成オプションを設定するには、このダイアログ・ボックスの設定を使用します。

モデルの設定

定数項を含める

定数項 (モデルの定数項) を含めると、解の精度を全体的に向上させることができます。データが原点を通ると仮定できる場合は、切片を除外できます。

自己回帰の最大次数

自己回帰の次数は、以前のどの値を使用して現在の値を予測するかを指定します。このオプションを使用すると、新しい値の計算に使用するために以前のレコードの数を指定できます。1 から 5 までの間の任意の整数を選択できます。

空間共分散

推定方法

使用する推定方法を選択します。「パラメトリック」または「ノンパラメトリック」を選択できます。「パラメトリック」の方法の場合は、以下の 3 つの「モデル」タイプの中から 1 つを選択できます。

- ガウス
- 指数
- 指数のべき乗: このオプションを選択する場合は、使用する「べき乗」レベルも指定する必要があります。このレベルには、0.1 きざみで 1 から 2 までの間の任意の値を指定できます。

時空間予測 - 高度な作成オプション

STP に関して詳細な知識を持つユーザーは、以下のオプションを使用して、モデル作成プロセスを微調整できます。

欠損値の最大パーセンテージ
: モデルに含めることのできる、欠損値のあるレコードの最大パーセンテージを指定します。

モデル構築での仮説検証の有意水準
: STP モデル推定のすべての検定 (適合度検定、効果 F 検定、係数 T 検定を含む) に使用する有意水準値を指定します。この水準には、0.01 きざみで 0 から 1 までの間の任意の値を指定できます。

時空間予測 - 出力

モデルを構築する前に、このペインのオプションを使用して、出力ビューアーに含める出力を選択します。

モデル情報

モデル指定
: このオプションを選択すると、モデル指定の情報がモデル出力に含まれます。

時間的情報の要約
: このオプションを選択すると、時間的情報の要約がモデル出力に含まれます。

評価

モデル品質
: このオプションを選択すると、モデル品質がモデル出力に含まれます。

平均構造モデルでの効果のテスト
: このオプションを選択すると、効果のテストの情報がモデル出力に含まれます。

解釈

平均構造モデル係数
: このオプションを選択すると、平均構造モデル係数の情報がモデル出力に含まれます。

自己回帰係数
: このオプションを選択すると、自己回帰係数の情報がモデル出力に含まれます。

空間上での減衰のテスト
: このオプションを選択すると、空間共分散または空間上での減衰のテストの情報がモデル出力に含まれます。

パラメトリック空間共分散モデルのパラメーター・プロット
: このオプションを選択すると、パラメトリック空間共分散モデルのパラメーター・プロットの情報がモデル出力に含まれます。

  注: このオプションは、「基本」タブで「パラメトリック」推定方法を選択した場合にのみ使用可能になります。

相関ヒート・マップ
: このオプションを選択すると、対象値のマップがモデル出力に含まれます。

  注: モデル内に 500 を超える場所がある場合、マップ出力は作成されません。

相関マップ
: このオプションを選択すると、相関のマップがモデル出力に含まれます。

  注: モデル内に 500 を超える場所がある場合、マップ出力は作成されません。

場所のクラスター
: このオプションを選択すると、場所のクラスタリング出力がモデル出力に含まれます。マップ データへのアクセスが不要な出力のみがクラスター出力の一部として含められます。

  注: この出力は、非パラメトリック空間共分散モデルに対してのみ作成できます。

  このオプションを選択する場合は、以下を設定できます。

  - 類似性閾値: ここで選択した閾値に達すると、出力クラスターは類似性が十分であると見なされ、単一クラスターにマージされます。
  - 表示するクラスターの最大数: モデル出力に含めることのできるクラスター数の上限を設定します。

時空間予測 - モデル・オプション

モデル名: 対象フィールドに基づいて自動的にモデル名を生成することも、カスタム名を指定することもできます。自動的に生成された名前は、対象フィールド名です。

不確実性因子 **(%)**: 不確実性因子とは、将来を予測する際の不確実性の増加を表すパーセント値です。将来に向けたステップごとに、このパーセンテージ分だけ、予測の不確実性の上限と下限が増加します。モデル出力に適用する不確実性因子を設定します。これにより、予測値の上限分割点と下限分割点が設定されます。

時空間予測モデル・ナゲット

時空間予測 (STP) モデル・ナゲットは、出力ビューアーの「モデル」タブにモデルの詳細を表示します。ビューアーの使用方法について詳しくは、「Modeler ユーザーズ・ガイド」(ModelerUsersGuide.pdf) の『出力の処理』を参照してください。

時空間予測 (STP) モデリング操作では、以下の表に示すように接頭辞 $STP- が付いた新規フィールドが多数作成されます。

表 *26. STP* モデリング操作で作成された新規フィールド

フィールド名	説明
$STP-<Time>	モデル構築の一環として作成される時間フィールドです。このフィールドの作成方法は、「作成オプション」タブの「時間区分」ペインの設定で決定されます。 <Time> は、「フィールド」タブで「時間フィールド」として選択されたフィールドの元の名前です。 **注:** このフィールドは、モデル構築の一環として元の「時間フィールド」を変換した場合にのみ作成されます。
$STP-<Target>	このフィールドには、対象値の予測が含まれます。 <Target> は、モデルの元の「対象」フィールドの名前です。
$STPVAR-<Target>	このフィールドには、VarianceOfPointPrediction 値が含まれます。 <Target> は、モデルの元の「対象」フィールドの名前です。

表 26. STP モデリング操作で作成された新規フィールド (続き)

$STPLCI-<Target>	このフィールドには、LowerOfPredictionInterval 値 (確信度の下限値) が含まれます。 <Target> は、モデルの元の「対象」フィールドの名前です。
$STPUCI-<Target>	このフィールドには、UpperOfPredictionInterval 値 (確信度の上限値) が含まれます。 <Target> は、モデルの元の「対象」フィールドの名前です。

時空間予測モデルの設定

モデリング操作で受け入れ可能であると見なす不確実性のレベルを制御するには、「設定」タブを使用します。

不確実性因子 **(%)**: 不確実性因子とは、将来を予測する際の不確実性の増加を表すパーセント値です。将来に向けたステップごとに、このパーセンテージ分だけ、予測の不確実性の上限と下限が増加します。モデル出力に適用する不確実性因子を設定します。これにより、予測値の上限分割点と下限分割点が設定されます。

TCM ノード

このノードは、時間的因果モデル (TCM) を作成する場合に使用します。

時間的因果モデル

時間的因果モデリングでは、時系列データの重要な因果関係の発見が試みられます。時間的因果モデリングでは、対象系列のセットとそれらの対象への入力の候補のセットを指定します。そうすると、プロシージャにより、各対象の自己回帰時系列モデルが作成され、対象に対して因果関係を持つ入力のみが含まれます。このアプローチは、対象系列の予測値を明示的に指定する必要がある従来型の時系列モデリングとは異なります。通常、時間的因果モデリングでは、関連する複数の時系列のモデルを作成します。そのため、結果はモデル システム と呼ばれます。

時間的因果モデリングのコンテキストでは、因果 という用語は、グレンジャー因果関係を指します。X および Y の両方の過去の値の観点で Y について回帰した方が、Y の過去の値のみを回帰するよりもより良い Y のモデルが作成される場合、時系列 X は、「グレンジャー原因」の別の時系列 Y に伝えられます。

注**:** 時間的因果モデル作成ノードは、IBM SPSS Collaboration and Deployment Services 内のモデル評価ステップまたは Champion Challenger ステップをサポートしません。

例

ビジネスの意思決定者は、時間的因果モデリングを使用して、ビジネスを説明する一連の大規模な時間ベースのメトリック内の因果関係を発見できます。この分析により、重要業績評価指標に重大な影響を及ぼすいくつかの制御可能な入力が明らかになることがあります。

巨大な IT システムのマネージャーは、時間的因果モデリングを使用して、相互に関連する一連の大規模な操作メトリックの異常値を検査できます。したがって、因果モデルでは、異常値を検査するだけではなく、最も可能性の高い異常値の根本原因を発見できます。

フィールドの要件

少なくとも 1 つの対象が必要です。デフォルトでは、定義済み役割が「なし」のフィールドは使用されません。

データ構造

時間的因果モデリングでは、2 つのタイプのデータ構造がサポートされます。

列ベースのデータ
: 列ベースのデータの場合、各時系列フィールドには、単一の時系列のデータが含まれます。この構造は、時系列モデラーで使用される、従来型の時系列データの構造です。

多次元データ
: 多次元データの場合、各時系列フィールドには、複数の時系列のデータが含まれます。特定のフィールド内の個々の時系列は、*dimension* フィールドと呼ばれるカテゴリー型フィールドの一連の値によって識別されます。例えば、2 つの異なる販売チャネル (小売および Web) の販売データが単一の *sales* フィールドに格納されることがあります。「小売」および「Web」という値を持つ *channel* という名前の次元フィールドは、これらの 2 つの各販売チャネルに関連付けられたレコードを識別します。

注**:** 時間的因果モデルを作成するには、十分なデータ・ポイントが必要です。製品では、次の制約が使用されます。

```
m>(L + KL + 1)
```

ここで、m はデータ・ポイント数、L はラグ数、K は予測値の数を表します。データ・ポイント数 (m) が条件を満たすように、データ・セットのサイズが十分に大きいことを確認してください。

モデルに対する時系列

「フィールド」タブで「時系列」の設定を使用すると、モデル システムに含める系列を指定できます。

データに適用するデータ構造のオプションを選択します。多次元データの場合は、「次元の選択」をクリックして、次元フィールドを指定します。次元フィールドを指定した順序により、後続のすべてのダイアログおよび出力に次元フィールドが表示される順序が定義されます。次元フィールドを並べ替えるには、「次元の選択」サブダイアログの上下の矢印ボタンを使用します。

列ベースのデータの場合、系列 という語は、フィールド という語と同じ意味を持ちます。多次元データの場合、時系列を含むフィールドは、メトリック フィールドと呼ばれます。多次元データの時系列は、メトリック フィールドと各次元フィールドの値によって定義されます。列ベースのデータと多次元データの両方に以下の考慮事項があります。

- 入力の候補として、または対象と入力の両方として指定された系列は、各対象のモデルに含めるものとして考慮されます。各対象のモデルには、対象自身の遅れた値が必ず含まれます。
- 強制入力として指定された系列は、必ず各対象のモデルに含められます。
- 少なくとも一つの系列を対象、または対象と入力の両方として指定する必要があります。
- 「定義済みの役割を使用」を選択した場合、入力の役割を持つフィールドは、入力の候補として設定されます。定義済みの役割が強制入力にマップされることはありません。

多次元データ

多次元データの場合は、メトリック フィールドと関連役割をグリッドに指定します (グリッドの各行には、単一のメトリックと役割が指定されます)。デフォルトでは、モデル システムには、グリッドの各行の

次元フィールドのすべての組み合わせの系列が含まれます。例えば、*region* および *brand* の次元がある場合、デフォルトでは、メトリック *sales* を対象として指定することは、*region* および *brand* の組み合わせごとに別々の販売対象系列があることを意味します。

グリッドの行ごとに、次元の省略記号ボタンをクリックすることにより、任意の次元フィールドの値のセットをカスタマイズできます。このアクションにより、「次元の値の選択」サブダイアログが開きます。グリッドの行を追加、削除、またはコピーすることもできます。

「系列の数」列には、関連メトリックに現在指定されている次元の値のセットの数が表示されます。表示される値は、実際の系列の数 (セット当たり 1 つの系列) よりも大きい場合があります。この条件が発生するのは、指定した次元の値の組み合わせのいくつかが、関連メトリックに含まれる系列に対応していないときです。

次元の値の選択**:** 多次元データの場合、特定の役割を持つ特定のメトリック フィールドに適用する次元の値を指定して、分析をカスタマイズできます。例えば、*sales* がメトリック フィールドで、*channel* が値「retail」および「web」を持つ次元の場合は、「web」の販売を入力にし、「retail」の販売を対象にすることを指定できます。分析で使用されるすべてのメトリック フィールドに適用される次元のサブセットを指定することもできます。例えば、*region* が地理的領域を示す次元フィールドの場合は、分析を特定の領域に限定できます。

すべての値
: 現在の次元フィールドのすべての値を含めることを指定します。このオプションはデフォルトです。

含める値または除外する値の選択
: このオプションは、現在の次元フィールドの値のセットを指定する場合に使用します。「モード」に「含める」が選択されている場合は、「選択した値」リストに指定されている値のみが含まれます。「モード」に対して「除外」が選択されている場合は、「選択した値」リストに指定されている値以外のすべての値が含まれます。

  選択元の値のセットをフィルタリングすることが可能です。フィルター条件に一致した値は、「一致」タブに表示されます。フィルター条件に一致しない値は、「選択されていない値」リストの「不一致」タブに表示されます。「すべて」タブには、フィルター条件にかかわらず、選択されていない値がすべてリストされます。

  - フィルターを指定するときは、ワイルドカード文字を示すアスタリスク (*) を使用できます。
  - 現在のフィルターをクリアするには、「表示された値のフィルター」ダイアログで検索語に空の値を指定します。

観測

「フィールド」タブにある「観測」の設定を使用すると、観測を定義するためのフィールドを指定できます。

日付/時刻によって定義された観測
: 観測を日付、時刻、またはタイムスタンプのフィールドで定義することを指定できます。観測を定義するフィールドに加えて、観測を記述する適切な時間区分を選択します。指定した時間区分によっては、観測間の区分 (増分) や週当たりの日数などの他の設定も指定できます。時間区分には、以下の考慮事項があります。

  - 観測の時間区分が不規則のときは (例えば、販売注文の処理時)、「不規則」の値を使用します。「不規則」を選択したときは、「データ指定」タブの「時間区分」の設定で、分析に使用する時間区分を指定する必要があります。

- 観測が日付と時刻を表しており、時間区分が時、分または秒の場合は、「**1** 日あたりの時間数」、「**1** 日あたりの分数」、または「**1** 日あたりの秒数」を使用します。観測が日付を参照しない時刻 (期間) を表しており、時間区分が時、分、または秒の場合は、「時間数 **(**非周期的**)**」、「分数 **(**非周期的**)**」、または「秒数 **(**非周期的**)**」を使用します。
- 選択した時間区分に基づき、手続きで欠損観測値を検出できます。欠損観測値の検出が必要であるのは、すべての観測の時間区分が等しく、欠損観測値がないものと手続きで想定されているためです。例えば、時間区分が日であり、日付 2014-10-27 の後に 2014-10-29 がある場合は、2014-10-28 が欠損観測値です。欠損観測値がある場合は、値が代入されます。欠損値を処理するための設定は、「データ指定」タブから指定できます。
- 指定した時間区分により、手続きは、同じ時間区分内の集計が必要な複数の観測値を検出し、区分境界 (月の先頭など) の観測値を調整できます。これにより、観測値の区分が等しくなります。例えば、時間区分が月の場合は、同じ月の複数の日付が集計されます。このタイプの集計は、グループ化 と呼ばれます。デフォルトでは、観測値はグループ化のときに合計されます。別のグループ化の方法 (観測値の平均など) を指定するには、「データ指定」タブの「集計と分布」の設定を使用します。
- 時間区分によっては、追加設定により、通常の等間隔の区分の区切りを定義できます。例えば、時間区分が日であるが、平日のみが有効の場合は、1 週間の日数が 5 日で、週が月曜日に始まることを指定できます。

期間または循環する期間によって定義される観測

観測は、任意の数の循環レベルまでの期間または期間の繰り返しサイクルを表す 1 つ以上の整数フィールドによって定義できます。この構造では、標準時間区分の一つを適合させない観測の系列を記述できます。例えば、月数が 10 カ月のみの会計年度を、年を表す循環フィールドと月を表す期間フィールドで記述できます (1 つの循環の長さは 10)。

循環する期間を指定するフィールドにより、周期的レベルの階層が定義されます。ここで、最低レベルは「期間」フィールドで定義されます。次の最高のレベルは、レベルが 1 の循環フィールド、レベルが 2 の循環フィールド... という順で指定されます。各レベルのフィールド値は (最高レベルを除く)、次の最高のレベルに関して周期的でなければなりません。最高レベルの値は、周期的にできません。例えば、10 カ月の会計年度の場合、月は年内で周期的であり、年は周期的ではありません。

- 特定レベルの循環の長さが、次の最低レベルの周期性になります。会計年度の例では、循環レベルは 1 つのみあり、循環の長さは 10 でした。これは、次の最低レベルが月を表し、指定した会計年度の月数が 10 カ月であるからです。
- 1 から始まらないすべての周期的フィールドに対して、開始値を指定します。この設定は、欠損値を検出するために必要です。例えば、周期的フィールドが 2 から始まるが、開始値が 1 として指定されている場合、手続きでは、そのフィールドの各循環の最初の期間に欠損値があるものと想定されます。

分析の時間区分

分析に使用される時間区分は、観測の時間区分と異なる場合があります。例えば、観測の時間区分が日の場合に、分析の時間区分に月を選択することがあります。その場合は、データが日次データから月次データに集計されてからモデルが構築されます。データの分布を長い時間区分から短い時間区分にすることも選択できます。例えば、観測が四半期の場合、データを四半期から月次データに分布できます。

分析を実行する時間区分で使用できる選択肢は、観測の定義方法とそれらの観測の時間区分によって決まります。特に、循環する期間によって観測が定義されている場合は、集計のみがサポートされます。その場合、分析の時間区分は、観測の時間区分と等しいか、それより長い必要があります。

分析の時間間隔は、「データ仕様 (Data Specifications)」タブの「時間区分」設定から指定されます。データを集計または分布する方法は、「データ指定」タブの「集計と分布」の設定から指定します。

集計と分布

集計関数

分析に使用する時間区分が観測の時間区分よりも長い場合は、入力データが集計されます。例えば、観測の時間区分が日で、分析の時間区分が月のときは集計が実行されます。使用可能な集計関数は、平均、合計、モード、最小、または最大です。

分布関数

分析に使用する時間区分が観測の時間区分よりも短い場合は、入力データが分布します。例えば、観測の時間区分が四半期で、分析の時間区分が月のときは分布が実行されます。使用できる分布関数は、平均または合計です。

グループ化関数

観測が日付/時刻によって定義されており、複数の観測が同じ時間区分で実行される場合は、グループ化が適用されます。例えば、観測の時間区分が月の場合は、同じ月の複数の日付がグループ化され、観測が実行される月に関連付けられます。使用できるグループ化関数は、平均、合計、モード、最小、または最大です。観測が日付/時刻によって定義されており、観測の時間区分が不規則として指定されている場合は、グループ化が必ず実行されます。

注**:** グループ化は、集計の一つの形式ですが、欠損値を処理する前に実行されます。一方、公式の集計は、欠損値を処理した後に実行されます。観測の時間間隔が不規則と指定されている場合、集計はグループ化関数でのみ実行されます。

日をまたぐ観測を前日に集計

時間が日の境界をまたぐ観測を前日の値に集計するかどうかを指定します。例えば、1 日 8 時間の毎時観測を 20:00 に開始する場合、00:00 から 04:00 の間の観測を前日の集計結果に含めるかどうかをこの設定で指定します。この設定が適用されるのは、観測の時間区分が 1 日あたりの時間数、1 日あたりの分数、または 1 日あたりの秒数で、分析の時間区分が日の場合に限られます。

指定されたフィールドのカスタム設定

集計関数、分布関数、およびグループ化関数をフィールドごとに指定できます。この設定により、集計関数、分布関数、およびグループ化関数のデフォルト設定は上書きされます。

欠損値

入力データの欠損値は、代入値で置換されます。使用できる置換方法を以下に示します。

線形補間

線形補間を使用して欠損値を置換します。補間には、欠損値の前の最後の有効な値と、欠損値の後の最初の有効な値が使用されます。系列の最初または最後の観測に欠損値がある場合は、その系列の先頭または末尾にある最も近い 2 つの欠損値以外の値が使用されます。

系列平均

欠損値を系列全体の平均値に置換します。

周囲平均値

欠損値を、有効な周囲の値の平均値に置換します。隣接ポイントのスパンは、その平均値の計算に使用された欠損値の前後の有効値の数です。

周囲中央値

欠損値を、有効な周囲の値の中央値に置換します。隣接ポイントのスパンは、その中央値の計算に使用された欠損値の前後の有効値の数です。

線形トレンド
: このオプションは、系列の欠損観測値以外のすべての値を使用して、単純な線型モデルを適合させます。次にこのモデルは、欠損値を代入するために使用されます。

その他の設定:

欠損値の最大パーセンテージ**(%)**
: すべての系列に許可される欠損値の最大パーセンテージを指定します。指定した最大数よりも欠損値が多い系列は、分析から除外されます。

一般的なデータ オプション

次元フィールドあたりの異なる数値の最大数
: この設定は、多次元データに適用されるものであり、いずれか一つの次元フィールドに許可される異なる数値の最大数を指定します。デフォルトでは、この制限は 10000 に設定されていますが、必要に応じて大きな数値に増やすことができます。

一般的な作成オプション

信頼区間の幅 **(%)**
: この設定では、予測およびモデル・パラメーターの両方の信頼区間を制御します。100 未満の正の値を指定できます。デフォルトでは、95% の信頼区間が使用されます。

対象ごとの入力フィールドの最大数
: この設定では、各対象のモデルで許可される入力の最大数を指定します。1 から 20 までの範囲の整数を指定できます。各対象のモデルには、それ自身の遅れた値が必ず含まれます。そのため、この値を 1 に設定すると、その入力のみが対象自身になることが指定されます。

モデル許容度
: この設定では、各対象の最適な入力セットを判定するために使用される反復プロセスを制御します。ゼロよりも大きい任意の数を指定できます。デフォルトは 0.001 です。モデル許容度は、予測値の選択の停止基準です。最終モデルに組み込まれる予測値の数に影響する可能性があります。しかし、対象自体がうまく予測することができる場合は、その他の予測値が最終モデルに組み込まれない可能性があります。多少の試行錯誤が必要な場合があります (例えば、この値を大きい値に設定した場合、それよりも小さい値を設定してみることにより、その他の予測値を組み込むことが可能かどうかを判断できます)。

外れ値のしきい値**(%)**
: モデルから計算された外れ値の確率がこのしきい値を超えると、観測に外れ値を示すフラグが立てられます。50 から 100 までの範囲の値を指定できます。

各入力のラグ数 **(Number of Lags for Each Input)**
: この設定は、各対象のモデル内の各入力のラグ項目数を指定します。デフォルトでは、ラグ項目数は分析に使用される時間間隔から自動的に決定されます。例えば、時間間隔が (月単位の増分である) 月である場合のラグ数は 12 になります。オプションで、明示的にラグ数を指定することもできます。指定する値は、1 から 20 の範囲内の整数でなければなりません。

既存のモデルを使用して推定を続行
: 時間的因果モデルが既に生成されている場合、新規モデルを作成するのではなく、そのモデルに指定された基準設定を再利用するには、このオプションを選択します。この方法で、以前と同じモデル設定でも最新データを使用してそのモデルに基づいて新しい予測を再推定および作成できるので、時間の節約になります。

表示する系列

以下のオプションで、出力を表示する系列 (対象または入力) を指定します。指定した系列の出力の内容は、「出力オプション」の設定で決定されます。

最適合モデルに関連付けられた対象の表示

デフォルトでは、R2 乗の値によって決定される、上位 10 の最適合モデルに関連した対象の出力が表示されます。最適合モデルの別の固定数を指定することも、最適合モデルのパーセンテージを指定することもできます。また、以下の適合度から選択することもできます。

R 2 乗

線型モデルの適合度。決定係数とも呼びます。ターゲット変数の変動のうち、モデルによって説明される割合です。値の範囲は 0 から 1 までです。値が小さい場合は、モデルが十分にデータに適合していないことを示します。

平均平方根パーセント誤差

モデル予測値が系列の観測値とどれほど異なるかを示す指標。使用する単位に依存しないので、異なる単位の系列との比較に使用することができます。

平均 **2** 乗誤差平方根

平均平方誤差の平方根。モデルによって予測されるレベルから従属系列がどの程度外れているかを、従属系列と同じ単位を使用して表した指標。

BIC ベイズ情報量基準。-2 縮小対数尤度に基づいてモデルを選択し、比較するための指標。値が小さいほどモデルが良好であることを示します。BIC もパラメータが過多のモデル (例えば、大量の入力がある複雑なモデル) にペナルティを科しますが、AIC よりも厳密にそれを行います。

AIC 赤池情報量基準。-2 縮小対数尤度に基づいてモデルを選択し、比較するための指標。値が小さいほどモデルが良好であることを示します。AIC は、パラメータが過多のモデル (例えば、大量の入力がある複雑なモデル) にペナルティを科します。

個別の系列の指定

出力が必要な個々の系列を指定できます。

- 列ベースのデータの場合は、必要な系列を含むフィールドを指定します。指定したフィールドの順序によって、フィールドが出力に表示される順序が定義されます。
- 多次元データの場合は、系列を含むメトリック フィールドのグリッドにエントリーを追加することにより、特定の系列を指定します。次に、系列を定義する次元フィールドの値を指定します。
  - 各次元フィールドの値を直接グリッドに入力することも、使用可能な次元の値のリストから選択することもできます。使用可能な次元の値のリストから選択するには、必要な次元のセルで省略記号ボタンをクリックします。このアクションにより、「次元の値の選択」サブダイアログが開きます。
  - 次元の値のリストを検索するには、「次元の値の選択」サブダイアログで双眼鏡アイコンをクリックし、検索語を指定します。スペースは、検索語の一部として扱われます。検索語内のアスタリスク (*) は、ワイルドカード文字を示しません。
  - グリッド内の系列の順序により、系列が出力に表示される順序が定義されます、

列ベースのデータの場合も、多次元データの場合も、出力は 30 系列に制限されます。この制限には、指定した個々の系列 (入力または対象) と、最適合モデルに関連付けられた対象が含まれます。個々に指定した系列は、最適合モデルに関連付けられた対象よりも優先されます。

出力オプション

以下のオプションでは、出力の内容を指定します。「対象の出力」グループのオプションでは、「表示する系列」の設定で最適合モデルに関連付けられた対象の出力が生成されます。「系列の出力」グループのオプションでは、「表示する系列」の設定で指定された個々の系列の出力が生成されます。

全体のモデル システム

モデル システムの系列間の因果関係がグラフィカル表現で表示されます。表示する対象のモデル適合度統計および外れ値の両方のテーブルが出力項目の一部として含まれます。このオプションを「系列の出力」グループで選択した場合は、「表示する系列」の設定で指定された個々の系列ごとに別個の出力項目が作成されます。

系列間の因果関係には、関連する有意水準があります。ここで、有意水準が低いほど、関連の重要度が高くなります。指定した値より有意水準が高い関連は、非表示にすることを選択できます。

モデルの適合度の統計量と外れ値

表示用に選択した対象系列のモデル適合度統計および外れ値のテーブル。これらのテーブルには、「全体のモデル システム」視覚化のテーブルと同じ情報が含まれます。これらのテーブルでは、ピボットおよび編集を行うための標準機能がすべてサポートされています。

モデル効果とモデルのパラメータ

表示用に選択した対象系列のモデル効果検定とモデルのパラメーターのテーブル。モデル効果検定には、モデルに含まれる各入力の F 統計および関連する有意確率値が含まれます。

影響図

対象となる系列と、その影響を受ける他の系列またはそれに影響を与える他の系列との間の因果関係がグラフィカル表現で表示されます。対象となる系列に影響を与える系列は、原因 と呼ばれます。「効果」を選択すると、効果を表示するように初期化された影響図が生成されます。「原因」を選択すると、原因を表示するように初期化された影響図が生成されます。「原因と効果の両方」を選択すると、原因に初期化された影響図と効果に初期化された影響図の 2 つの別個の影響図が生成されます。影響図を表示する出力項目では、原因と効果をインタラクティブに切り替えることができます。

表示する原因または効果のレベル数を指定できます。ここで、第 1 レベルは、対象となる系列のみです。その他の各レベルは、対象となる系列のより間接的な原因または効果を示します。例えば、効果の表示の 3 番目のレベルは、2 番目のレベルの系列を直接入力として含む系列から構成されます。3 番目のレベルの系列は、対象となる系列から間接的に影響を受けます。これは、対象となる系列は 2 番目のレベルの系列に対する直接入力であるからです。

系列プロット

表示用に選択した対象系列の観測値および予測値のプロット。予測が要求された場合、プロットには、予測値と予測の信頼区間も表示されます。

残差プロット

表示用に選択した対象系列のモデル残差のプロット。

上位入力

表示する各対象 (長期間) と、対象の上位 3 件の入力のプロット。上位入力は、有意確率値が最も低い入力です。入力と対象で異なるスケールに対応するために、y 軸は、各系列の z スコアを表します。

予測テーブル

表示用に選択した対象系列の予測値とそれらの予測の信頼区間のテーブル。

外れ値の根本原因分析

対象となる系列の各外れ値の原因になる可能性が最も高い系列を判定します。「表示する系列」の設定の個々の系列のリストに含まれる対象系列ごとに、外れ値の根本原因分析が実行されます。

出力

対話式の外れ値のテーブルと図表

対象となる各系列の外れ値とそれらの外れ値の根本原因のテーブルおよび図表。このテーブルには、外れ値ごとに 1 つの行が含まれます。この図表は影響図です。テーブルで行を選択すると、影響図でパスが強調表示されます。このパスは、対象となる系列から、関連する外れ値が発生する可能性の最も高い系列までのパスです。

外れ値のピボット テーブル

対象となる各系列の外れ値とそれらの外れ値の根本原因のテーブル。このテーブルには、インタラクティブ表示のテーブルと同じ情報が含まれています。このテーブルでは、ピボットおよび編集を行うためのすべての標準機能がサポートされています。

因果関係レベル

根本原因の検索に含めるレベル数を指定できます。ここで使用されるレベルの概念は、影響図で説明するものと同じです。

すべてのモデル間でのモデルの適合度

すべてのモデルおよび選択した適合度統計に対するモデル適合度のヒストグラム。使用できる適合度統計を以下に示します。

R 2 乗

線型モデルの適合度。決定係数とも呼びます。ターゲット変数の変動のうち、モデルによって説明される割合です。値の範囲は 0 から 1 までです。値が小さい場合は、モデルが十分にデータに適合していないことを示します。

平均平方根パーセント誤差

モデル予測値が系列の観測値とどれほど異なるかを示す指標。使用する単位に依存しないので、異なる単位の系列との比較に使用することができます。

平均 **2** 乗誤差平方根

平均平方誤差の平方根。モデルによって予測されるレベルから従属系列がどの程度外れているかを、従属系列と同じ単位を使用して表した指標。

BIC ベイズ情報量基準。-2 縮小対数尤度に基づいてモデルを選択し、比較するための指標。値が小さいほどモデルが良好であることを示します。BIC もパラメータが過多のモデル (例えば、大量の入力がある複雑なモデル) にペナルティを科しますが、AIC よりも厳密にそれを行います。

AIC 赤池情報量基準。-2 縮小対数尤度に基づいてモデルを選択し、比較するための指標。値が小さいほどモデルが良好であることを示します。AIC は、パラメータが過多のモデル (例えば、大量の入力がある複雑なモデル) にペナルティを科します。

長期間の外れ値

推定期間の時間区分ごとのすべての対象に渡る外れ値の数の棒グラフ。

系列の変換

モデル システムの系列に適用されたすべての変換のテーブル。実行できる変換は、欠損値の代入、集計、および分布です。

推定期間

デフォルトでは、推定期間は、最も早い観測の時刻から始まり、すべての系列における最も遅い観測の時刻に終わります。

開始時刻と終了時刻で指定

推定期間の開始と終了の両方を指定することも、開始のみまたは終了のみを指定することもできます。推定期間の開始または終了を省略した場合は、デフォルト値が使用されます。

- 観測が「日付/時刻」フィールドによって定義されている場合は、「日付/時刻」フィールドに使用されているのと同じ形式で開始値と終了値を入力します。
- 循環する期間によって観測が定義されている場合は、それぞれの「循環する期間」フィールドの値を指定します。各フィールドは、別々の列に表示されます。

最も遅い時間区分または最も早い時間区分で指定

データの最も早い時間区分で開始または最も遅い時間区分で終了する時間区分の指定回数として推定期間を定義します。必要に応じてオフセットを指定することができます。このコンテキストでは、時間区分は、分析の時間区分を参照します。例えば、観測は月 1 回であるが、分析の時間区分は四半期であると想定します。「最新」を指定し、「時間区分の数 **(Number of time intervals)**」の値として 24 を指定することは、最新の 24 個の四半期を意味します。

必要な場合は、指定した数の時間区分を除外することもできます。例えば、最新の 24 個の時間区分を指定し、除外する数値として 1 を指定した場合、推定期間は、最新の区分の前の 24 個の区分から構成されます。

モデル・オプション

モデル名

モデルのカスタム名を指定することも、自動生成された名前 (*TCM*) を受け入れることもできます。

予測　「レコードの将来への拡張」のオプションで、推定期間の終わりを超えて予測する時間区分の数を設定します。この場合の時間区分は、「データ指定」タブで指定した分析の時間区分です。予測が要求されると、対象でもない任意の入力系列に対して、自己回帰モデルが自動的に構築されます。次にこれらのモデルは、予測期間のこれらの入力系列の値を生成するために使用されます。この設定の最大値制限はありません。

インタラクティブ出力

時間的因果モデリングからの出力には、多数のインタラクティブ出力オブジェクトが含まれます。インタラクティブ機能を使用するには、出力ビューアーでオブジェクトをアクティブ化 (ダブルクリック) します。

全体のモデル システム

モデル システムの系列間の因果関係が表示されます。同じ特定のターゲットをその入力に接続する線には、すべて同じ色が使用されます。線の太さは因果関係の有意性を示しており、線が太いほど有意性のより高い接続を表します。ターゲットでもない入力は黒い四角形で示されます。

- 上位のモデル、指定した系列、すべての系列、または入力のないモデルの関係を表示できます。上位のモデルは、「表示する系列」の設定で最適合モデルに指定された基準を満たすモデルです。
- 図表内の系列名を選択して右クリックし、コンテキスト メニューから「影響図の作成」を選択することで、1 つ以上の系列の影響図を生成できます。
- 指定した値より有意水準が高い因果関係は、非表示にすることを選択できます。有意水準が低いほど、因果関係の重要度が高くなります。

- 図表内の系列名を選択して右クリックし、次にコンテキスト メニューから「系列の関係を強調表示」を選択することで、特定の系列の関係を表示できます。

影響図
: 対象となる系列と、その影響を受ける他の系列またはそれに影響を与える他の系列との間の因果関係がグラフィカル表現で表示されます。対象となる系列に影響を与える系列は、原因 と呼ばれます。
  - 対象となる系列を変更するには、必要な系列の名前を指定します。影響図で任意のノードをダブルクリックすると、対象となる系列が、そのノードに関連付けられた系列に変更されます。
  - 原因と効果の表示は切り替え可能です。また、表示する原因または効果のレベル数は変更が可能です。
  - 任意のノードを 1 回クリックすると、ノードに関連付けられた系列の詳細な時系列図が開きます。

外れ値の根本原因分析
: 対象となる系列の各外れ値の原因になる可能性が最も高い系列を判定します。
  - 任意の外れ値の根本原因を表示するには、外れ値のテーブルで外れ値の行を選択します。根本原因は、時系列グラフの外れ値のアイコンをクリックして表示することもできます。
  - 任意のノードを 1 回クリックすると、ノードに関連付けられた系列の詳細な時系列図が開きます。

全体のモデル品質
: 特定の適合度統計におけるすべてのモデルのモデル適合度のヒストグラム。棒グラフの棒をクリックすると、ドット図がフィルタリングされ、選択した棒に関連付けられたモデルのみが表示されます。ドット図の特定の対象系列のモデルを探すには、系列の名前を指定します。

外れ値分布
: 推定期間の時間区分ごとのすべての対象に渡る外れ値の数の棒グラフ。棒グラフの棒をクリックすると、ドット図がフィルタリングされ、選択した棒に関連付けられた外れ値のみが表示されます。

TCM モデル・ナゲット

TCM モデリング操作で、次の表に示すように接頭辞 $TCM- が付いた多数の新規フィールドを作成します。

表 *27. TCM* モデリング操作で作成された新規フィールド

フィールド名	説明
$TCM-*colname*	各対象系列のモデルに予測された値。
$TCMLCI-*colname*	予測された各列の信頼区間の始めの値。
$TSUCI-*colname*	予測された各列の信頼区間の上限。
$TCMResidual-*colname*	生成されたモデル・データの各列の、ノイズの残差値。

TCM モデル・ナゲットの設定

「設定」タブには、TCM モデル・ナゲットに関する追加のオプションが用意されています。

予測

「レコードの将来への拡張」のオプションで、推定期間の終わりを超えて予測する時間区分の数を設定します。この場合の時間区分は、TCM ノードの「データ指定」タブで指定した分析の時間区分です。予測が要求されると、対象でもない任意の入力系列に対して、自己回帰モデルが自動的に構築されます。次にこれら

のモデルは、予測期間のこれらの入力系列の値を生成するために使用されます。

スコアリングで使用可能にする

スコアリングされる各モデルに新規フィールドの作成 **:** スコアリングされる各モデルに作成する新規フィールドを指定できるようにします。

- ノイズの残差**:** これを選択すると、各対象フィールドのモデル残差に対応する新規フィールド (デフォルトの接頭辞 $TCM- が付きます) が、これらの値の合計とともに作成されます。
- 確信度の上限および下限**:** これを選択すると、各対象フィールドの信頼区間の上限と下限にそれぞれ対応する新規フィールド (デフォルトの接頭辞 $TCM- が付きます) が、これらの値の合計とともに作成されます。

スコアリングに含める対象**:** モデル・スコアに含める使用可能な対象を選択します。

時間的因果モデルのシナリオ

「時間的因果モデルのシナリオ」手続きでは、アクティブなデータセットからのデータを使用して、時間的因果モデル システムのユーザー定義シナリオを実行します。シナリオ は、時系列 (ルート系列 と呼ばれる) と、指定した時刻範囲に渡るその系列のユーザー定義値のセットによって定義されます。指定した値は、ルート系列の影響を受ける時系列の予測を生成するために使用されます。この手続きでは、「時間的因果モデリング」手続きで作成されたモデル システム ファイルが必要です。アクティブなデータセットは、モデル システム ファイルの作成に使用されたのと同じデータであると想定されます。

例

「時間的因果モデリング」手続きを使用して、ビジネスの意思決定者が、多数の重要な業績評価指標に影響を与える主要なメトリックを発見しました。このメトリックは制御可能であるため、意思決定者は、次の四半期に渡るメトリックのさまざまな値のセットの影響を調査したいと考えています。この調査は簡単に行うことができ、モデル システム ファイルを「時間的因果モデルのシナリオ」手続きにロードし、主要なメトリックの値のセットを指定することにより行うことができます。

シナリオ期間の定義

シナリオ期間は、シナリオの実行に使用される値を指定する期間です。この期間は、推定期間の終わりの前または後に開始できます。必要な場合は、シナリオ期間の終わりを超えて予測することも指定できます。デフォルトでは、予測は、シナリオ期間の終わりから生成されます。すべてのシナリオでは、同じシナリオ期間と予測期間の指定が使用されます。

注**:** 予測は、シナリオ期間の開始後の最初の期間に開始されます。例えば、シナリオ期間が 2014-11-01 に開始され、時間間隔が月の場合、その最初の予測対象は 2014-12-01 になります。

開始時刻、終了時刻、および指定による予測の時刻で指定

- 観測が「日付/時刻」フィールドによって定義されている場合は、「日付/時刻」フィールドに使用されているのと同じ形式で開始、終了、および予測の値を入力します。「日付/時刻」フィールドの値は、関連する時間区分の先頭に調整されます。例えば、分析の時間区分が月の場合、値 10/10/2014 は、月の先頭である 10/01/2014 に調整されます。
- 循環する期間によって観測が定義されている場合は、それぞれの「循環する期間」フィールドの値を指定します。各フィールドは、別々の列に表示されます。

推定期間の終了からの相対的時間間隔で指定

推定期間の終了からの相対的時間間隔の数の観点で開始と終了を定義します。ここで、時間間隔は、分析の時間間隔です。推定期間の終わりは、時間間隔 0 として定義されます。推定期間の終わりより前の時間間隔は負の値で、推定期間の終わりより後の時間間隔は正の値になります。シナリオ期間の終わりを超えて予測する間隔の数を指定することもできます。デフォルトは 0 です。

例えば、分析の時間間隔が月であり、開始間隔に 1、終了間隔に 3、それを超えた予測期間に 1 を指定することを想定します。シナリオ期間は、推定期間の終わりに続く 3 カ月です。予測は、シナリオ期間の 2 および 3 カ月と、シナリオ期間の終わりを超えてさらに 1 カ月分、生成されます。

シナリオおよびシナリオ グループの追加

「シナリオ」タブでは、実行するシナリオを指定します。シナリオを定義するには、最初に「シナリオ期間の定義」をクリックしてシナリオ期間を定義します。シナリオおよびシナリオ グループ (多次元データにのみ適用) を作成するには、関連する「シナリオの追加」ボタンまたは「シナリオ グループの追加」ボタンをクリックします。関連するグリッドで特定のシナリオまたはシナリオ グループを選択すると、そのシナリオまたはシナリオ グループの編集、コピー、または削除を行うことができます。

列ベースのデータ

グリッドの「ルート フィールド」列では、値がシナリオ値で置換される時系列フィールドを指定します。「シナリオ値」列には、指定したシナリオ値が古いものから新しいものの順に表示されます。シナリオ値が式で定義された場合は、列にその式が表示されます。

多次元データ

個別のシナリオ

「個別のシナリオ」グリッドの各行では、指定したシナリオ値で値が置換される時系列を指定します。系列は、「ルート メトリック」列に指定したフィールドと、各次元フィールドに指定した値の組み合わせによって定義されます。「シナリオ値」列の内容は、列ベースのデータの場合と同じです。

シナリオ グループ

「シナリオ グループ」は、単一のルート メトリック フィールドの値と複数の次元の値のセットに基づく一連のシナリオを定義します。指定したメトリック フィールドの次元の値の各セット (次元フィールド当たり 1 つの値) によって時系列が定義されます。そのような時系列ごとに個々のシナリオが生成されます。これらの時系列の値は、シナリオ値で置換されます。シナリオ グループのシナリオ値は、式によって指定されます。この式は次に、グループ内の各時系列に適用されます。

「系列の数」列には、シナリオ グループに関連付けられた次元の値のセット数が表示されます。表示される値は、シナリオ グループに関連付けられた時系列の実際の数 (セット当たり 1 つの系列) よりも大きい場合があります。この条件が発生するのは、指定した次元の値の組み合わせのいくつかが、グループのルート メトリックに含まれる系列に対応していないときです。

シナリオ グループの例として、1 つのメトリック フィールド *advertising* と 2 つの次元フィールド *region* および *brand* を検討します。*region* と *brand* のすべての組み合わせを含む、ルート メトリックとしての *advertising* に基づくシナリオ グループを定義できます。*advertising* フィールドに関連付けられた各時系列について、*advertising* を 20 パーセント増やした影響を調査するには、`advertising*1.2` を式として指定します。*region* に 4 つの値、*brand* に 2 つの値がある場合は、そのような 8 つの時系列と、その結果 8 つのシナリオがグループで定義されます。

シナリオ定義: シナリオを定義するための設定は、データが列ベースであるのか、それとも多次元であるのかによって異なります。

ルート系列
: シナリオのルート系列を指定します。各シナリオは、単一のルート系列に基づきます。列ベースのデータの場合は、ルート系列を定義するフィールドを選択します。多次元データの場合は、系列を含むメトリック フィールドのグリッドにエントリーを追加することにより、ルート系列を指定します。次に、ルート系列を定義する次元フィールドの値を指定します。次元の値を指定するときは、以下のことが該当します。

- 各次元フィールドの値を直接グリッドに入力することも、使用可能な次元の値のリストから選択することもできます。使用可能な次元の値のリストから選択するには、必要な次元のセルで省略記号ボタンをクリックします。このアクションにより、「次元の値の選択」サブダイアログが開きます。
- 次元の値のリストを検索するには、「次元の値の選択」サブダイアログで双眼鏡アイコンをクリックし、検索語を指定します。スペースは、検索語の一部として扱われます。検索語内のアスタリスク (*) は、ワイルドカード文字を示しません。

影響を受ける対象の指定
: このオプションは、ルート系列の影響を受ける特定の対象が分かっており、それらの対象に対する影響を調査する場合にのみ使用します。デフォルトでは、ルート系列の影響を受ける対象が自動的に判定されます。「オプション」タブの設定を使用して、シナリオの影響を受ける系列の幅を指定できます。

  列ベースのデータの場合は、必要な対象を選択します。多次元データの場合は、系列を含む対象メトリック フィールドのグリッドにエントリーを追加することにより、対象系列を指定します。 デフォルトでは、指定したメトリック フィールド内のすべての系列が含まれます。含まれる系列のセットをカスタマイズするには、1 つ以上の次元フィールドに含まれる値をカスタマイズします。含まれる次元の値をカスタマイズするには、必要な次元の省略記号ボタンをクリックします。このアクションにより、「次元の値の選択」ダイアログが開きます。

  (多次元データの) 「系列の数」列には、関連する対象メトリックに現在指定されている次元の値のセットの数が表示されます。表示される値は、影響を受ける対象系列の実際の数 (セット当たり 1 つの系列) よりも大きい場合があります。この条件が発生するのは、指定した次元の値の組み合わせのいくつかが、関連する対象メトリックに含まれる系列に対応していないときです。

シナリオ **ID**
: 各シナリオには、固有の ID が必要です。ID は、シナリオに関連付けられた出力に表示されます。ID の値には、固有性以外の制限はありません。

ルート系列のシナリオ値の指定
: このオプションは、シナリオ期間のルート系列に明示的な値を指定する場合に使用します。グリッドにリストされる時間区分ごとに数値を指定する必要があります。シナリオ期間内の各間隔のルート系列 (実際または予測) の値を取得するには、「読み取り」、「予測」、または「読み取り/予測」をクリックします。

ルート系列のシナリオ値の式を指定
: シナリオ期間のルート系列の値を計算するための式を定義できます。式は直接入力することも、計算機ボタンをクリックしてシナリオ値の式ビルダーから作成することもできます。

- この式には、モデル システム内の任意のターゲットまたは入力を含めることができます。
- シナリオ期間が既存のデータを超えて拡張すると、式は、式のフィールドの予測値に適用されます。

- 多次元データの場合、フィールドで定義される時系列と、ルート メトリックに指定される次元の値を式の各フィールドで指定します。これは、式の評価に使用される時系列です。

  例として、ルート フィールドが *advertising*、式が `advertising*1.2` であると想定します。シナリオで使用される *advertising* の値は、既存の値の 20% 増加を表しています。

注**:** シナリオを作成するには、「シナリオ」タブで「シナリオの追加」をクリックしてください。

次元の値の選択*:* 多次元データの場合は、シナリオまたはシナリオ グループの影響が及ぶ対象を定義する次元の値をカスタマイズできます。また、シナリオ グループのルート系列のセットを定義する次元の値をカスタマイズすることもできます。

すべての値
: 現在の次元フィールドのすべての値を含めることを指定します。このオプションはデフォルトです。

値の選択
: このオプションは、現在の次元フィールドの値のセットを指定する場合に使用します。選択元の値のセットをフィルタリングすることが可能です。フィルター条件に一致した値は、「一致」タブに表示されます。フィルター条件に一致しない値は、「選択されていない値」リストの「不一致」タブに表示されます。「すべて」タブには、フィルター条件にかかわらず、選択されていない値がすべてリストされます。

  - フィルターを指定するときは、ワイルドカード文字を示すアスタリスク (*) を使用できます。
  - 現在のフィルターをクリアするには、「表示された値のフィルター」ダイアログで検索語に空の値を指定します。

影響を受ける対象の次元の値をカスタマイズするには、以下を実行します。

1. 「シナリオ定義」ダイアログまたは「シナリオ グループの定義」ダイアログで、次元の値をカスタマイズする対象メトリックを選択します。
2. カスタマイズする次元の列にある省略記号ボタンをクリックします。

シナリオ グループのルート系列の次元の値をカスタマイズするには、以下を実行します。

1. 「シナリオ グループの定義」ダイアログで、カスタマイズする次元の (ルート系列グリッドにある) 省略記号ボタンをクリックします。

シナリオ グループの定義**:**

ルート系列
: シナリオ グループのルート系列のセットを指定します。セットの時系列ごとに個々のシナリオが生成されます。ルート系列を指定するには、必要な系列を含むメトリック フィールドのグリッドにエントリーを追加します。次に、セットを定義する次元フィールドの値を指定します。デフォルトでは、指定したルート メトリック フィールド内のすべての系列が含まれます。含まれる系列のセットをカスタマイズするには、1 つ以上の次元フィールドに含まれる値をカスタマイズします。含まれる次元の値をカスタマイズするには、次元の省略記号ボタンをクリックします。このアクションにより、「次元の値の選択」ダイアログが開きます。

  「系列の数」列には、関連するルート メトリックに現在含まれている次元の値のセットの数が表示されます。表示される値は、シナリオ グループのルート系列の実際の数 (セット当たり 1 つの系列) よりも大きい場合があります。この条件が発生するのは、指定した次元の値の組み合わせのいくつかが、ルート メトリックに含まれる系列に対応していないときです。

影響を受ける対象系列の指定

このオプションは、ルート系列のセットの影響を受ける特定の対象が分かっており、それらの対象に対する影響を調査する場合にのみ使用します。デフォルトでは、各ルート系列の影響を受ける対象が自動的に判定されます。「オプション」タブの設定を使用して、個々のシナリオの影響を受ける系列の幅を指定できます。

対象系列を指定するには、系列を含むメトリック フィールドのグリッドにエントリーを追加します。 デフォルトでは、指定したメトリック フィールド内のすべての系列が含まれます。含まれる系列のセットをカスタマイズするには、1 つ以上の次元フィールドに含まれる値をカスタマイズします。含まれる次元の値をカスタマイズするには、必要な次元の省略記号ボタンをクリックします。このアクションにより、「次元の値の選択」ダイアログが開きます。

「系列の数」列には、関連する対象メトリックに現在指定されている次元の値のセットの数が表示されます。表示される値は、影響を受ける対象系列の実際の数 (セット当たり 1 つの系列) よりも大きい場合があります。この条件が発生するのは、指定した次元の値の組み合わせのいくつかが、関連する対象メトリックに含まれる系列に対応していないときです。

シナリオ **ID** 接頭辞

各シナリオ グループには、固有の接頭辞が必要です。接頭辞は、シナリオ グループの個々のシナリオに関連付けられている出力に表示される ID を構成するために使用されます。個々のシナリオの ID は、接頭辞の後に下線が続き、その後にルート系列を識別する各次元フィールドの値が続いたものです。次元の値は下線で区切られます。接頭辞の値には、固有性以外の制限はありません。

ルート系列のシナリオ値の式

シナリオ グループのシナリオ値は、式によって指定されます。この式は、次にグループの各ルート系列の値を計算するために使用されます。式は直接入力することも、計算機ボタンをクリックしてシナリオ値の式ビルダーから作成することもできます。

- この式には、モデル システム内の任意のターゲットまたは入力を含めることができます。
- シナリオ期間が既存のデータを超えて拡張すると、式は、式のフィールドの予測値に適用されます。
- グループのルート系列ごとに、式のフィールドで定義される時系列と、ルート系列を定義する次元の値を式のフィールドで指定します。これは、式の評価に使用される時系列です。例えば、ルート系列が `region='north'` および `brand='X'` で定義されている場合、式で使用される時系列は、これらの同じ次元の値で定義されます。

例として、ルート メトリック フィールドが *advertising* であり、2 つの次元フィールド *region* および *brand* が存在することを想定します。また、シナリオ グループに、次元フィールドの値のすべての組み合わせが含まれていることも想定します。*advertising* フィールドに関連付けられた各時系列について、*advertising* を 20 パーセント増やした影響を調査するには、`advertising*1.2` を式として指定します。

注**:** シナリオ グループは、多次元データにのみ適用されます。シナリオ グループを作成するには、「シナリオ」タブで「シナリオ グループの追加」をクリックしてください。

オプション

影響を受ける対象の最大レベル

影響を受ける対象のレベルの最大数を指定します。それぞれの連続したレベルには、ルート系列の影響をより間接的に受ける対象が最大 5 つ含まれます。具体的には、1 番目のレベルには、直接入力としてルート系列を持つ対象が含まれます。2 番目のレベルの対象には、直接入力として 1 番目のレベルの対象が含まれます。4 番目以降も同様です。この設定の値を増やすと、計算の複雑さが増し、パフォーマンスに影響が及ぶことがあります。

自動検出される対象の最大数
: ルート系列ごとに自動検出される影響を受ける対象の最大数を指定します。この設定の値を増やすと、計算の複雑さが増し、パフォーマンスに影響が及ぶことがあります。

影響図
: 各シナリオのルート系列とその影響を受ける対象系列間の因果関係がグラフィカル表現で表示されます。影響を受ける対象のシナリオ値と予測値の両方のテーブルが、出力項目の一部として含まれます。グラフには、影響を受ける対象の予測値のプロットが含まれます。影響図の任意のノードを 1 回クリックすると、ノードに関連付けられている系列の詳細な時系列図が開きます。シナリオごとに個別の影響図が生成されます。

系列図
: 各シナリオの影響を受ける対象ごとに予測値の系列プロットが生成されます。

予測とシナリオのテーブル
: 予測値および各シナリオのシナリオ値のテーブル。これらのテーブルには、影響図のテーブルと同じ情報が含まれます。これらのテーブルでは、ピボットおよび編集を行うための標準機能がすべてサポートされています。

プロットとテーブルに信頼区間を含める
: シナリオ予測の信頼区間をグラフおよびテーブル出力の両方に含めるかどうかを指定します。

信頼区間の幅 **(%)**
: この設定では、シナリオ予測の信頼区間を制御します。100 未満の正の値を指定できます。デフォルトでは、95% の信頼区間が使用されます。

時系列ノード

時系列ノードは、ローカル環境または分散環境のいずれかのデータと共に使用できます。分散環境では、IBM SPSS Analytic Server の機能を使用できます。このノードを使用して、時系列から指数平滑法、1 変量の自己回帰積分移動平均法 (ARIMA)、および多変量 ARIMA (または伝達関数) モデルを推測して構築し、その時系列データに基づいて予測を作成することを選択できます。

指数平滑法は、前の時系列の観測結果に重み付けされた値を使用して将来の値を予測する方法です。指数平滑法自体は、データの理論的解釈に基づいてはいません。一度に 1 つのポイントを予測し、新しいデータが投入されるごとに予測を調整します。この方法は、トレンド、季節性、またはその両方を示す時系列を予測する場合に役立ちます。トレンドと季節性の扱い方が異なる各種の指数平滑法モデルから選択することができます。

ARIMA モデルには、トレンドおよび季節性のコンポーネントのモデル作成に指数平滑法モデルよりも洗練された方法が用意されています。特に、モデル内に独立 (予測フィールド) 変数を含むことが可能になりました。これは、差異の程度と同様に自己回帰および移動平均の順序を明示して指定することと関連します。予測変数を含んでその変数のいくつかまたはすべてに伝達関数を定義し、外れ値または明示した外れ値セットの自動検出を指定できます。

注**:** 実際面では、郵送するカタログの数または会社の Web ページのヒット数など、予測対象の一連の性質を説明する上で役立つ可能性がある予測値を含める場合は、ARIMA モデルが最も有用です。指数平滑法モデルは、性質や傾向の理由を理解しようとしないで、時系列の性質や傾向を記述します。例えば、歴史的に 12 カ月ごとにピークが来る時系列は、その理由がわからなくても、おそらくそのようにその性質が継続します。

また、1 つ以上の目標変数に対して最も適合する ARIMA または指数平滑化モデルを自動的に特定し推定しようとするエキスパート・モデラーも利用できるため、試行錯誤しながら適切なモデルを特定する必要がなくなります。確信が持てない場合は、エキスパート・モデラー・オプションを使用してください。

予測変数が指定された場合、依存する時系列と統計的に顕著な関係があるこれらの変数を ARIMA モデル内に含めるために、エキスパート・モデラーはそのような変数を選択します。必要に応じて差分を取ることによって、また必要に応じて平方根変換または自然対数変換を使用して、モデル変数は変換されます。デフォルトでは、エキスパート・モデラーがすべての指数平滑化モデルと ARIMA モデルを考慮し、各対象フィールドに最も適したモデルを選択します。ただし、最適な指数平滑化モデルのみを取り上げるか、ARIMA モデルのみを取り上げるか、エキスパート・モデラーを制限することができます。また、外れ値の自動検出も指定できます。

時系列ノード - フィールド・オプション

「フィールド」タブで、上流のノードですでに定義されているフィールドの役割設定を使用するか、手動でフィールドの割り当てを行うかを選択します。

定義済みの役割を使用: このオプションを選択すると、上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) が使用されます。

カスタム・フィールド割り当ての使用: 対象、予測、およびその他の役割を手動で割り当てるには、このオプションを選択します。

フィールド: 矢印ボタンを使用して、このリストの項目を画面右側のさまざまな役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。

リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

対象: 1 つ以上のフィールドを予測の対象として選択します。

入力の候補: 1 つ以上のフィールドを予測の入力として選択します。

イベントおよび干渉: 特定の入力フィールドをイベント・フィールドまたは干渉フィールドとして指定する場合に、この領域を使用します。この指定により、イベント (販売促進などの予測可能な繰り返し発生する状況) または干渉 (停電や従業員のストライキなど、一時的な出来事) に影響を受ける時系列データを含むものとして、フィールドが識別されます。選択するフィールドは、整数のストレージを持つフラグでなければなりません。

時系列ノード - データ指定オプション

「データ指定」タブで、データをモデルに含めるためのすべてのオプションを設定します。「日付/時刻フィールド」と「時間区分」の両方を指定すると、「実行」 ボタンをクリックして、すべてデフォルト・オプションを含むモデルを構築できます。ただし、通常は、それぞれの目的でビルドをカスタマイズする必要があります。

このタブに含まれる数種類のペインを使用して、モデルに固有のカスタマイズを設定します。

時系列ノード - 観測

このペインの設定を使用して、観測を定義するフィールドを指定します。

日付/時刻フィールドによって指定される観測

観測を日付、時刻、またはタイムスタンプのフィールドによって定義することを指定できます。観測を定義するフィールドに加えて、観測を記述する適切な時間区分を選択します。指定した時間区分によっては、観測間の区分 (増分) や週当たりの日数などの他の設定も指定できます。時間区分には、以下の考慮事項があります。

- 観測の時間区分が不規則のときは (例えば、販売注文の処理時)、「不規則」の値を使用します。「不規則」を選択したときは、「データ指定」タブの「時間区分」の設定で、分析に使用する時間区分を指定する必要があります。
- 観測が日付と時刻を表しており、時間区分が時、分または秒の場合は、「**1** 日あたりの時間数」、「**1** 日あたりの分数」、または「**1** 日あたりの秒数」を使用します。観測が日付を参照しない時刻 (期間) を表しており、時間区分が時、分、または秒の場合は、「時間数 **(**非周期的**)**」、「分数 **(**非周期的**)**」、または「秒数 **(**非周期的**)**」を使用します。
- 選択した時間区分に基づき、手続きで欠損観測値を検出できます。欠損観測値の検出が必要であるのは、すべての観測の時間区分が等しく、欠損観測値がないものと手続きで想定されているためです。例えば、時間区分が日であり、日付 2015-10-27 の後に 2015-10-29 がある場合は、2015-10-28 が欠損観測値です。欠損観測値に対して値が代入されます。「データ指定」タブの「欠損値の処理」エリアを使用して、欠損値を処理するための設定を指定してください。
- 指定した時間区分により、手続きは、同じ時間区分内の集計が必要な複数の観測値を検出し、区分境界 (月の先頭など) の観測値を調整できます。これにより、観測値の区分が等しくなります。例えば、時間区分が月の場合は、同じ月の複数の日付が集計されます。このタイプの集計は、グループ化 と呼ばれます。デフォルトでは、観測値はグループ化のときに合計されます。別のグループ化の方法 (観測値の平均など) を指定するには、「データ指定」タブの「集計と分布」の設定を使用します。
- 時間区分によっては、追加設定により、通常の等間隔の区分の区切りを定義できます。例えば、時間区分が日であるが、平日のみが有効の場合は、1 週間の日数が 5 日で、週が月曜日に始まることを指定できます。

期間または循環する期間として定義される観測

観測は、任意の数の循環レベルまでの期間または期間の繰り返しサイクルを表す 1 つ以上の整数フィールドによって定義できます。この構造では、標準の時間区分に適合しない観測の系列を記述できます。例えば、月数が 10 カ月のみの会計年度を、年を表す循環フィールドと月を表す期間フィールドで記述できます (1 つの循環の長さは 10)。

循環する期間を指定するフィールドにより、周期的レベルの階層が定義されます。ここで、最低レベルは「期間」フィールドで定義されます。次の最高のレベルは、レベルが 1 の循環フィールド、レベルが 2 の循環フィールド... という順で指定されます。各レベルのフィールド値は (最高レベルを除く)、次の最高のレベルに関して周期的でなければなりません。最高レベルの値は、周期的にできません。例えば、10 カ月の会計年度の場合、月は年内で周期的であり、年は周期的ではありません。

- 特定レベルの循環の長さが、次の最低レベルの周期性になります。会計年度の例では、循環レベルは 1 つのみあり、循環の長さは 10 でした。これは、次の最低レベルが月を表し、指定した会計年度の月数が 10 カ月であるからです。
- 1 から始まらないすべての周期的フィールドに対して、開始値を指定します。この設定は、欠損値を検出するために必要です。例えば、周期的フィールドが 2 から始まるが、開始値が 1 として指定されている場合、手続きでは、そのフィールドの各循環の最初の期間に欠損値があるものと想定されます。

時系列ノード - 分析の時間区分

分析で使用する時間区分は、観測の時間区分とは異なる場合があります。例えば、観測の時間区分が日の場合に、分析の時間区分に月を選択することがあります。その場合は、データが日次データから月次データに集計されてからモデルが構築されます。データの分布を長い時間区分から短い時間区分にすることも選択できます。例えば、観測が四半期の場合、データを四半期から月次データに分布できます。

このペインの設定を使用して、分析の時間区分を指定します。データを集計または分布する方法は、「データ指定」タブの「集計と分布」の設定から指定します。

分析を実行する時間区分で使用できる選択肢は、観測の定義方法とそれらの観測の時間区分によって決まります。特に、循環する期間によって観測が定義されている場合は、集計のみがサポートされます。その場合、分析の時間区分は、観測の時間区分と等しいか、それより長い必要があります。

時系列ノード - 集計オプションおよび分布オプション

このペインの設定を使用して、観測の時間区分に関する入力データの集計または分布を行うための設定を指定します。

集計関数
: 分析に使用する時間区分が観測の時間区分よりも長い場合は、入力データが集計されます。例えば、観測の時間区分が日で、分析の時間区分が月のときは集計が実行されます。使用可能な集計関数は、平均、合計、モード、最小、または最大です。

分布関数
: 分析に使用する時間区分が観測の時間区分よりも短い場合は、入力データが分布します。例えば、観測の時間区分が四半期で、分析の時間区分が月のときは分布が実行されます。使用できる分布関数は、平均または合計です。

グループ化関数
: 観測が日付/時刻によって定義されており、複数の観測が同じ時間区分で実行される場合は、グループ化が適用されます。例えば、観測の時間区分が月の場合は、同じ月の複数の日付がグループ化され、観測が実行される月に関連付けられます。使用できるグループ化関数は、平均、合計、モード、最小、または最大です。観測が日付/時刻によって定義されており、観測の時間区分が不規則として指定されている場合は、グループ化が必ず実行されます。

  注: グループ化は、集計の一つの形式ですが、欠損値を処理する前に実行されます。一方、公式の集計は、欠損値を処理した後に実行されます。観測の時間間隔が不規則と指定されている場合、集計はグループ化関数でのみ実行されます。

日をまたぐ観測を前日に集計
: 時間が日の境界をまたぐ観測を前日の値に集計するかどうかを指定します。例えば、1 日 8 時間の毎時観測を 20:00 に開始する場合、00:00 から 04:00 の間の観測を前日の集計結果に含めるかどうかをこの設定で指定します。この設定が適用されるのは、観測の時間区分が 1 日あたりの時間数、1 日あたりの分数、または 1 日あたりの秒数で、分析の時間区分が日の場合に限られます。

指定されたフィールドのカスタム設定
: 集計関数、分布関数、およびグループ化関数をフィールドごとに指定できます。この設定により、集計関数、分布関数、およびグループ化関数のデフォルト設定は上書きされます。

時系列ノード - 欠損値オプション

このペインの設定を使用して、入力データの欠損値を代入値で置換する方法を指定します。使用できる置換方法を以下に示します。

線形補間
: 線形補間を使用して欠損値を置換します。補間には、欠損値の前の最後の有効な値と、欠損値の後の最初の有効な値が使用されます。系列の最初または最後の観測に欠損値がある場合は、その系列の先頭または末尾にある最も近い 2 つの欠損値以外の値が使用されます。

系列平均
: 欠損値を系列全体の平均値に置換します。

周囲平均値
: 欠損値を、有効な周囲の値の平均値に置換します。隣接ポイントのスパンは、その平均値の計算に使用された欠損値の前後の有効値の数です。

周囲中央値
: 欠損値を、有効な周囲の値の中央値に置換します。隣接ポイントのスパンは、その中央値の計算に使用された欠損値の前後の有効値の数です。

線形トレンド
: このオプションは、系列の欠損観測値以外のすべての値を使用して、単純な線型モデルを適合させます。次にこのモデルは、欠損値を代入するために使用されます。

その他の設定:

最小データ品質スコア **(%)**
: 各時系列に対応する時刻変数および入力データのデータ品質指標を計算します。データ品質スコアがこのしきい値よりも低い場合、対応する時系列は破棄されます。

時系列ノード - 推定期間

「推定期間」ペインで、モデルの推定に使用するレコードの範囲を指定することができます。デフォルトでは、推定期間は、最も早い観測の時刻から始まり、すべての系列における最も遅い観測の時刻に終わります。

開始時刻と終了時刻で指定
: 推定期間の開始と終了の両方を指定することも、開始のみまたは終了のみを指定することもできます。推定期間の開始または終了を省略した場合は、デフォルト値が使用されます。

- 観測が「日付/時刻」フィールドによって定義されている場合は、「日付/時刻」フィールドに使用されているのと同じ形式で開始値と終了値を入力します。
- 循環する期間によって観測が定義されている場合は、それぞれの「循環する期間」フィールドの値を指定します。各フィールドは、別々の列に表示されます。

最も遅い時間区分または最も早い時間区分で指定
: データの最も早い時間区分で開始または最も遅い時間区分で終了する時間区分の指定回数として推定期間を定義します。必要に応じてオフセットを指定することができます。このコンテキストでは、時間区分は、分析の時間区分を指します。例えば、観測は月 1 回であるが、分析の時間区分は四半期であると想定します。「最新」を指定し、「時間区分の数 **(Number of time intervals)**」の値として 24 を指定することは、最新の 24 個の四半期を意味します。

  必要な場合は、指定した数の時間区分を除外することもできます。例えば、最新の 24 個の時間区分を指定し、除外する数値として 1 を指定した場合、推定期間は、最新の区分の前の 24 個の区分から構成されます。

時系列ノード - 作成オプション

「作成オプション」タブで、モデルを作成するためのすべてのオプションを設定します。もちろん、「実行」 ボタンをクリックすると、すべてデフォルト・オプションのモデルが構築されますが、通常は、それぞれの目的で作成をカスタマイズする必要があります。

このタブには、2 つの異なるペインが含まれています。これらのペインで、ご使用のモデルに固有のカスタマイズを設定します。

時系列ノード - 一般的な作成オプション

このペインで使用可能なオプションは、「方法」リストから以下の 3 つの設定のうち、どれを選択したかによって異なります。

- エキスパート・モデラー **:** エキスパート・モデラーを使用するには、このオプションを選択します。これで、独立した各時系列に最も適合するモデルが自動的に検出されます。
- 指数平滑法。このオプションは、カスタム指数平滑法モデルを指定する場合に使用します。
- **ARIMA**。このオプションは、カスタム ARIMA モデルを指定する場合に使用します。

エキスパート モデラー

「モデル タイプ」で、作成するモデルのタイプを以下から選択します。

- 「すべてのモデル」。エキスパート・モデラーは、ARIMA と指数平滑法モデルの両方を考慮します。
- 指数平滑法モデルのみ。エキスパート・モデラーは、指数平滑法モデルのみを考慮します。
- 「**ARIMA** モデルのみ」。エキスパート・モデラーは ARIMA モデルのみを考慮します。

エキスパート・モデラーが季節モデルを考慮する。このオプションは、アクティブなデータ・セットに周期性が定義されている場合にのみ有効です。このオプションがオンの場合、エキスパート・モデラーは季節性および非季節性の両方のモデルを検討します。このオプションを選択しないと、エキスパート・モデラーは非季節モデルのみを考慮します。

エキスパート モデラーが指数平滑化モデルを考慮。 このオプションを選択した場合、エキスパート モデラーは合計 13 個の指数平滑法モデル (このうちの 7 個は元の時系列ノードに存在していたもので、6 個はバージョン 18.1 で追加されたもの) を検索します。このオプションを選択しないと、エキスパート・モデラーは元の 7 個の指数平滑化モデルのみを検索します。

「外れ値」で、以下のオプションから選択します。

「自動的に外れ値を検出」。デフォルトでは、外れ値の自動検出は実行されません。外れ値の自動検出を実行するには、このオプションをオンにしてから、希望する外れ値のタイプを選択します。

入力フィールドは、測定の尺度がフラグ型、名義型、または順序型である必要があり、このリストに含まれる前に数値 (フラグ型フィールドの場合は True/Falseでなく 1/0) になっていなければなりません。

エキスパート・モデラーは単純な回帰のみを検討し、「フィールド」タブのイベント・フィールドまたは干渉フィールドとして識別される入力フィールドの任意の伝達関数を考慮しません。

指数平滑化

モデル・タイプ**:** 指数平滑化モデルは、季節性または非季節性のどちらかに分類されます。[1] 季節性モデルは、「データ指定」タブの「時間区分」ペインを使用して定義される周期性が季節性である場合にのみ使用できます。季節的な周期性には、循環する期間、年数、四半期数、月数、曜日数、1 日あたりの時間数、1 日あたりの分数、1 日あたりの秒数があります。 選択可能なモデル タイプは、以下のとおりです。

- 単純: このモデルは、トレンドまたは季節性のない時系列に適しています。関連する平滑化パラメーターは水準のみです。単純指数平滑化は、0 次の自己回帰、1 次の差分、1 次の移動平均、および定数なしの ARIMA に最もよく似ています。
- 「**Holt** の線型トレンド」。このモデルは、線型トレンドがあり、季節性がない系列に適しています。関連する平滑化パラメーターは水準パラメーターとトレンド・パラメーターであり、このモデル内では、互いの値に制約を受けません。Holt のモデルは Brown のモデルよりも一般的ですが、大きな系列の推定値の計算には余計に時間がかかる場合があります。Hplt の指数平滑化は、0 次の自己回帰、2 次の差分、移動平均が 2 次の ARIMA に最もよく似ています。
- 「減衰トレンド」。このモデルは、減衰する線型トレンドがあり、季節性のない時系列に適しています。関連する平滑化パラメーターは、水準、トレンド、および減衰トレンドです。減衰指数平滑化は、1 次の自己回帰、1 次の差分、および 2 次の移動平均の ARIMA に最もよく似ています。
- 「乗法トレンド」。このモデルは、系列の水準に依存するトレンドがあり、季節性のない系列に適しています。関連する平滑化パラメーターは、水準パラメーターとトレンド・パラメーターです。「乗法トレンド」の指数平滑法は、いかなる ARIMA モデルにも類似しません。
- 「**Brown** の線型トレンド」。このモデルは、線型トレンドがあり、季節性がない系列に適しています。関連する平滑化パラメーターは水準パラメーターとトレンド・パラメーターで、モデル内では等しいと見なされます。したがって、Brown モデルは Holt モデルの特別なケースです。Brown の指数平滑化は、ARIMA に最もよく似ています。0 次の自己回帰、2 次の差異、および 2 次の移動平均があり、移動平均の 2 次目の係数が一次の二乗の係数の 1/2 です。
- 「単純な季節性」。このモデルは、トレンドがなく常に一定の季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準パラメーターと季節パラメーターです。季節指数平滑化は、0 次の自己回帰、1 次の差分、1 次の季節差分、および 1、*p*、および移動平均が *p*+1 の ARIMA に最もよく似ています。この *p* は、季節区間 (季節的な間隔) の周期数です。月次データの場合、*p* = 12 です。
- 「**Winters** 加法」。このモデルは、線型トレンドおよび常に一定の季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、および季節です。Winters の加法指数平滑化は、0 次の自己回帰、1 次の差分、1 次の季節差分、および移動平均が *p*+1 の ARIMA に最もよく似ています。この *p* は、季節区分 (季節的な間隔) の周期数です。月次データの場合、*p* = 12 です。
- 「減衰トレンドと加法的季節性」。このモデルは、減衰する線型トレンドおよび常に一定の季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、減衰トレンド、および季節です。「減衰トレンドと加法的季節性」の指数平滑法は、いかなる ARIMA モデルにも類似しません。
- 「乗法トレンドと加法的季節性」。このモデルは、系列の水準に依存するトレンドおよび常に一定の季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、および季節です。「乗法トレンドと加法的季節性」の指数平滑法は、いかなる ARIMA モデルにも類似しません。
- 「乗法的季節性」。このモデルは、トレンドがなく系列の水準に依存する季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準パラメーターと季節パラメーターです。「乗法的季節性」の指数平滑法は、いかなる ARIMA モデルにも類似しません。

---

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

- 「**Winters** 乗法」。このモデルは、線型トレンドおよび系列の水準に依存する季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、および季節です。Winters の相乗指数平滑法は、いかなる ARIMA モデルにも類似しません。
- 「減衰トレンドと乗法的季節性」。このモデルは、減衰する線型トレンドおよび系列の水準に依存する季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、減衰トレンド、および季節です。「減衰トレンドと乗法的季節性」の指数平滑法は、いかなる ARIMA モデルにも類似しません。
- 「乗法トレンドと乗法的季節性」。このモデルは、系列の水準に依存するトレンドおよび季節的効果がある系列に適しています。関連する平滑化パラメーターは、水準、トレンド、および季節です。「乗法トレンドと乗法的季節性」の指数平滑法は、いかなる ARIMA モデルにも類似しません。

「対象の変換」。各従属変数に、モデル化される前に実行される変換を指定できます。

- なし**:** 変換は実行されません。
- 平方根: 平方根変換が実行されます。
- 自然対数: 自然対数変換が実行されます。

ARIMA

カスタム ARIMA モデルの構造を指定します。

「**ARIMA** の順序」。モデルのさまざまな ARIMA 成分の値を、グリッドの対応するセルに入力します。すべての値は負でない整数にする必要があります。自己回帰と移動平均の成分については、値は最大次数を表します。すべての正の低い次数はモデルに含まれます。例えば、2 を指定すると、モデルには次数 2 と 1 が含まれます。「季節性」列のセルは、周期性がアクティブ データ セットに定義されている場合にのみ有効です。

- 「自己回帰 **(p)**」。モデル内の自己回帰の次数の数値です。自己回帰の次数は、現在の値を予測するために使用される系列からの前 (過去) の値を指定します。例えば、自己回帰の次数 2 は、現在の値を予測するために系列の値を過去の 2 期間使用するように指定します。
- 差分 **(d)**。モデルを推定する前に系列に適用される差分の次数を指定します。トレンドが存在する場合は差分を取る必要があり (トレンドの存在する系列は通常非定常性であり、ARIMA モデルは定常性を前提としている)、その効果を取り除くために行います。差分の次数は、系列のトレンドの次数に対応しています (1 次差分は線型トレンドを表し、2 次差分は 2 次トレンドを表す、など)。
- 「移動平均 **(q)**」。モデル内の移動平均の次数の数値。移動平均の次数は、過去の値の系列平均の偏差が、現在の値を予測するためにどのように使用されるかを指定します。例えば、移動平均の次数 1 および 2 は、系列の現在の値を予測する際に最近の 2 期間のそれぞれから取得した系列の平均値の偏差を考慮することを指定します。

季節性。季節型の自己回帰、移動平均、および差分成分は、対応する非季節の成分と同様の役割を果たします。ただし、季節次数については、現在の系列値は、1 つ以上の季節期間で区切られた過去の系列値に影響されます。例えば、毎月のデータ (季節期間 12) については、季節次数 1 は、現在の系列値は現在の期間より 12 期間以前の系列値により影響されることを意味しています。毎月のデータについて、季節次数 1 は、非季節次数 12 を指定するのと同じことになります。

「自動的に外れ値を検出」。外れ値の自動検出を実行するには、このオプションをオンにしてから、使用可能な外れ値のタイプを 1 つ以上選択します。

検出する外れ値の型。検出する外れ値の型を選択します。サポートされるタイプは、次のとおりです。

- 相加的 (デフォルト)

- レベル・シフト (デフォルト)
- 技術革新的
- 過渡
- 季節性相加
- 局所トレンド
- 相加的パッチ

転送関数の順序と変換。ARIMA モデル内の任意またはすべての入力フィールドに対する変換の指定および伝達関数の定義を行うには、「設定」をクリックします。別のダイアログ ボックスが表示され、そこで転送と変換の詳細を入力します。

「モデル内に定数項を含める」。系列値の全体平均が 0 だという確信がない限り、通常は定数を含めます。差分を適用する場合は、定数を除外することをお勧めします。

詳細情報

- 外れ値のタイプについて詳しくは、 308 ページの『外れ値』を参照してください。
- 伝達関数および変換関数について詳しくは、『伝達関数および変換関数』を参照してください。

伝達関数および変換関数**:** 伝達関数の次数および「変換」ダイアログ ボックスを使用して、ARIMA モデル内の任意またはすべての入力フィールドに対する変換の指定および伝達関数の定義を行います。

対象の変換。このペインでは、各目標変数に、モデル化される前に実行される変換を指定できます。

- なし**:** 変換は実行されません。
- 平方根: 平方根変換が実行されます。
- 自然対数: 自然対数変換が実行されます。

転送の関数と変換の入力の候補。伝達関数を使用して、入力フィールドの過去の値が対象系列の将来の値を予測するために使用される方法を指定します。ペインの左側のリストには、すべての入力フィールドが表示されます。このペインのその他の情報は、選択した入力フィールドに固有のものです。

「伝達関数の次数」。伝達関数のさまざまな成分の値を、「構造」グリッドの対応するセルに入力します。すべての値は負でない整数にする必要があります。分子と分母の成分については、値は最大次数を表します。すべての正の低い次数はモデルに含まれます。さらに、次数 0 は常に分子成分に含まれます。例えば、分子に 2 を指定すると、モデルには次数 2、1 および 0 が含まれます。分母に 3 を指定するとモデルには次数 3、2、および 1 が含まれます。「季節性」列のセルは、周期性がアクティブ データ セットに定義されている場合にのみ有効です。

「分子」。伝達関数の分子の次数で、従属系列の現在の値を予測するために使用される選択した独立 (予測値) 系列からの前 (過去) の値を指定します。例えば、分子次数 1 は、独立系列の現在の値だけでなく、過去の 1 期間における独立系列の値が各従属系列の現在の値を予測するために使用されることを指定します。

「分母」。伝達関数の分母の次数で、従属系列の現在の値を予測するために、選択された独立 (予測値) 系列の前 (過去) の値に対して系列の平均からどのくらいの偏差が使用されるかを指定します。例えば、分母次数 1 は、各従属系列の現在の値を予測する際に、過去の 1 期間における独立系列の平均値の偏差が考慮されることを指定します。

「差分」。モデルを推定する前に、選択された独立 (予測) 系列に適用される差分の次数を指定します。トレンドが存在する場合は差分を取る必要があり、トレンドの効果を取り除くために差分を使用します。

季節性。季節分子、分母、および差分成分は、対応する非季節の成分と同様の役割を果たします。ただし、季節次数については、現在の系列値は、1 つ以上の季節期間で区切られた過去の系列値に影響されます。例えば、毎月のデータ (季節期間 12) については、季節次数 1 は、現在の系列値は現在の期間より 12 期間以前の系列値により影響されることを意味しています。毎月のデータについて、季節次数 1 は、非季節次数 12 を指定するのと同じことになります。

「遅延」。遅延を設定すると、指定された間隔数だけ、入力フィールドの影響が遅延させられます。例えば遅延が 5 に設定された場合、時間 *t* での入力フィールドの値は、5 期間が経過するまで (*t* + 5) 予測に影響しません。

変換。独立変数のセットに対する伝達関数の仕様にも、そのような変数に実行されるオプションの変換が含まれます。

- なし**:** 変換は実行されません。
- 平方根: 平方根変換が実行されます。
- 自然対数: 自然対数変換が実行されます。

時系列ノード - 作成出力オプション

ACF および **PACF** 出力内の最大ラグ数 **:** 自己相関 (ACF) および 偏自己相関 (PACF) は、現在と過去の時系列値の関連性の測定で、どの過去の時系列値が将来値の予測に最も役立つかを示します。自己相関および偏自己相関のテーブルおよびプロットに表示されるラグの最大数を設定できます。

予測変数の重要度を計算**:** 重要度の適切な測定基準を作成するモデルの場合、モデル推定時に各予測値の相対重要度を示すグラフを表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない予測値を削除または無視したいと考えます。モデルによっては、特に大きなデータ セットを使用する場合、予測変数の重要度の計算に時間がかかることがあります。そのため、一部のモデルではデフォルトでオフになっています。

時系列ノード - モデル・オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

信頼限界幅 **(%)**: 信頼区間は、モデルの予測と残差自己相関に対して計算されます。100 未満の正の値を指定できます。デフォルトでは、95% の信頼区間が使用されます。

既存のモデルを使用して推定を続行: 時系列モデルが既に生成されている場合は、新しいモデルを始めから構築するのではなく、このオプションを選択し、そのモデルに指定された基準の設定値を再使用して、新しいモデル・ノードをモデル・パレットで生成します。この方法で、以前と同じモデル設定でも最新データを使用してそのモデルに基づいて新しい予測を再推定および作成できるので、時間の節約になります。したがって、たとえば、特定の時系列の元のモデルが Holt の線形トレンドだった場合、同じタイプのモデルがデータの再推定と予測に使用されます。システムは、新しいデータに対する最適なモデル タイプの検出を再試行しません。

スコアリング モデルのみを作成**:** モデルに格納されるデータの量を削減するには、このボックスにチェックマークを付けます。このオプションを使用すると、多数 (1 万単位) の時系列のモデルを作成する場合にパフォーマンスを向上させることができます。データは、依然として通常の方法でスコアリングできます。

レコードの将来への拡張**:** 以下の「予測で使用する将来の値」セクションを有効にします。このセクションでは、推定期間の終わりを超えて予測する時間区分の数を設定できます。この場合の時間区分は、「データ指定」タブで指定した分析の時間区分です。この設定の最大値制限はありません。以下のオプションを使用して、将来の入力値を自動的に計算したり、予測する値を 1 つ以上の予測値に手動で指定したりできます。

予測で使用する将来の値

- 将来の入力値を計算**:** このオプションを選択した場合、予測、ノイズ予測、分散推定、および将来の時間値の予測値が自動的に計算されます。予測が要求されると、対象でもない任意の入力系列に対して、自己回帰モデルが自動的に構築されます。次にこれらのモデルは、予測期間のこれらの入力系列の値を生成するために使用されます。
- 値をデータに追加するフィールドの選択**:** 予測する各レコードに対し (ホールド・アウトは除く)、予測フィールド (役割を**入力**に設定) を使用する場合、各予測の予測期間に対し、推定値を指定できます。手動で値を指定することも、リストから選択することもできます。
  - フィールド**:** フィールド選択ボタンをクリックし、予測として使用するフィールドを選択します。ここで選択したフィールドは、モデル作成で使用されることも、使用されないこともあります。フィールドを予測フィールドとして実際に使用するには、下流のモデル作成ノードで選択する必要があります。このダイアログ・ボックスは将来の値を指定する便利な場所であり、下流のモデル作成ノードによって共有できるので、各ノードで将来の値を個別に指定しなくても済みます。また、利用できるフィールドの一覧が「作成オプション」タブでの選択に制約される場合があります。

    将来の値がストリーム内で以後使用されないフィールドに指定された場合 (削除されたか、「作成オプション」タブで選択が更新されたことが原因)、そのフィールドは赤で表示されることに注意してください。
  - 値**:** 各フィールドに対し、関数のリストから選択するか、または「指定」をクリックして手動で入力または事前に定義された値から選択することができます。予測フィールドが、管理するまたは事前に検知できる項目と関連する場合、値を手動で入力する必要があります。例えば、部屋の予約数に基づいてホテルの翌月の収益を予測する場合、当月実際に取得した予約数を指定することができます。それに対し、予測フィールドが株価など管理外のものと関連する場合、最も最近使用した値や最近使用したポイントの平均などの関数を使用することができます。

    利用できる関数は、フィールドの尺度によって異なります。

表 *28.* 測定の尺度に使用可能な関数

尺度	関数
連続型フィールドまたは名義型フィールド	ブランク 最近使用したポイントの平均 最も最近使用した値 指定
フラグ型フィールド	ブランク 最も最近使用した値 真 偽 指定

「最近使用したポイントの平均」は、最近の 3 つのデータ・ポイントの平均から将来の値を計算します。

「最も最近使用した値」は、将来の値を最新のデータ・ポイントの値に設定します。

「真/偽」は、フラグ型フィールドの将来の値を、指定した真または偽に設定します。

「指定」を選択すると、手動で将来の値を指定するため、または事前定義されたリストから値を選択するためのダイアログ・ボックスが開きます。

スコアリングに使用できるようにする

モデル・ナゲットのダイアログ・ボックスに表示されるスコアリング・オプションのデフォルト値を設定できます。

- 確信度の上限および下限の計算: このオプションを選択すると、各対象フィールドの信頼区間の上限と下限にそれぞれ対応する新規フィールド (デフォルトの接頭辞 $TSLCI- および $TSUCI- が付きます) が作成されます。
- ノイズの残差の計算: このオプションを選択すると、各対象フィールドのモデル残差に対応する新規フィールド (デフォルトの接頭辞 $TSResidual- が付きます) が、これらの値の合計とともに作成されます。

モデルの設定

出力に表示するモデルの最大数: 出力に含めるモデルの最大数を指定します。構築されたモデルの数がこのしきい値を超えると、モデルは出力に表示されませんが、スコアリングには引き続き使用できるということに注意してください。デフォルト値は 10 です。多数のモデルを表示すると、パフォーマンスが低下したり、アプリケーションが不安定になったりする可能性があります。

時系列モデル・ナゲット

時系列モデル・ナゲット出力

時系列モデルを作成すると、以下の情報が出力ビューアーに表示されます。時系列モデルの出力ビューアーに表示できるモデルは、10 個に制限されていることに注意してください。

時間的情報の要約

この要約には以下の情報が表示されます。

- 時間フィールド
- 増分
- 始点および終点
- 固有ポイントの数

要約はすべての対象に適用されます。

「モデル情報」テーブル

各対象に対して繰り返される「モデル情報」テーブルは、モデルについての重要な情報を提供します。このテーブルには、常に以下の上位モデル設定が含まれています。

- データ型ノードまたは時系列ノードのいずれかの「フィールド」タブで選択された対象フィールドの名前。
- モデル構築方法 - 例えば、指数平滑化、または ARIMA。
- モデルへ入力される予測値の数。

- モデル タイプを適合させるために使用されたレコードの数。モデルのさまざまなタイプの例には、RMSE、MAE、AIC、BIC、R2 乗 などがあります。

また、必要な条件を満たすデータであれば、Ljunq-Box Q 統計も表示される可能性があります。この統計は、以下の条件の下では使用できません。

- 非欠損データ ポイントの数が、必要な合計項の数以下の場合 (18 に固定)。
- パラメーターの数が、必要な合計項の数以上の場合。
- 計算される合計項の数が、許容可能な K の最小値未満の場合 (7 に固定)。
- 各対象でそのテーブルが繰り返される場合。

予測変数の重要度

各対象に対して繰り返される「予測変数の重要度」グラフは、モデル内の上位 10 個の入力 (予測値) の重要度を棒グラフとして表示します。

グラフ内に 10 個を超えるフィールドがある場合は、グラフの下のスライダーを使用して、グラフ内に含まれる予測値の選択を変更できます。スライダー上のインディケーター・マークは固定幅であり、スライダー上の各マークは 10 個のフィールドを表します。スライダーに沿ってインディケーター・マークを移動して、予測変数の重要度の順序で並べられた次の 10 個または前の 10 個のフィールドを表示できます。

グラフをダブルクリックして、グラフ設定を編集するための別個のダイアログ・ボックスを開くことができます。例えば、グラフのサイズ、使用されるフォントのサイズと色などの項目を修正できます。この別個の編集ダイアログ・ボックスを閉じると、「出力」タブに表示されるグラフに変更が適用されます。

相関曲線

相関曲線、つまり自己相関プロットは、対象ごとに表示され、残差 (期待値と実際値の間の差異) の自己相関関数 (ACF) または偏自己相関関数 (PACF) およびそれに対する時間ラグを表示します。信頼区間は、グラフ全体にわたって強調表示されます。

パラメーター推定値

各対象に対して繰り返されるパラメーター推定値テーブルには、以下の詳細が表示されます (適用可能な場合)。

- 対象名
- 適用されている変換
- モデル (ARIMA) のこのパラメーターに使用されているラグ
- 係数値
- パラメーター推定値の標準誤差
- パラメーター推定値を標準誤差で割った値
- パラメーター推定値の有意水準

時系列モデルナゲット設定

「設定」タブには、時系列モデル・ナゲットの追加オプションが用意されています。

予測

「レコードの将来への拡張」のオプションで、推定期間の終わりを超えて予測する時間区分の数を設定します。この場合の時間区分は、時系列ノードの「データ指定」タブで指定した分析の時間区分です。予測が要

求されると、対象でもない任意の入力系列に対して、自己回帰モデルが自動的に構築されます。次にこれらのモデルは、予測期間のこれらの入力系列の値を生成するために使用されます。

将来の入力値を計算: このオプションを選択した場合、予測、ノイズ予測、分散推定、および将来の時間値の予測値が計算されます。

予測で使用する将来の値

- 将来の入力値を計算: このオプションを選択した場合、予測、ノイズ予測、分散推定、および将来の時間値の予測値が自動的に計算されます。予測が要求されると、対象でもない任意の入力系列に対して、自己回帰モデルが自動的に構築されます。次にこれらのモデルは、予測期間のこれらの入力系列の値を生成するために使用されます。
- 値をデータに追加するフィールドの選択: 予測する各レコードに対し (ホールド・アウトは除く)、予測フィールド (役割を**入力**に設定) を使用する場合、各予測の予測期間に対し、推定値を指定できます。手動で値を指定することも、リストから選択することもできます。
  - フィールド: フィールド選択ボタンをクリックし、予測として使用するフィールドを選択します。ここで選択したフィールドは、モデル作成で使用されることも、使用されないこともあります。フィールドを予測フィールドとして実際に使用するには、下流のモデル作成ノードで選択する必要があります。このダイアログ・ボックスは将来の値を指定する便利な場所であり、下流のモデル作成ノードによって共有できるので、各ノードで将来の値を個別に指定しなくても済みます。また、利用できるフィールドの一覧が「作成オプション」タブでの選択に制約される場合があります。

    将来の値がストリーム内で以後使用されないフィールドに指定された場合 (削除されたか、「作成オプション」タブで選択が更新されたことが原因)、そのフィールドは赤で表示されることに注意してください。
  - 値: 各フィールドに対し、関数のリストから選択するか、または「指定」をクリックして手動で入力または事前に定義された値から選択することができます。予測フィールドが、管理するまたは事前に検知できる項目と関連する場合、値を手動で入力する必要があります。例えば、部屋の予約数に基づいてホテルの翌月の収益を予測する場合、当月実際に取得した予約数を指定することができます。それに対し、予測フィールドが株価など管理外のものと関連する場合、最も最近使用した値や最近使用したポイントの平均などの関数を使用することができます。

    利用できる関数は、フィールドの尺度によって異なります。

表 *29.* 測定の尺度に使用可能な関数

尺度	関数
連続型フィールドまたは名義型フィールド	ブランク 最近使用したポイントの平均 最も最近使用した値 指定
フラグ型フィールド	ブランク 最も最近使用した値 真 偽 指定

「最近使用したポイントの平均」は、最近の 3 つのデータ・ポイントの平均から将来の値を計算します。

「最も最近使用した値」は、将来の値を最新のデータ・ポイントの値に設定します。

「真/偽」は、フラグ型フィールドの将来の値を、指定した真または偽に設定します。

「指定」を選択すると、手動で将来の値を指定するため、または事前定義されたリストから値を選択するためのダイアログ・ボックスが開きます。

スコアリングで使用可能にする

スコアリングされる各モデルに新規フィールドの作成 **:** スコアリングされる各モデルに作成する新規フィールドを指定できるようにします。

- ノイズの残差**:** これを選択すると、各対象フィールドのモデル残差に対応する新規フィールド (デフォルトの接頭辞 $TSResidual- が付きます) が、これらの値の合計とともに作成されます。
- 確信度の上限および下限**:** このオプションを選択すると、各対象フィールドの信頼区間の上限と下限にそれぞれ対応する新規フィールド (デフォルトの接頭辞 $TSLCI- および $TSUCI- が付きます) が、これらの値の合計とともに作成されます。

スコアリングに含める対象**:** モデル・スコアに含める使用可能な対象を選択します。

第 14 章 自己学習応答ノードモデル

SLRM ノード

自己学習応答モデル (SLRM) ノードでは、データ・セット全体を使用するたびにモデルを再構築する必要のないデータ・セットとして、継続的に更新したりあるいは再推定したりできるモデルを構築できます。例えば複数の製品があり、顧客にオファーする場合にどの製品を顧客が購入するのかを識別する際に有用です。このモデルにより、顧客にとって最も適切な提案および受け入れられる提案の確率を予測できます。

モデルは、任意に行われる提案およびその提案に対する応答により、小さなデータ・セットを使用して最初に構築できます。データ・セットが増大するにつれてモデルを更新できるため、年齢、性別、仕事、および収入の入力フィールドに基づいて、モデルは顧客にとって最適な提案および受け入れられる確率を予測できるようになります。データ・セットの対象フィールドを変更する代わりに、ノード・ダイアログ・ボックス内で追加や削除を行うことにより、利用可能な提案を変更できます。

IBM SPSS Collaboration and Deployment Services と組み合わせると、モデルに対して自動定期更新を設定できます。このプロセスは、人間による監視や活動の必要なしに、データ・マイニングによるユーザー定義の介入が不可能なまたは不要な組織とアプリケーション向けに、柔軟性があり低コストのソリューションを提供します。

例: 金融機関は、それぞれの顧客に受け入れられる提案を行うことで、さらに収益を上げることを望んでいます。自己学習モデルを使用すると、以前の販売促進を基に顧客が最も好意的な反応を示す特徴を識別し、最新の顧客の反応に基づいてリアル タイムでモデルを更新できます。

SLRM ノードのフィールド・オプション

SLRM ノードを実行する前に、そのノードの「フィールド」タブで対象および対象回答の両方のフィールドを指定する必要があります。

対象フィールド**:** 例えば顧客に提供したい異なる製品を含む名義型 (セット型) フィールドを選択します。

注 : 対象フィールドには、数値ではなく文字列を格納する必要があります。

対象回答フィールド **:**リストから対象回答フィールドを選択します。例えば、承認または拒否です。

注 : このフィールドは、フラグ型である必要があります。このフラグの真 (true) の値は受け入れられたオファーを表し、偽 (false) の値は拒否されたオファーを表します。

このダイアログ・ボックスのその他のフィールドは、IBM SPSS Modeler では使用する標準的なものです。 詳しくは、トピック 31 ページの『モデル作成ノードのフィールド・オプション』を参照してください。

注 : 連続型 (数値範囲型) 入力フィールドとして使用する範囲をソース・データが含んでいる場合は、それぞれの範囲の最小と最大の両方の詳細をメタデータが含んでいることを確認する必要があります。

SLRM ノードのモデル・オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

既存モデルの学習を継続: デフォルトでは、モデル作成ノードが実行されるごとに、まったく新しいモデルが作成されます。このオプションを選択すると、ノードによって正常に生成された最後のモデルで学習が続行されます。元のデータにアクセスすることなく既存のモデルを更新またはリフレッシュできます。また、新規レコードまたは更新されたレコードのみ がストリームに適用されるため、パフォーマンスが大幅に向上します。以前のモデルの詳細はモデル作成ノードで保存されるので、以前のモデル・ナゲットがストリームまたは「モデル」パレットでもう使用できない場合でも、このオプションを使用することができます。

対象フィールド値:デフォルトでは、これは 「すべて使用」 に設定され、選択された対象フィールド値に関連するそれぞれのオファーを含むモデルが構築されるということを意味します。可能性のある対象フィールドのオファーのいくつかのみを含むモデルを生成する場合、「指定」 をクリックし、「追加」、「編集」、および 「削除」 のボタンを使用して、モデルを構築するのに使用するオファーの名前を入力または修正します。例えば、供給するすべての製品を表示する対象を選択する場合、このフィールドを使用して提供する製品をここで入力する数に制限することができます。

モデルの評価:このパネルのフィールドはモデルから独立していて、スコアリングには影響を与えません。その代わり、このフィールドにより、モデルがどのようにして結果を予測するかを視覚的に表示できます。

注 : モデル・ナゲットにおけるモデル評価結果を表示するには、「モデル評価の表示」ボックスも選択する必要があります。

- モデルの評価を含める:このボックスを選択して、それぞれの選択したオファーに対するモデルの予測精度を示すグラフを作成します。
- ランダム シードの設定: 無作為なパーセンテージに基づいてレコードの精度を推定する場合、このオプションで、別のセッションに同じ結果を複製できるようになります。乱数ジェネレータに使用される開始値を指定することで、ノードが実行されるごとに毎回同じレコードが割り当てられることが保証されます。希望のシード値を入力します。このオプションが選択されないと、ノードが実行されるごとに異なるサンプルが生成されます。
- シミュレーション サンプル数:モデルを評価する場合にサンプルで使用するレコード数を指定します。デフォルトは 100 です。
- 反復数: これにより、指定された反復数の後にモデル評価の作成を停止できます。最大反復数を指定します。デフォルトは 20 です。

  注 : サンプル数や反復数が増加すると、モデルを構築するのに時間がかかるようになります。

モデル評価の表示 :モデル・ナゲットにおける結果をグラフィカルに表示するには、このオプションを選択します。

SLRM ノードの設定オプション

ノードの設定オプションを使用すると、モデル構築プロセスを微調整できます。

レコードあたりの最大予測数: このオプションを使用すると、データセットの各レコードに作成される予測フィールドの数を制限できます。デフォルトは 3 です。

例えば、6 件のオファー (預金、住宅ローン、カー・ローン、年金、クレジット カード、保険) があり、お勧めの 2 件のみを知りたい場合があるとします。その場合、このフィールドを 2 にセットします。モデ

ルを構築してテーブルに添付した場合、レコードごとに予測列を 2 つ (およびオファーが受け入れられる確率の関連する確信度) が表示されます。予測は 6 つのオファー候補のうちいずれかを使用して行います。

ランダム化のレベル: 例えば小規模なデータセットや不完全なデータセットでバイアスを回避し、すべての潜在オファーを同様に扱うために、オファーの選択や推奨されたオファーとして出現する可能性にランダム化のレベルを追加することができます。ランダム化のレベルは、0.0 (ランダム化なし) ～ 1.0 (完全ランダム化) の小数の割合で表されます。デフォルトは 0.0 です。

ランダム シードの設定: オファーの選択にランダム化のレベルを追加する場合、このオプションを使用すると別のセッションに同じ結果を複製することができます。乱数ジェネレータに使用される開始値を指定することで、ノードが実行されるごとに毎回同じレコードが割り当てられることが保証されます。希望のシード値を入力します。このオプションが選択されないと、ノードが実行されるごとに異なるサンプルが生成されます。

注: データベースから読み込まれたレコードで「ランダム・シードの設定」オプションを使用する場合は、ノードを実行するたびに同じ結果になるように、サンプリングの前にソート・ノードが必要になることがあります。この理由は、ランダム シードがレコードの順序に依存しているためです。各レコードがリレーショナル・データベース内で同じ位置に留まる保証はありません。

ソート順: オファーが次のような作成モデルに表示される順序を選択します。

- 降順: このモデルでは、最大スコアを持つオファーから順に表示します。これらは承認される確率の最も高いオファーです。
- 昇順: このモデルでは、最小スコアを持つオファーから順に表示します。これらは拒否される確率の最も高いオファーです。このオプションは、例えば特定のオファーでマーケティング・キャンペーンからどの顧客を削除するかを決定する場合に有用です。

対象フィールドの優先度: モデルを作成する場合、 促進または削除するデータの特定の側面が存在する場合があります。例えば、最善の財務上のオファーを選択して顧客に販売促進するモデルを作成する場合、各顧客に対して特定のオファーがどれほどスコアリングするかにかかわらず、1 つの特定のオファーが含まれていることを確認したい場合があります。

このパネルにオファーを追加してその優先度を編集するには、「追加」 をクリックし、オファー名 (「貯金」または「住宅ローン」など) を入力して 「**OK**」 をクリックします。

- 値。 このオプションでは、追加したオファーの名前を表示します。
- 優先度: オファーに適用する優先度のレベルを指定します。優先度のレベルは、0.0 (優先度無し) ～ 1.0 (優先度最大) の小数の割合で表されます。デフォルトは 0.0 です。
- 常に表示: 特定のオファーが予測フィールドに常に表示されていることを確認するには、このボックスをオンにします。

  注: 「優先度」が 0.0 に設定されている場合、「常に表示」の設定は無視されます。

モデルの信頼性を考慮: 何度かの再生成によって調整された、構造化され、データの豊富なモデルは、データの少ない新規モデルに比べ、より正確な結果を常に生み出す必要があります。より成熟してモデルの高い信頼度を利用するには、このボックスをオンにします。

SLRM モデル・ナゲット

注 : 「モデル・オプション」タブで「モデルの評価を含める」と「モデル評価の表示」の両方を選択すると、結果はこのタブにのみ表示されます。

SLRM モデルを含んでいるストリームを実行すると、ノードは各対象フィールド値 (オファー) の精度および使用した各予測フィールドの重要度を推定します。

注 : モデル作成ノードの「モデル」タブで「既存モデルの学習を継続」を選択すると、このモデル・ナゲットに関して表示される情報は、モデルを再生成するたびに更新されます。

IBM SPSS Modeler 12.0 以降を使用して構築されたモデルの場合、モデル・ナゲットの「モデル」タブは次の 2 つの列に分割されます。

左側の列 :

- 表示: 複数のオファーがある場合、結果を表示するオファーを選択します。
- モデルのパフォーマンス :これは、それぞれの提案の推定モデル精度を示します。テスト・セットがシミュレーションによって生成されます。

右側の列 :

- 表示: 「応答との関連」か「変数の重要度」のいずれの詳細を表示するかを選択します。
- 応答との関連: それぞれの予測フィールドと目標変数との関連性 (相関) を示します。
- 予測変数の重要度: モデルを推定する際の、各予測値の相対的な重要度を示します。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。このグラフは、予測変数の重要度を表示するその他のモデルと同じ方法で解釈できますが、SLRM の場合は、グラフは SLRM アルゴリズムによるシミュレーションによって生成されます。これは、各予測フィールドを順にモデルから削除してモデルの精度にどのような影響を与えるかを確認することによって行います。 詳しくは、トピック 44 ページの『予測変数の重要度』を参照してください。

SLRM モデル設定

SLRM モデル・ナゲットの「設定」タブは、構築したモデルを修正するオプションを指定します。例えば、同じデータと設定を用いていくつかの異なるモデルを構築するために SLRM ノードを使用し、設定を少しだけ修正して結果に及ぼす影響を確認するにはそれぞれのモデルの同じタブを使用します。

注: このタブは、モデル・ナゲットがストリームに追加された後にのみ使用されます。

レコードあたりの最大予測数: このオプションを使用すると、データセットの各レコードに作成される予測フィールドの数を制限できます。デフォルトは 3 です。

例えば、6 件のオファー (預金、住宅ローン、カー・ローン、年金、クレジット カード、保険) があり、お勧めの 2 件のみを知りたい場合があるとします。その場合、このフィールドを 2 にセットします。モデルを構築してテーブルに添付した場合、レコードごとに予測列を 2 つ (およびオファーが受け入れられる確率の関連する確信度) が表示されます。予測は 6 つのオファー候補のうちいずれかを使用して行います。

ランダム化のレベル: 例えば小規模なデータセットや不完全なデータセットでバイアスを回避し、すべての潜在オファーを同様に扱うために、オファーの選択や推奨されたオファーとして出現する可能性にランダム化のレベルを追加することができます。ランダム化のレベルは、0.0 (ランダム化なし) ～ 1.0 (完全ランダム化) の小数の割合で表されます。デフォルトは 0.0 です。

ランダム シードの設定**:** オファーの選択にランダム化のレベルを追加する場合、このオプションを使用すると別のセッションに同じ結果を複製することができます。乱数ジェネレータに使用される開始値を指定することで、ノードが実行されるごとに毎回同じレコードが割り当てられることが保証されます。希望のシード値を入力します。このオプションが選択されないと、ノードが実行されるごとに異なるサンプルが生成されます。

注**:** データベースから読み込まれたレコードで「ランダム・シードの設定」オプションを使用する場合は、ノードを実行するたびに同じ結果になるように、サンプリングの前にソート・ノードが必要になることがあります。この理由は、ランダム シードがレコードの順序に依存しているためです。各レコードがリレーショナル・データベース内で同じ位置に留まる保証はありません。

ソート順**:** オファーが次のような作成モデルに表示される順序を選択します。

- 降順**:** このモデルでは、最大スコアを持つオファーから順に表示します。これらは承認される確率の最も高いオファーです。
- 昇順**:** このモデルでは、最小スコアを持つオファーから順に表示します。これらは拒否される確率の最も高いオファーです。このオプションは、例えば特定のオファーでマーケティング・キャンペーンからどの顧客を削除するかを決定する場合に有用です。

対象フィールドの優先度**:** モデルを作成する場合、 促進または削除するデータの特定の側面が存在する場合があります。例えば、最善の財務上のオファーを選択して顧客に販売促進するモデルを作成する場合、各顧客に対して特定のオファーがどれほどスコアリングするかにかかわらず、1 つの特定のオファーが含まれていることを確認したい場合があります。

このパネルにオファーを追加してその優先度を編集するには、「追加」 をクリックし、オファー名 (「貯金」または「住宅ローン」など) を入力して 「**OK**」 をクリックします。

- 値。 このオプションでは、追加したオファーの名前を表示します。
- 優先度**:** オファーに適用する優先度のレベルを指定します。優先度のレベルは、0.0 (優先度無し) ～ 1.0 (優先度最大) の小数の割合で表されます。デフォルトは 0.0 です。
- 常に表示**:** 特定のオファーが予測フィールドに常に表示されていることを確認するには、このボックスをオンにします。

  注**:** 「優先度」が 0.0 に設定されている場合、「常に表示」の設定は無視されます。

モデルの信頼性を考慮**:** 何度かの再生成によって調整された、構造化され、データの豊富なモデルは、データの少ない新規モデルに比べ、より正確な結果を常に生み出す必要があります。より成熟してモデルの高い信頼度を利用するには、このボックスをオンにします。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。

- データベースの外部でスコアリング: このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

第 15 章 サポート・ベクター・マシン・モデル

SVM について

サポート・ベクター・マシン (SVM) は、学習データをオーバーフィットすることなくモデルの予測精度を最大化する、堅牢な分類および回帰の技術です。SVM は特に、予測フィールド数が非常に多い (例えば、数千の) データを分析するのに適しています。

SVM には、カスタマ リレーションシップ マネージメント (CRM)、顔面およびその他の画像認識、バイオインフォマティクス、テキスト・マイニング・コンセプト抽出、侵入検知、タンパク質構造の予測、音声認識など、多くの分野のアプリケーションが含まれています。

SVM の動作方法

SVM は、データを高い次元の特徴空間にマップすることで動作するため、データを線状に分かることができない場合であっても、データ・ポイントをカテゴリー別に分けることができます。カテゴリー間の区切りが検出された後、区切りを超平面として描画することができる方法でデータが変換されます。これにより、新しいデータの特性を利用して、新しいレコードが属するグループを予測できます。

例えば、データ・ポイントが 2 つの異なるカテゴリーに含まれる次の図について考えます。

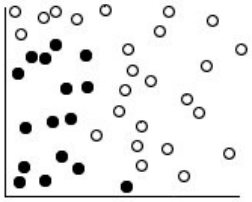


図 *59.* 元のデータセット

次の図に示すように、2 つのカテゴリーは曲線で分けることができます。

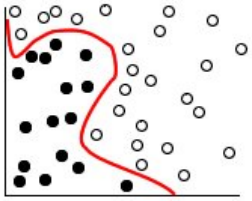


図 *60.* 区切りを追加したデータ

次の図に示すように、変換後、2 つのカテゴリー間の境界を超平面で定義できます。

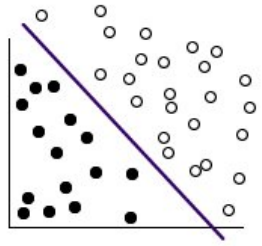


図 *61.* 変換されたデータ

変換に使用される数学関数は、カーネル関数と呼ばれています。IBM SPSS Modeler の SVM は、次のカーネル タイプをサポートしています。

- 線型
- 多項式
- 放射基底関数 (RBF)
- シグモイド

線型カーネル関数は、データの線型区分が直線的である場合にお勧めします。その他の場合は、他の関数のいずれかが使用されます。異なるアルゴリズムおよびパラメーターが使用されているため、さまざまな関数を試して各ケースの最良のモデルを取得する必要があります。

SVM モデルの調整

カテゴリー間の区分線に加え、分類 SVM モデルは 2 つのカテゴリー間の空間を定義する境界線を検出します。

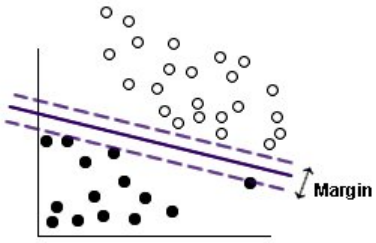


図 *62.* 予備的モデルでのデータ

余白上のデータ・ポイントは、サポート・ベクター として知られています。

2 つのカテゴリー間の余白が広くなると、モデルは新規レコードのカテゴリーの予測がより正確になります。前述の例では、余白があまり広くないため、このモデルはオーバーフィットしているといわれます。少ない数の誤分類を受け入れて余白を広くすることができます。この例が以下の図に示されています。

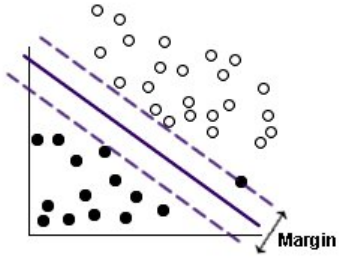


図 *63.* 改良されたモデルを含むデータ

線型区分がより難しい場合があります。この例が以下の図に示されています。

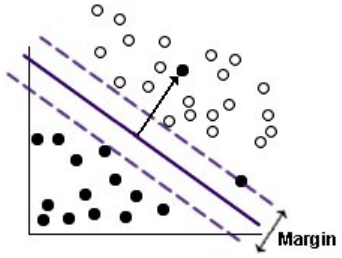


図 *64.* 線型区分の問題

このようなケースでは、広い余白および少数の誤分類データ・ポイント間の最適なバランスを見つけることが目的です。カーネル関数には、これら 2 つの値の間のトレードオフを制御する正則化パラメーター (C) があります。最良のモデルを見つけるために、この異なる値とその他のカーネル・パラメーターを試す必要がある場合があります。

SVM ノード

SVM ノードを使用すると、サポート・ベクトル・マシンを使用してデータを分類できます。SVM は、特に広範なデータセット、つまり多くの予測値フィールドを持つデータセットを使用する場合に適しています。ノードにデフォルト設定を使用して比較的迅速に基本的なモデルを作成できます。またはエキスパート設定を使用して、異なるタイプの SVM モデルを試すことができます。

モデルが構築されると、以下のことができます。

- モデル・ナゲットを参照して、モデル構築における入力フィールドの相対重要度を表示します。
- テーブル・ノードをモデル・ナゲットに追加して、モデル出力を表示します。

例: ある医学研究者が、ガン発症の危険性があると考えられる患者から採取した多くのヒト細胞サンプルの特性を含むデータセットを取得しています。元のデータの分析では、良性と悪性のサンプルの間で、多数の特性が大きく異なることがわかりました。研究者は、他の患者から採取したサンプルの類似した細胞の特性の値を使用できる SVM モデルを開発し、サンプルが良性または悪性かを早期に特定できるようにしたいと考えています。

SVM ノードの「モデル」オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

分割モデルを作成 **:** 分割フィールドとして指定される入力フィールドの各値の個別モデルを作成します。詳しくは、 28 ページの『分割モデルの作成』を参照してください。

SVM ノードの「エキスパート」オプション

サポート・ベクトル・マシンをよく理解している場合は、エキスパート・オプションを使用して学習過程を微調整できます。エキスパート・オプションを利用するには、「エキスパート」タブで「モード」に 「エキスパート」 を設定してください。

すべての確率を追加 **(**カテゴリー・ターゲットにのみ有効**) :**このオプションがオン (チェックマークが入る) の場合、 名義型またはフラグ型のターゲット・フィールドの各値の確率をノードで処理される各レコードに表示することを指定します。このオプションがオフの場合、予測値のみの確率が名義型またはフラグ型対象フィールドに表示されます。このチェック・ボックスの設定により、モデル・ナゲット表示の対応するチェック・ボックスのデフォルト状態を決定します。

停止基準: 最適化アルゴリズムをいつ停止するかを決定します。値の範囲は 1.0E–1 から 1.0E–6 までで、デフォルトは 1.0E–3 です。値を小さくするとモデルはより正確になりますが、モデルは学習に時間がかかるようになります。

正則化パラメーター **(C)**:余白の最大化と学習エラー項の最小化の間のトレードオフを制御します。値は通常、1 以上 10 以下で、デフォルトは 10 です。値を大きくすると、学習データの分類精度が向上 (または回帰エラーが減少) しますが、オーバーフィットする場合もあります。

回帰の精度 **(**イプシロン**)**:対象フィールドの尺度が連続型の場合にのみ使用されます。ここで指定された値より小さい場合、エラーが受け取られます。値を大きくすると、モデル作成の速度が上がりますが、精度は犠牲になります。

カーネル タイプ:変換に使用されるカーネル関数のタイプを指定します。異なるカーネル タイプを使用すると、区切りがさまざまな方法で計算されるため、あらゆるオプションを試すことをお勧めします。デフォルトは **RBF** (Radial Basis Function) です。

RBF ガンマ:カーネル タイプが **RBF** に設定されている場合にのみ有効です。値は通常、$3/k$ ～ $6/k$ で、k は入力フィールドの数を表します。例えば、12 の入力フィールドがある場合、0.25 ～ 0.5 の値を試す価値があります。値を大きくすると、学習データの分類精度が向上 (または回帰エラーが減少) しますが、オーバーフィットする場合もあります。

ガンマ:カーネル タイプが 多項式 または **Sigmoid** に設定されている場合にのみ有効です。値を大きくすると、学習データの分類精度が向上 (または回帰エラーが減少) しますが、オーバーフィットする場合もあります。

Bias:カーネル タイプが 多項式 または **Sigmoid** に設定されている場合にのみ有効です。カーネル関数で `coef0` 値を設定します。デフォルト値 0 は、多くの場合に適しています。

程度:カーネル タイプが 多項式 に設定されている場合にのみ有効です。マッピング空間の複雑さ (次元) を制御します。通常、10 を超える値は使用しません。

SVM モデル・ナゲット

SVM モデルで、多くのフィールドを新規作成します。これらのフィールドで最も重要なのは **$S-fieldname** フィールドで、モデルに予測された対象フィールドの値を示します。

モデルによって作成された新しいフィールドの数および名前は、対象フィールドの尺度によって異なります (このフィールドは次の表の *fieldname* で表示)。

新しいフィールドおよび値を確認するには、テーブル・ノードを SVM モデル・ナゲットに追加してテーブル・ノードを実行します。

表 *30.* 対象フィールドの尺度が「名義型」または「フラグ型」

新規フィールド名	説明
$S-*fieldname*	対象フィールドの予測値。
$SP-*fieldname*	予測値の確率。
$SP-*value*	名義型またはフラグ型の値の確率 (モデル・ナゲットの「設定」タブの「すべての確率を追加」がチェックされている場合にのみ表示)。
$SRP-*value*	(フラグ型対象のみ) 対象フィールドの真 (true) の結果の対数尤度を示す、生 (SRP) および調整された (SAP) 傾向スコア。これらのスコアは、モデルが生成される前に SVM モデル作成ノードの「分析」タブの対応するチェックボックスがオンである場合にのみ表示されます。 詳しくは、トピック 35 ページの『モデル作成ノードの分析オプション』を参照してください。
$SAP-*value*	

表 *31.* 対象フィールドの尺度が「連続型」

新規フィールド名	説明
$S-*fieldname*	対象フィールドの予測値。

予測変数の重要度

オプションで、モデルの推定時に各予測値の相対的重要度を示すグラフを「モデル」タブに表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。このグラフは、モデル生成前に「精度分析」タブで 「予測変数の重要度を計算」が選択されている場合にのみ使用できます。 詳しくは、トピック 44 ページの『予測変数の重要度』を参照してください。

注 : SVM の場合、他のタイプのモデルに比べて予測変数の重要度の計算に時間がかかるため、デフォルトでは「分析」タブで選択されていません。このオプションを選択すると、特に大きなデータセットを含む場合にパフォーマンスの速度が遅くなる場合があります。

SVM モデル設定

「設定」タブを使用すると、結果を表示する場合に追加のフィールドを表示するよう指定することができます (例えば、ナゲットに接続されたテーブル・ノードを実行)。これらのオプションを選択して「プレビュー」ボタンをクリックしてオプションの効果を確認できます。プレビュー出力の右側にスクロールして追加フィールドを表示します。

すべての確率を追加 **(**カテゴリ ターゲットにのみ有効**):** このオプションがオンの場合、 名義型またはフラグ型のターゲット・フィールドの各値の確率をノードで処理される各レコードに表示することを指定します。このオプションがオフの場合、予測された値と確率のみが名義型およびフラグ型の対象フィールドに表示されます。

このチェック・ボックスのデフォルト設定は、モデル作成ノードの対応するチェック・ボックスによって決まります。

未調整傾向スコアを計算**:** フラグ型対象 (yes または no の予測を返す) を持つモデルの場合、対象フィールドに指定された真 (true) の結果の尤度を示す傾向スコアが必要な場合があります。また、スコアリング時に生成することができるその他の予測および確信度値があります。

調整済み傾向スコアを計算 **:** 未調整傾向スコアは、学習データにのみ基づき、このデータがオーバーフィットする多くのモデルの傾向によって過度に楽観的な場合があります。調整済み傾向は、テストまたは検証用データ区分に対してモデルのパフォーマンスを評価することによって補正しようとします。このオプションでは、モデルの生成前にデータ区分フィールドをストリーム内で定義し、調整済み傾向スコアがモデル作成ノードで有効化されている必要があります。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL 生成の実行方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- データベースの外部でスコアリング**:** このオプションを選択すると、データベースからデータが再度フェッチされ、SPSS Modeler 内でスコアリングが行われます。

LSVM ノード

LSVM ノードでは、線型サポート ベクター マシンを使用してデータを分類できます。LSVM は、広範なデータセット、つまり多数の予測値フィールドがあるデータセットでの使用に特に適しています。ノードにデフォルト設定を使用すると、基本的なモデルを比較的迅速に作成できます。あるいは、作成オプションを使用してさまざまな設定を試すこともできます。

LSVM ノードは SVM ノードに類似していますが、線形であり、多数のレコードを処理するのに優れています。

モデルが構築されると、以下のことができます。

- モデル・ナゲットを参照して、モデル構築における入力フィールドの相対重要度を表示します。
- テーブル・ノードをモデル・ナゲットに追加して、モデル出力を表示します。

例: ある医学研究者が、ガン発症の危険性があると考えられる患者から採取した多くのヒト細胞サンプルの特性を含むデータセットを取得しています。元のデータの分析では、良性と悪性のサンプルの間で、多数の

特性が大きく異なることがわかりました。この研究者は、他の患者から採取したサンプルの類似した細胞の特性の値を使用できる LSVM モデルを開発して、サンプルが良性または悪性かを早期に特定できるようにしたいと考えています。

LSVM ノードのモデル オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

予測変数の重要度を計算**:** 重要度の適切な測定基準を作成するモデルの場合、モデル推定時に各予測値の相対重要度を示すグラフを表示することができます。通常、ユーザーはモデル作成の目標を最も重要な予測値に焦点を当て、最も重要でない変数を削除または無視したいと考えます。モデルによっては、特に大きなデータセットを使用する場合、予測変数の重要度の計算に時間がかかることがあります。そのため、一部のモデルではデフォルトでオフになっています。予測変数の重要度は、ディシジョン・リスト・モデルには使用できません。 詳しくは、 44 ページの『予測変数の重要度』を参照してください。

LSVM 作成オプション

モデルの設定

定数項を含める: 定数項 (モデルの定数項) を含めると、解の精度を全体的に向上させることができます。データが原点を通ると仮定できる場合は、切片を除外できます。

カテゴリ型対象のソート順: カテゴリ型対象のソート順を指定します。連続型対象の場合は、この設定が無視されます。

回帰の精度 **(**イプシロン**)**:対象フィールドの尺度が連続型の場合にのみ使用されます。ここで指定された値より小さい場合、エラーが受け取られます。値を大きくすると、モデル作成の速度が上がりますが、精度は犠牲になります。

欠損値があるレコードを除外: **True** に設定すると、欠損値が 1 つでもある場合にレコードが除外されます。

ペナルティの設定

ペナルティ関数: オーバーフィットの尤度を削減するために使用されるペナルティ関数のタイプを指定します。オプションは、**L1** または **L2** です。

L1 および **L2** は、係数にペナルティを追加することにより、オーバーフィットの可能性を削減します。これらのオプションの違いは、特徴量が多数ある場合、**L1** は、モデル構築時にいくつかの係数を 0 に設定することにより、特徴量選択を使用するという点です。**L2** には、この機能がないため、特徴量が多数あるときは使用しないでください。

ペナルティ パラメータ **(**ラムダ**)**: ペナルティ (正規化) パラメータを指定します。この設定は、ペナルティ関数 が設定されている場合に有効となります。

LSVM モデル ナゲット **(**対話式の出力**)**

LSVM モデルを実行すると、以下の出力が得られます。

モデル情報

「モデル情報」ビューは、モデルについての重要な情報を提供します。テーブルは次のようなハイレベルなモデル設定を特定します。

- 「フィールド」タブで指定されている対象の名前
- モデル選択設定で指定されたモデル構築方法
- 予測値入力の数
- 最終モデル内の予測値の数
- 正規化タイプ (L1 または L2)
- ペナルティ パラメータ (ラムダ)。これは正則化パラメータです。
- 回帰精度 (イプシロン)。エラー数がこの値より少なければ、エラーは許容されます。値を大きくすると、モデル作成の速度が上がりますが、精度は犠牲になります。対象フィールドの尺度が連続型 の場合にのみ使用されます。
- 分類精度の割合。これは分類にのみ関係します。
- 平方平均エラー。これは回帰にのみ関係します。

レコード要約

「レコード要約」ビューは、モデルに組み込まれたレコード (ケース) とモデルから除外されたレコード (ケース) の数および割合についての情報を提供します。

予測変数の重要度

一般にモデリングの作業では、最も重要な予測フィールドの編集に集中して取り組み、最も重要でない予測フィールドは削除するか無視してしまいたくなるものです。予測値の重要度グラフを使用すると、モデル推定時に各予測値の相対重要度が示されるので便利です。値が相対的であるため、表示されるすべての予測変数の値の合計は 1.0 となります。予測値の重要度は、モデルの精度に関連しません。予測が正確かどうかに関係なく、予測時における各予測の重要度にのみ関連します。

予測対観測

縦軸の予測値に対し横軸に観測値を示した分割散布図が表示されます。点は 45 度の線上にあるのが理想です。このビューで、モデルによるレコードの予測にとりわけ問題があるかがわかります。

注: LSVM および SVM では、他のタイプのモデルに比べて予測変数の重要度の計算に時間がかかる場合があります。このオプションを選択すると、特に大きなデータセットを含む場合にパフォーマンスの速度が遅くなる場合があります。

混同マトリックス

混同マトリックスは、集計表とも呼ばれ、LSVM 分析に基づいて各グループに正しくまたは誤って割り当てられたケースの数を示します。

LSVM モデルの設定

LSVM モデル ナゲットの「設定」タブでは、モデル スコアリング時の未調整傾向のオプションおよび SQL 生成のオプションを指定します。このタブは、モデル・ナゲットがストリームに追加された後にのみ使用できます。

未調整傾向スコアの計算: フラグ型対象のみを含むモデルについては、対象フィールドに true の結果が指定されている尤度を示す未調整傾向スコアを要求できます。これらは、標準の予測値と確信度値に追加されています。調整済み傾向スコアは使用できません。

このモデルの **SQL** を生成**:** データベースのデータを使用する場合に、SQL コードをデータベースにプッシュ バックして実行することができます。これにより、多くの操作のパフォーマンスを向上させることができます。

SQL の生成方法を指定するには、次のオプションのいずれかを選択します。

- デフォルト**: Server Scoring Adapter (**インストールされている場合**)** を使用してスコアリング **(**インストールされていない場合はインプロセス**):** スコアリング・アダプターがインストールされたデータベースに接続した場合は、スコアリング・アダプターおよび関連付けられたユーザー定義関数 (UDF) を使用して SQL を生成し、データベース内でモデルをスコアリングします。使用可能なスコアリング・アダプターがない場合、このオプションは、データベースからデータを再度フェッチし、SPSS Modeler でそのデータをスコアリングします。
- データベースの外部でスコアリング**:** 選択した場合、このオプションは、データベースからデータをフェッチし、SPSS Modeler でそのデータをスコアリングします。

第 16 章 最近傍モデル

KNN ノード

最近傍分析は、そのほかのケースに対する類似性に基づいてケースを分類する方法です。マシン学習で、保存されたパターン、またはケースに完全に一致する必要なくデータのパターンを認識する方法として開発されました。類似したケースはお互いに近く、類似していないケースはお互いに離れています。つまり、2 つのケース間の距離は、それらの非類似度の尺度です。

互いに近いケースを「近傍」と呼びます。新しいケース (ホールドアウト) が存在する場合、モデル内の各ケースからその新しいケースへの距離が計算されます。最も類似した分類「最近傍」が集計され、新しいケースが、最大数の最近傍を含むカテゴリーに投入されます。

検証する最近傍の数を指定できます。この値は k となります。図は、新しいケースが 2 つの異なる値の k を使用してどのように分類されるかを示します。$k = 5$ の場合、最近傍の大部分はカテゴリー *1* に属するため、新しいケースはカテゴリー *1* にあります。ただし $k = 9$ の場合、最近傍の大部分はカテゴリー *0* に属するため、新しいケースはカテゴリー *0* にあります。

また、最近傍分析を使用して、連続型対象値を計算することもできます。この場合、最近傍の平均または中央の対象値を使用して、新しいケースの予測値を取得します。

KNN ノードの目的オプション

「目的」タブで、最近傍の値に基づいて入力データの対象フィールドの値を予測するモデルを作成するか、単に関心のある特定のケースの最近傍を検索するかを選択することができます。

どの種類の分析を実行しますか?

対象フィールドを予測**:** 最近傍の値に基づいて対象フィールドの値を予測する場合に選択します。

最近傍のみを識別**:** 特定の入力フィールドの最近傍の確認を行う場合にのみ選択します。

最近傍のみを識別する場合、残りのオプションは対象の予測にのみ関連するため、精度および速度に関連するタブのそれらは無効となります。

目的は何ですか?

対象フィールドを予測する場合、このグループのオプションを使用すると、対象フィールド予測時に最も重要な因子となるのは速度、精度、またはこれらの組み合わせのどれになるのかを決定することができます。または、設定をカスタマイズすることもできます。

「バランス」、「速度」、または「精度」オプションを選択すると、アルゴリズムはそのオプションに最も適切な組み合わせの設定を事前に選択します。高度なユーザーは、これらの選択を上書きしたい場合があります。「設定」のさまざまなパネルで実行可能です。

速度と精度のバランス**:** 小さな領域内で最適な数の近傍を選択します。

速度**:** 固定された近傍数を検索します。

精度**:** 大きい領域内で最適な数の近傍を選択し、距離の計算時に予測変数の重要度を使用します。

カスタム分析**:** このオプションを選択して、「設定」タブでアルゴリズムを調整します。

注**:** 他の多くのモデルとは異なり、KNN モデルのサイズは、学習データの量によって直線的に増大します。KNN モデルを作成しようとするときに、「メモリー不足」のエラー・レポートが表示された場合、IBM SPSS Modeler で使用する最大システム メモリー容量を増やしてください。容量を増やすには、次のメニューを選択します。

「ツール」 > 「オプション」 > 「システム オプション」

「最大メモリー」 フィールドに新しいサイズを入力します。「システム オプション」ダイアログで行った変更を有効にするには、IBM SPSS Modeler を再起動します。

KNN ノード設定

「設定」タブで、最近傍分析に固有のオプションを指定します。画面の左側にある再度バーには、オプションの指定に使用するパネルが表示されます。

モデル

「モデル」パネルでは、モデルの作成方法 (例: データ区分または分割モデルのどちらを使用するか、すべてが同じ領域になるよう数値型入力フィールドを変換するかどうか、関心のあるケースの管理方法) を制御するオプションを提供します。また、モデルのカスタム名を選択することもできます。

注**:** 「区分されたデータを使用」および「ケース ラベルの使用」で同じフィールドを使用することはできません。

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

データ区分データを使用。 データ区分フィールドが定義されている場合、このオプションでは学習用データ区分からのデータのみがモデル構築に使用されるようにします。

分割モデルを作成 **:** 分割フィールドとして指定される入力フィールドの各値の個別モデルを作成します。詳しくは、 28 ページの『分割モデルの作成』を参照してください。

手動でフィールドを選択するには **:** デフォルトでは、ノードはデータ区分を使用してフィールド設定 (あれば) をデータ型ノードから分割しますが、これらの設定をここで上書きすることができます。 「データ区分」 および 「分割」 フィールドを有効にするには、「フィールド」 タブを選択して 「ユーザー設定を使用」 を選択し、ここに戻ります。

- データ区分**:** このフィールドでは、モデル構築の学習、テスト、および検証の各ステージ用に、データを独立したサブセット (サンプル) に分割するフィールドを指定できます。1 組のサンプルをモデルの生成に使用し、別のサンプルで生成したモデルをテストすることにより、そのモデルが、このデータに似た性質を持つより大きなデータセットにどの程度適用できるかについての良い目安を得ることができます。データ型ノードまたはデータ区分ノードを使用して複数のデータ区分フィールドが定義されている場合、データ区分を使用する各モデル作成ノードの「フィールド」タブで単一のデータ区分フィールドを選択する必要があります (1 つのデータ区分だけが存在している場合、データ分割を有効にすると、そのデータ区分が必ず自動的に使用されます)。また、選択したデータ区分を分析に適用するには、そのノードの「モデル・オプション」タブでデータ区分が有効になっている必要があります (このオプションの選択を解除すると、フィールド設定を変更することなくデータ区分を無効にすることができます)。
- 分割。分割モデルについて、分割フィールドを選択します。これは、データ型ノードのフィールドの役割を 「分割」 に設定するのと似ています。「フラグ型」、「名義型」 または 「順序型」 のフィールドのみ、分割フィールドとして指定できます。分割フィールドとして選択されたフィールドは、対象フ

ィールド、入力フィールド、データ区分フィールド、度数フィールドまたは重みフィールドとして使用できません。 詳しくは、トピック 28 ページの『分割モデルの作成』を参照してください。

範囲入力の正規化**:** 連続型入力フィールドの値を正規化します。正規化機能には同じ範囲の値があり、推定アルゴリズムのパフォーマンスを改善できます。調整済み正規化、 [2*(x–min)/(max–min)]–1 が使用されます。調整済み正規化の値は -1 ～ 1 です。

ケース ラベルの使用**:** モデル・ビューアーでドロップダウン・リストが有効化され、予測領域のグラフ、ピア・グラフ、象限情報マップで、関心のあるケースを識別するラベルとして値を使用するフィールドを選択することができます。ラベル付けフィールドとして使用するフィールドの測定の尺度は、名義型、順序型、またはフラグ型のどれでもかまいません。ここでフィールドを選択しない場合、入力データの行番号で識別される最近傍を使用して、レコードがモデル・ビューアーのグラフに表示されます。モデル作成後にデータを処理する場合、ケース・ラベルを使用して、表示でケースを特定するごとに入力データを参照しないようにします。

重要レコードの特定**:** ドロップダウン・リストを有効にし、特別に関心のある入力フィールドをマークすることができます (フラグ型フィールドのみ)。ここでフィールドを指定すると、モデル作成時にモデルビューアーでフィールドを示すポイントが最初に選択されます。重要レコードの選択はオプションです。モデル・ビューアーで手動で選択すれば、どのポイントも一時的に重要レコードとなります。

近傍

「近傍」パネルには、計算される最近傍の数を制御する一連のオプションが含まれています。

最近傍の数 **(k):**特定のケースの最近傍数を指定します。より大きな数の近傍を使用すると、必ずしも正確なモデルが作成されるとは限りません。

目的が対象を予測する場合、次の 2 つの選択肢があります。

- 固定値の **K** を選択**:** 検出する最近傍の固定数を指定する場合、このオプションを使用します。
- 自動的に **k** を選択**:** 「最小値」および「最大値」フィールドを使用して、値の範囲を指定し、その範囲内にある最適な近傍数を選択することもできます。最近傍の数を決定する方法は、「特徴量選択」パネルが「特徴量選択」で要求されているかどうかによって異なります。

  特徴量選択が有効である場合、特徴量選択は要求された範囲の *k* の各値に実行され、 最も低い誤差率 (または対象が連続型の場合、最も低い平方和の誤差) の *k* および付随する特徴量のセットが選択されます。

  特徴量選択が有効ではない場合、*V* 群交差検証を使用して、「最適な」近傍数が選択されます。群の割り当てのコントロールについては「交差検証」タブを参照してください。

奥行きの計算: ケースの類似度の測定に使用される距離基準を指定するための計量です。

- ユークリッド メトリック**:** x と y の 2 つのケース間の距離は、すべての次元において、それらのケースの値の差の平方和の平方根になります。
- 都市ブロック メトリック**:** 2 つのケースの間の距離は、すべての次元において、それらのケースの値の絶対差の合計になります。Manhattan 距離とも呼ばれます。

オプションで、対象の予測を目的としている場合、奥行き計算時の正規化重要度によって特徴を重み付けすることができます。予測の特徴の重要度は、モデルから削除された予測変数を持つモデルの誤差率または平方和の誤差の、完全モデルの誤差率または平方和の誤差に対する比率によって計算されます。正規化された重要度は、合計が 1 となるよう、特徴重要度の値を再度重み付けして計算します。

距離計算時の重要度による重み付け機能**:** (目的が「対象の予測」となっている場合にのみ表示されます。) 予測変数の重要度が、近傍間の距離を計算する場合に使用されます。予測変数の重要度がモデル・ナゲットに表示され、予測に使用されます (また、予測に影響が与えられます)。 詳しくは、トピック 44 ページの『予測変数の重要度』を参照してください。

範囲目標の予測**:** (目的が「対象の予測」となっている場合にのみ表示されます。)連続型 (数値範囲) 対象が指定されている場合、このオプションでは予測された値が最近傍の平均値または中央値のどちらに基づいて計算するかを定義します。

特徴量選択

このパネルは、目的が「対象の予測」となっている場合にのみ有効化されます。特徴量選択のオプションを要求および指定できます。デフォルトでは、すべての特徴量が特徴量選択用に考慮されていますが、オプションで、特徴量のサブセットを選択してモデル内に適用することができます。

特徴量選択の実行**:** 特徴量選択のオプションを有効にします。

- 強制投入法 **:** このボックスの隣にあるフィールド・ピッカー・ボタンをクリックして、モデルに強制する変数を選択します。

停止基準: 各ステップで、モデルへの追加により誤差が最も小さくなる (カテゴリー型対象の誤差率および連続型対象の誤差の平方和として計算) 変数がモデル・セットに選択すると見なされます。変数増加法は、指定された条件を満たすまで続行します。

- 指定した数の変数を選択したときに停止。 アルゴリズムでは、モデルに強制的に投入された変数に加え、固定数の変数を追加します。正の整数を指定してください。選択する数値を減らすと、より節約的なモデルが作成され、重要な変数が欠損するというリスクがあります。選択する数値を増やすと、すべての重要な変数を取得しますが、モデル誤差が増加する変数を追加するというリスクがあります。
- 絶対誤差率の変化が最小値以下となった場合に停止。 絶対誤差比の変化量が、これ以上変数を追加してもモデルが改善されないことを示す場合、アルゴリズムは停止します。正の数値を指定してください。変化の最小値を減少させると、より多くの変数を追加しますが、モデルに多くの値を追加しない変数を追加してしまうというリスクがあります。最小変化量の値を大きくすると、より多くの変数を除外しますが、モデルに重要な変数を失うというリスクがあります。最小変化量の「最適な」値は、データおよびアプリケーションによって異なります。どの特徴量が最も重要か評価する方法については、出力の特徴量選択エラー・ログを参照してください。 詳しくは、トピック 372 ページの『予測値選択エラー・ログ』を参照してください。

交差検証

このパネルは、目的が「対象の予測」となっている場合にのみ有効化されます。このパネルのオプションで、最近傍の計算時に交差検証を使用するかどうかを制御します。

交差検証では、サンプルを群と呼ばれる複数のサブサンプルに分割します。分割の後、最近傍モデルが生成されますが、各サブサンプルのデータは除外されます。つまり、最初のモデルは最初のサブサンプル以外のすべてのケースを基に生成され、2 番目のモデルは 2 番目のサブサンプル以外のすべてのケースを基に生成されます。それぞれのモデルを、そのモデルの生成時に除外したサブサンプルに適用し、誤差を推定します。最近傍の「最適な」数は、群全体で最も誤差が少ない数です。

交差検証分割**:***V* 群交差検証を使用して、近傍の「最適な」数が判断されます。パフォーマンス上の理由で、特徴量選択と組み合わせて使用することはできません。

- ケースを **fold** に無作為に割り当て**:**交差検証で使用する群の数を指定します。この手続きでは、1 から *V* (群の数) まで、ケースを群に割り当てます。

- ランダム シードの設定**:** 無作為なパーセンテージに基づいてレコードの精度を推定する場合、このオプションで、別のセッションに同じ結果を複製できるようになります。乱数ジェネレータに使用される開始値を指定することで、ノードが実行されるごとに毎回同じレコードが割り当てられることが保証されます。希望のシード値を入力します。このオプションが選択されないと、ノードが実行されるごとに異なるサンプルが生成されます。
- フィールドを使用してケースを割り当て**:** 群にアクティブなデータセットの各ケースを割り当てる数値型フィールドを指定します。このフィールドには、1 から *V* までの数値を指定を指定する必要があります。値がこの範囲外にあるときに、任意の分割フィールドで分割モデルが有効な場合は、エラーが発生します。

分析

「分析」パネルは、目的が「対象の予測」となっている場合にのみ有効化されます。このオプションを使用して、次を含む追加の変数を使用するかどうかを指定します。

- 対象フィールド値の確率
- ケースと最近傍との間の距離
- 未調整および調整済み傾向スコア (フラグ型対象のみ):

すべての確率を追加**:** このオプションがオンの場合、 名義型またはフラグ型のターゲット・フィールドの各値の確率をノードで処理される各レコードに表示することを指定します。このオプションがオフの場合、予測された値と確率のみが名義型およびフラグ型の対象フィールドに表示されます。

ケースと **k** 最近傍との距離を保存**:** 各重要レコードについて、重要レコードの (学習サンプルからの) *k* の最近傍と、対応する *k* の最短距離のそれぞれの変数が作成されます。

傾向スコア

傾向スコアは、モデル作成ノードで、またはモデル・ナゲットの「設定」タブで有効にできます。この機能は、選択された対象がフラグ型フィールドである場合にのみ使用できます。 詳しくは、トピック 36 ページの『傾向スコア』を参照してください。

未調整傾向スコアを計算**:** 生の傾向スコアは学習データだけに基づいたモデルから得られます。モデルが *true* 値 (応答する) を予測する場合、傾向は P と同じになります。ここで P は、予測値の確率です。モデルが false 値を予測する場合、傾向は (1 - P) と算出されます。

- モデルを構築する際にこのオプションを選択すると、傾向スコアはそのモデル・ナゲット内でデフォルトで有効になります。ただし、モデル作成ノードで選択したかどうかにかかわらず、モデル・ナゲット内でいつでも生の傾向スコアを有効にできます。
- モデルをスコアリングする際、生の傾向スコアは、標準の接頭辞に *RP* が追加されてフィールドに追加されます。例えば、予測値が *$R-churn* という名前のフィールドにある場合は、傾向スコア フィールドの名前は *$RRP-churn* となります。

調整済み傾向スコアを計算 **:** 生の傾向スコアはモデルによって与えられた推定値のみに基づいて算出されますが、これはオーバフィットしている可能性があり、極端に楽観的な傾向が推定されることがあります。調整済み傾向スコアは、テスト・データ区分や検証データ区分に対するモデルの成果を調べて、傾向を調整することによって、より的確な推定を行うものです。

- この設定では、ストリームに有効なデータ区分フィールドが存在している必要があります。
- 生の傾向スコアと違い、調整済み傾向スコアは、モデルを構築するときに計算されなければなりません。そうでなければ、モデル・ナゲットをスコアリングするときにそれらを使用することはできません。

- モデルをスコアリングする際、調整済み傾向スコアは、標準の接頭辞に *AP* が追加されてフィールドに追加されます。例えば、予測値が *$R-churn* という名前のフィールドにある場合は、傾向スコア フィールドの名前は *$RAP-churn* となります。調整済み傾向スコアは、ロジスティック回帰モデルには使用できません。
- 調整済み傾向スコアを計算する場合、計算に使用するテスト・データ区分または検証データ区分はバランス化されていてはいけません。そのため、上流のバランス・ノードで 「学習データのみをバランス」オプションを必ず選択します。さらに、複雑なサンプルが上流にとられた場合は、それによって調整済み傾向スコアが無効になります。
- 調整済み傾向スコアは、「ブーストされた」ツリーまたはルールセット・モデルには使用できません。詳しくは、トピック 130 ページの『ブーストされた C5.0 モデル』を参照してください。

KNN モデル・ナゲット

KNN モデルは、次の表に示されているように、多くの新しいフィールドを作成します。新しいフィールドおよび値を確認するには、テーブル・ノードを KNN モデル・ナゲットに追加してテーブル・ノードを実行するか、ナゲットの「プレビュー」ボタンをクリックします。

表 *32. KNN* モデル・フィールド

新規フィールド名	説明
$KNN-*fieldname*	対象フィールドの予測値。
$KNNP-*fieldname*	予測値の確率。
$KNNP-*value*	名義型またはフラグ型の各値の確率。「すべての確率を追加」 がモデル・ナゲットの「設定」タブでオンになっている場合にのみ含まれます。
$KNN-neighbor-*n*	重要レコードに対する *n* 番目の最近傍の名前。モデル・ナゲットの「設定」タブで「最近傍の表示」 がゼロ以外の値に設定されている場合にのみ含まれます。
$KNN-distance-*n*	重要レコードに対する *n* 番目の最近傍の重要レコードからの相対距離。モデル・ナゲットの「設定」タブで 「最近傍の表示」 がゼロ以外の値に設定されている場合にのみ含まれます。

最近傍モデル・ビュー

モデル・ビュー

このモデル・ビューには、以下の 2 つのパネルで構成されたウィンドウがあります。

- 最初のパネルはメイン・ビューと呼ばれ、モデルの概要が表示されます。
- 2 番目のパネルには、次の 2 種類のビューのいずれかが表示されます。

  モデルの詳細が表示され、モデル自体には焦点を当てていない補助的なモデル・ビュー。

  ユーザーがメイン・ビューの一部について掘り下げた場合、モデルのある特徴についての詳細を示すリンク ビュー。

デフォルトでは、1 つめのパネルで予測領域を示し、2 つめのパネルで予測変数の重要度グラフを表示します。予測変数の重要度のグラフが使用できない場合、つまり「設定」タブの「近傍」パネルで「重要度による重み付け機能」が選択されていない場合は、「ビュー」ドロップダウンで最初に使用できるビューが表示されます。

ビューに使用できる情報がない場合、「ビュー」ドロップダウンには表示されません。

予測値の領域: 予測領域のグラフは、予測領域 (または、3 件を上回る予測値がある場合、部分空間) のインタラクティブ グラフです。それぞれの軸はモデルの予測値を示し、グラフの点の場所は、学習およびホールドアウト分割のケースにおけるこれらの予測値を示します。

キー: 予測値のほか、図表内の点はその他の情報を示します。

- 形状は、点が属する分割 (学習またはホールドアウト) を示します。
- 点の色/網掛けはそのケースの目標の値を示します。それぞれの色でカテゴリー目標のカテゴリーを示し、網掛けは連続型目標の値の範囲を示します。学習分割に示された値は観測値で、ホールドアウト分割は、予測値となります。目標が指定されていない場合、このキーは表示されません。
- 太い枠線は、そのケースが中心ケースであることを示しています。重要レコードは、*k* 最近傍へのリンクを示します。

コントロールおよび双方性: 図表内の多くのコントロールを使用して、予測領域を調べることができます。

- グラフ内に表示する予測のサブセットを選択でき、また次元で表示される予測を変更できます。
- 「重要レコード」は予測空間のグラフに選択された点です。重要レコード変数を指定すると、重要レコードを示す点が最初に選択されます。ただし、いかなる点を選択しても、一時的に重要レコードとなります。ポイント選択の「通常の」コントロールが適用されます。点選択の「通常の」コントロールが適用されます。点をクリックすると、その点が選択され、それ以外の点がすべて選択解除されます。Ctrl キーを押しながら点をクリックすると、選択している一連の点にその点が追加されます。同位図などのリンク ビューは、予測領域で選択されたケースに基づいて自動的に更新されます。
- 最近傍の数 (*k*) を変更して重要レコードで表示することができます。
- カーソルを図内の点に移動すると、ケース・ラベルの値を含む tooltip 、またはケース・ラベルが定義されていない場合はケース数、そして観測目標値および予測目標値が表示されます。
- 「リセット」ボタンを使用して、予測領域を元の状態に戻すことができます。

予測領域グラフの軸の変更: 予測領域のグラフの軸にどの特徴を表示するかを制御できます。

軸の設定を変更する手順は、次のとおりです。

1. 左側のパネルの「編集モード」ボタン (刷毛のアイコン) をクリックして、予測領域の「編集」モードを選択します。
2. 右側のパネルのビューを変更します。「区域の表示」 パネルが、2 つのメイン・パネルの間に表示されます。
3. 「区域の表示」 チェック・ボックスをクリックします。
4. 予測領域の任意のデータ・ポイントをクリックします。
5. 軸を同じデータ型の予測と置き換えるには
   - 置き換える予測の区域ラベル (小さい X ボタンを持つ) に新しい予測をドラッグします。
6. 軸を異なるデータ型の予測と置き換えるには
   - 置きかえる予測の区域ラベルで、小さい X ボタンを押します。予測領域が二次元の表示に変わります。
   - 「次元の追加」 区域ラベルに新しい予測をドラッグします。
7. 左側のパネルの「検証モード」ボタン (矢印のアイコン) をクリックして、「編集」モードを終了します。

予測変数の重要度: 一般にモデリングの作業では、最も重要な予測フィールドの編集に集中して取り組み、最も重要でない予測フィールドは削除するか無視してしまいたくなるものです。予測値の重要度グラフを使用すると、モデル推定時に各予測値の相対重要度が示されるので便利です。値が相対的であるため、表

示されるすべての予測変数の値の合計は 1.0 となります。予測値の重要度は、モデルの精度に関連しません。予測が正確かどうかに関係なく、予測時における各予測の重要度にのみ関連します。

最近傍の距離: この表には、重要レコードのみの *k* 最近傍と距離が表示されます。この表は、重要レコードの識別子がモデル作成ノードで指定されている場合に使用することができ、この変数によって識別された重要レコードだけが表示されます。

各行の内容は以下のとおりです。

- 「重要レコード」 列には重要レコードのケース・ラベル変数の値が表示されます。ケースのラベルが定義されていない場合、この列には重要レコードのケース数が表示されます。
- 「最近傍」グループの *i* 番目の列には、重要レコードの *i* 番目の最近傍のケース・ラベル変数の値が含まれます。ケース・ラベルが定義されていない場合、この列には重要レコードの *i* 番目のケース番号が含まれます。
- 「最短距離」グループの *i* 番目の列には、重要レコードの *i* 番目の最近傍の距離が含まれます。

同位: この図は、各予測値および目標の中心ケースおよび *k* 最近傍を表示します。中心ケースが予測領域で選択されている場合に使用できます。

ピア・グラフは予測領域と、2 つの点でリンクしています。

- 予測領域で選択された (中心) ケースは、*k* 最近傍とともに同位図に表示されます。
- 予測領域で選択された *k* の値は、同位図で使用されます。

予測値を選択: ピア・グラフに表示する予測値を選択することができます。

象限マップ: この表には、中心ケースと *k* 最近傍が散布図 (または、目標の尺度に応じてドット プロット) で表示されます。*y* 軸には目標、*x* 軸には予測値を表示し、予測ごとにパネル表示します。目標があり、中心ケースが予測領域で選択されている場合に使用できます。

- 連続変数について参照線が、学習分割の変数の平均値で描画されます。

予測値を選択: 象限情報グラフに表示する予測値を選択することができます。

予測値選択エラー・ログ: 図内の点は、モデルの *y* 軸に誤差 (目標の尺度に応じて誤差率または誤差の平方和) を示し、*x* 軸は予測値を示します (*x* の左側にすべての変数が表示されます) 。目標があり、特徴量選択が有効である場合、この図を使用することができます。

分類表: このテーブルには、目標の観測値と予測値のクロス分類が分割ごとに表示されます。対象があり、それがカテゴリー (フラグ型、名義型、または順序型) である場合に使用できます。

- ホールドアウト分割の 「**(**欠損値**)**」 行には、目標に欠損値を持つホールドアウト ケースが表示されます。これらのケースはホールドアウト・サンプル: すべてのパーセントの値には寄与しますが、正分類パーセントの値には寄与しません。

誤差の集計: このテーブルは、目標変数が存在する場合に使用することができます。このテーブルには、モデルに関連する誤差が表示されます。連続型目標の場合は平方和が表示され、カテゴリー型目標の場合は誤差率 (100% – 全正分類パーセント) が表示されます。

KNN モデル設定

「設定」タブを使用すると、結果を表示する場合に追加のフィールドを表示するよう指定することができます (例えば、ナゲットに接続されたテーブル・ノードを実行)。これらのオプションを選択して「プレビュー」ボタンをクリックしてオプションの効果を確認できます。プレビュー出力の右側にスクロールして追加フィールドを表示します。

すべての確率を追加 (カテゴリ ターゲットにのみ有効): このオプションがオンの場合、 名義型またはフラグ型のターゲット・フィールドの各値の確率をノードで処理される各レコードに表示することを指定します。このオプションがオフの場合、予測された値と確率のみが名義型およびフラグ型の対象フィールドに表示されます。

このチェック・ボックスのデフォルト設定は、モデル作成ノードの対応するチェック・ボックスによって決まります。

未調整傾向スコアを計算: フラグ型対象 (yes または no の予測を返す) を持つモデルの場合、対象フィールドに指定された真 (true) の結果の尤度を示す傾向スコアが必要な場合があります。また、スコアリング時に生成することができるその他の予測および確信度値があります。

調整済み傾向スコアを計算 : 未調整傾向スコアは、学習データにのみ基づき、このデータがオーバーフィットする多くのモデルの傾向によって過度に楽観的な場合があります。調整済み傾向は、テストまたは検証用データ区分に対してモデルのパフォーマンスを評価することによって補正しようとします。このオプションでは、モデルの生成前にデータ区分フィールドをストリーム内で定義し、調整済み傾向スコアがモデル作成ノードで有効化されている必要があります。

最近傍の表示 : この値を *n* に設定し、*n* がゼロ以外の正の整数である場合、重要レコードに対する *n* 個の最近傍が、重要レコードからの相対距離とともにモデルに含まれます。

第 17 章 Python ノード

SPSS Modeler には、Python 固有のアルゴリズムを使用するためのノードが用意されています。「ノードパレット」の「**Python**」タブには次のノードがあり、これらを使用して Python のアルゴリズムを実行できます。これらのノードは、Windows 64、Linux64、および Mac でサポートされています。

SMOTE (Synthetic Minority Over-sampling Technique) ノードは不均衡データ・セットを扱うためのオーバーサンプリング・アルゴリズムを提供します。これにより、データの均衡化のための高度な手法が提供されます。SPSS Modeler の SMOTE プロセス ノードは Python で実装されており、`imbalanced-learn`© Python ライブラリーを必要とします。

XGBoost Linear© は、線型モデルを基本モデルとして使用する勾配ブースティング・アルゴリズムの高度な実装です。ブースティング・アルゴリズムでは、弱い分類子に繰り返し学習させ、それを最終的な強い分類子に追加します。SPSS Modeler の XGBoost Linear ノードは Python で実装されています。

XGBoost Tree© は、ツリー モデルを基本モデルとして使用する勾配ブースティング・アルゴリズムの高度な実装です。ブースティング・アルゴリズムでは、弱い分類子に繰り返し学習させ、それを最終的な強い分類子に追加します。XGBoost ツリーは柔軟性が極めて高く、多くのユーザーを圧倒するほどの多数のパラメータが用意されています。このため、SPSS Modeler の XGBoost ツリー・ノードでは、コア・フィーチャーおよびよく使用されるパラメータが公開されています。このノードは Python で実装されています。

t 分布 Stochastic Neighbor Embedding (t-SNE) は、高次元データの視覚化のためのツールです。t-SNE は、データ ポイントの類似性を確率に変換します。SPSS Modeler の t-SNE ノードは Python で実装されており、`scikit-learn`© Python ライブラリーを必要とします。

ガウス混合©モデルは、未知パラメータを持つ有限個数のガウス分布の混合からすべてのデータ ポイントが生成されると仮定する確率モデルです。混合モデルは、データの共分散構造および潜在ガウス分布の中心に関する情報を取り込むための一般化 K-means クラスタリングと考えることができます。SPSS Modeler のガウス混合ノードは、ガウス混合ライブラリーのコア機能およびよく使用されるパラメータを公開します。このノードは Python で実装されています。

カーネル密度推定 (KDE)© は、Ball Tree または KD Tree のアルゴリズムを使用してクエリを効率化し、教師なし学習、特徴量エンジニアリング、データのモデル化の概念を結合します。KDE などの近隣ベースの手法が、最もよく使用され、有用な密度推定手法です。SPSS Modeler の KDE モデル作成および KDE シミュレーションのノードは、KDE ライブラリーのコア機能およびよく使用されるパラメータを公開します。これらのノードは Python で実装されています。

ランダム・フォレスト・ノードは、ツリー・モデルを基本モデルとして使用するバギング・アルゴリズムの高度な実装を使用します。SPSS Modeler のランダム フォレスト モデル作成ノードは Python で実装されており、scikit-learn© Python ライブラリーを必要とします。

Hierarchical Density-Based Spatial Clustering (HDBSCAN)© は、教師なし学習を使用してデータ・セットのクラスター (つまり、密度の高い領域) を検出します。SPSS Modeler の HDBSCAN ノードは、HDBSCAN ライブラリーのコア機能およびよく使用されるパラメーターを公開します。このノードは Python で実装されており、最初にグループの性質が分からない場合にデータ・セットを異なるグループにクラスター化するために使用できます。

One-Class SVM ノードでは、教師なし学習アルゴリズムを使用します。このノードは、新規性検知の目的で使用できます。このノードは、与えられたサンプル・セットのソフト境界を検知し、新規ポイントがこのセットに属するか、属さないかを分類します。SPSS Modeler の One-Class SVM モデル作成ノードは Python で実装されており、scikit-learn© Python ライブラリーを必要とします。

SMOTE ノード

SMOTE (Synthetic Minority Over-sampling Technique) ノードは不均衡データ・セットを扱うためのオーバーサンプリング・アルゴリズムを提供します。これにより、データの均衡化のための高度な手法が提供されます。SMOTE プロセス ノードは Python で実装されており、imbalanced-learn© Python ライブラリーを必要とします。imbalanced-learn ライブラリーについて詳しくは、http://contrib.scikit-learn.org/imbalanced-learn/about.html[1] を参照してください。

「ノード パレット」の「Python」タブには、SMOTE ノードおよびその他の Python ノードがあります。

[1]Lemaître, Nogueira, Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research*, vol. 18, no. 17, 2017, pp. 1-5. (http://jmlr.org/papers/v18/16-365.html)

SMOTE ノードの設定

SMOTE ノードの「設定」タブで、次の設定を定義します。

対象の設定

対象フィールド**:** 対象フィールドを選択します。すべての尺度タイプ (フラグ型、名義型、順序型、および離散型) がサポートされます。「データ区分」セクションで「データ区分データを使用」オプションを選択した場合、学習データのみがオーバーサンプリングされます。

オーバーサンプリング率

「自動」を選択してオーバーサンプリング率を自動的に選択するか、「比率 **(**マジョリティーに対するマイノリティーの比率**)** の設定」を選択してカスタムの比率の値を選択します。これは、マジョリティー クラスのサンプル数に対するマイノリティー クラスのサンプル数の比率です。値は **0** より大きく **1** 以下でなければなりません。

ランダム シード

ランダム シードの設定**:** 乱数発生ルーチンによって使用されるシードを生成するには、このオプションを選択し、「生成」をクリックします。

方法

アルゴリズムの種類**:** 使用する SMOTE アルゴリズムの種類を選択します。

サンプルのルール

K Neighbours: 合成サンプルを作成するために使用する最近傍の数を指定します。

M Neighbours: マイノリティー サンプルが危険な状況にあるかをどうか判断するために使用する最近傍の数を指定します。これは、SMOTE アルゴリズムの種類「**Borderline1**」または「**Borderline2**」が選択された場合にのみ使用されます。

データ区分

データ区分データを使用。 学習データのオーバーサンプリングのみを行う場合は、このオプションを選択します。

SMOTE ノードは、`imbalanced-learn`© Python ライブラリーを必要とします。次の表に、SPSS Modeler の SMOTE ノードのダイアログの設定と Python アルゴリズムとの間の関係を示します。

表 *33.* ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**Python API** パラメータ名
オーバーサンプリング率 (数値の入力コントロール)	`sample_ratio_value`	`ratio`
ランダム シード	`random_seed`	`random_state`
K_Neighbours	`k_neighbours`	`k`
M_Neighbours	`m_neighbours`	`m`
アルゴリズムの種類	`algorithm_kind`	`kind`

XGBoost Linear ノード

XGBoost Linear© は、線型モデルを基本モデルとして使用する勾配ブースティング・アルゴリズムの高度な実装です。ブースティング・アルゴリズムでは、弱い分類子に繰り返し学習させ、それを最終的な強い分類子に追加します。SPSS Modeler の XGBoost Linear ノードは Python で実装されています。

ブースティング・アルゴリズムについて詳しくは、XGBoost のチュートリアル (http://xgboost.readthedocs.io/en/latest/tutorials/index.html) を参照してください。[1]

XGBoost の交差検証機能は、SPSS Modeler ではサポートされていません。この機能の代わりに、SPSS Modeler のデータ区分ノードを使用できます。SPSS Modeler の XGBoost では、カテゴリー変数の場合にワン・ホット・エンコーディングが自動的に実行されることにも留意してください。

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

XGBoost Linear ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

定義済みの役割を使用**:** 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: 対象および予測値を手動で割り当てる場合は、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側の「目標」役割フィールドおよび「予測値」役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

目標: この予測の目標として使用するフィールドを選択します。

予測値: 1 つ以上のフィールドを予測の入力として選択します。

XGBoost Linear ノードの作成オプション

XGBoost Linear ノードの作成オプションを指定するには、「作成オプション」タブを使用します。このタブには、線型ブースティングのパラメータやモデルの構築などの基本オプションや、目的のための学習タスク・オプションがあります。これらのオプションについて詳しくは、以下のオンライン情報源を参照してください。

- XGBoost のパラメータの解説[1]
- XGBoost Python API[2]
- XGBoost ホーム ページ[3]

基本

ハイパーパラメータ最適化 **(Rbfopt** に基づく**)**。Rbfopt に基づくハイパーパラメータ最適化を有効にするには、このオプションを選択します。有効にすると、パラメータの最適な組み合わせが自動的に検出され、サンプルに対するモデルの誤差率が予測値以下になります。Rbfopt について詳しくは、http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.htmlを参照してください。

アルファ**:** 重みに関する L1 正規化項。この値を大きくすると、モデルがより保守的になります。

ラムダ**:** 重みに関する L2 正規化項。この値を大きくすると、モデルがより保守的になります。

ラムダ バイアス**:** 偏りに関する L2 正規化項。(偏りに関する L1 正規化項は重要ではないため、存在しません。)

Number boost round。ブースティング反復回数。

学習タスク

目的**:** 学習タスクの目的タイプ **reg:linear**、**reg:logistic**、**reg:gamma**、**reg:tweedie**、**count:poisson**、**rank:pairwise**、**binary:logistic**、または **multi** から選択します。

ランダム シード**:** 「生成」をクリックすると、乱数発生ルーチンによって使用されるシードを生成できます。

次の表に、SPSS Modeler の XGBoost Linear ノードのダイアログの設定と Python XGBoost ライブラリーのパラメータとの間の関係を示します。

表 34. ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**XGBoost** パラメータ
対象	TargetField	
予測値	InputFields	
ラムダ	lambda	lambda
アルファ	alpha	alpha
ラムダ バイアス	lambdaBias	lambda_bias
Num boost round	numBoostRound	num_boost_round
目的	objectiveType	objective
ランダム シード	random_seed	seed

[1] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[2] "Plotting API" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

XGBoost Linear ノードのモデル オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

XGBoost ツリー・ノード

XGBoost Tree© は、ツリー モデルを基本モデルとして使用する勾配ブースティング・アルゴリズムの高度な実装です。ブースティング・アルゴリズムでは、弱い分類子に繰り返し学習させ、それを最終的な強い分類子に追加します。XGBoost ツリーは柔軟性が極めて高く、多くのユーザーを圧倒するほどの多数のパラメータが用意されています。このため、SPSS Modeler の XGBoost ツリー・ノードでは、コア・フィーチャーおよびよく使用されるパラメータが公開されています。このノードは Python で実装されています。

ブースティング・アルゴリズムについて詳しくは、XGBoost のチュートリアル (http://xgboost.readthedocs.io/en/latest/tutorials/index.html) を参照してください。[1]

XGBoost の交差検証機能は、SPSS Modeler ではサポートされていません。この機能の代わりに、SPSS Modeler のデータ区分ノードを使用できます。SPSS Modeler の XGBoost では、カテゴリー変数の場合にワン・ホット・エンコーディングが自動的に実行されることにも留意してください。

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

XGBoost ツリー・ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

定義済みの役割を使用**:** 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: 対象および予測値を手動で割り当てる場合は、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側の「目標」役割フィールドおよび「予測値」役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

目標: この予測の目標として使用するフィールドを選択します。

予測値: 1 つ以上のフィールドを予測の入力として選択します。

XGBoost ツリー・ノードの作成オプション

XGBoost ツリー・ノードの作成オプションを指定するには、「作成オプション」タブを使用します。このタブには、モデルの構築やツリーの成長のための基本オプションや、目的のための学習タスク・オプション、およびオーバーフィッティングの制御や不均衡データ・セットの処理のための詳細オプションが含まれています。これらのオプションについて詳しくは、以下のオンライン情報源を参照してください。

- XGBoost のパラメータの解説[1]
- XGBoost Python API[2]
- XGBoost ホーム ページ[3]

基本

ハイパーパラメータ最適化 **(Rbfopt** に基づく**):** Rbfopt に基づくハイパーパラメータ最適化を有効にするには、このオプションを選択します。有効にすると、パラメータの最適な組み合わせが自動的に検出され、サンプルに対するモデルの誤差率が予測値以下になります。Rbfopt について詳しくは、http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.htmlを参照してください。

Tree method: 使用する XGBoost ツリー構築アルゴリズムを選択します。

Num boost round: ブースティング反復回数を指定します。

Max depth: ツリーの最大深度を指定します。この値を大きくすると、モデルがより複雑になり、オーバーフィッティングになりやすくなります。

Min child weight: 子で必要なインスタンスの重み (ヘシアン) の最小合計を指定します。ツリーの分割ステップで生じた葉ノードのインスタンスの重みの合計が、この「**Min child weight**」より小さい場合、構築プロセスでそれ以上の分割は行いません。線型モードの場合は、単純に、各ノードに必要なインスタンスの最小値に対応しています。重みを大きくするほど、アルゴリズムがより保守的になります。

Max delta step: 各ツリーの重みを推定できるようにするには、Max delta step を指定します。**0** に設定した場合、何も制約はありません。正の値に設定した場合、更新ステップがより保守的になります。通常はこのパラメータは必要ありませんが、クラスのバランスが著しく悪いときに、ロジスティック回帰で役立つ可能性があります。

学習タスク

目的**:** 学習タスクの目的タイプ **reg:linear**、**reg:logistic**、**reg:gamma**、**reg:tweedie**、**count:poisson**、**rank:pairwise**、**binary:logistic**、または **multi** から選択します。

早期停止**:** 早期停止機能を使用する場合は、このオプションを選択します。「停止ラウンド **(stopping rounds)**」の場合、学習を続行するには、少なくとも早期停止ラウンドごとに検証エラーが減少する必要があります。「評価データの比率 **(Evaluation data ratio)**」は、検証エラーに使用する入力データの比率です。

ランダム シード**:** 「生成」をクリックすると、乱数発生ルーチンによって使用されるシードを生成できます。

詳細

サブサンプル**:** サブサンプルは、学習インスタンスの比率を示します。例えば、**0.5** に設定した場合、データ・インスタンスの半数をランダムに収集してツリーを成長させることで、オーバーフィッティングを防ぎます。

イータ**:** オーバーフィッティングを防ぐために更新ステップ中に使用するステップ サイズの収縮。各ブースティング ステップの後で、新規フィーチャーの重みを直接取得できます。イータはフィーチャーの重みも収縮させるため、ブースティングのプロセスがより保守的になります。

ガンマ**:** ツリーの葉ノードをさらに分割するために必要な最小の損失低減。ガンマの設定を大きくするほど、アルゴリズムがより保守的になります。

Colsample by tree: 各ツリーを構築する際の、列のサブサンプルの比率。

Colsample by level: 各分割における、各レベルでの列のサブサンプルの比率。

ラムダ**:** 重みに関する L2 正規化項。この値を大きくすると、モデルがより保守的になります。

アルファ**:** 重みに関する L1 正規化項。この値を大きくすると、モデルがより保守的になります。

Scale pos weight: 正の重みと負の重みのバランスを制御します。これは不均衡なクラスの場合に役立ちます。

次の表に、SPSS Modeler の XGBoost ツリー・ノードのダイアログの設定と Python XGBoost ライブラリーのパラメータとの間の関係を示します。

表 *35.* ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**XGBoost** パラメータ
対象	`TargetField`	
予測値	`InputFields`	
Tree method	`treeMethod`	`tree_method`
Num boost round	`numBoostRound`	`num_boost_round`
Max depth	`maxDepth`	`max_depth`
Min child weight	`minChildWeight`	`min_child_weight`
Max delta step	`maxDeltaStep`	`max_delta_step`
目的	`objectiveType`	`objective`
早期停止	`earlyStopping`	`early_stopping_rounds`
停止ラウンド (stopping rounds)	`stoppingRounds`	
評価データの比率 (Evaluation data ratio)	`evaluationDataRatio`	
ランダム シード	`random_seed`	`seed`

表 35. ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング (続き)

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**XGBoost** パラメータ
サブサンプル	sampleSize	subsample
イータ	eta	eta
ガンマ	gamma	gamma
Colsample by tree:	colsSampleRatio	colsample_bytree
Colsample by level	colsSampleLevel	colsample_bylevel
ラムダ	lambda	lambda
アルファ	alpha	alpha
Scale pos weight	scalePosWeight	scale_pos_weight

[1] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[2] "Plotting API" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

XGBoost ツリー・ノードのモデル・オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

---

t-SNE ノード

t 分布 Stochastic Neighbor Embedding (t-SNE)© は、高次元データの視覚化のためのツールです。t-SNE は、データ ポイントの類似性を確率に変換します。元の空間の類似性はガウス ジョイント確率によって表され、埋め込み空間の類似性はスチューデントの t 分布によって表されます。これにより、t-SNE はローカル構造に特に依存します。また、既存の技術に勝る利点が他にもいくつかあります。[1]

- 単一マップ上で、さまざまな尺度で構造を表示する
- 複数の異なる多様体またはクラスターに存在するデータを表示する
- 中央にポイントを集中させる傾向を削減する

SPSS Modeler の t-SNE ノードは Python で実装されており、scikit-learn© Python ライブラリーを必要とします。t-SNE および scikit-learn ライブラリーについて詳しくは、以下を参照してください。

- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE
- https://scikit-learn.org/stable/modules/manifold.html#t-sne

「ノード パレット」の「Python」タブには、このノードおよびその他の Python ノードがあります。また、「グラフ」タブには、t-SNE ノードがあります。

[1] 参照資料

van der Maaten, L.J.P.; Hinton, G. "Visualizing High-Dimensional Data using t-SNE". Journal of Machine Learning Research. 9:2579-2605, 2008.

van der Maaten, L.J.P. "t-Distributed Stochastic Neighbor Embedding".

van der Maaten, L.J.P. "Accelerating t-SNE using Tree-Based Algorithms". Journal of Machine Learning Research. 15(Oct):3221-3245, 2014.

t-SNE ノードのエキスパート オプション

t-SNE ノードにどちらのオプションを設定するかに応じて、「シンプル」モードまたは「エキスパート」モードを選択します。

視覚化タイプ: 「**2** 次元」または「**3** 次元」を選択して、2 次元または 3 次元のどちらでグラフを描画するかを指定します。

方法: 「**Barnes Hut**」または「正確確率」を選択します。デフォルトでは、勾配計算アルゴリズムは、正確確率メソッドより確実に速く実行される Barnes-Hut 近似を使用します。Barnes-Hut 近似により、t-SNE 技術を実際の大規模データ セットに適用できます。正確確率アルゴリズムは、最近隣のエラーを回避する上で、より優れています。

初期化 **(Init.):** 埋め込みの初期化について、「無作為」または「**PCA**」を選択します。

対象フィールド**:** 出力グラフのカラー マップとして表示する対象フィールドを選択します。ここで対象フィールドを指定しないと、グラフには 1 色が使用されます。

最適化

Perplexity: Perplexity は、他の多様体学習アルゴリズムで使用される最近隣の数に関連します。通常、データ セットが大きいほど、必要とされる Perplexity も大きくなります。**5** から **50** の間の値を選択することを考慮してください。デフォルトは **30**、範囲は **2** から **9999999** です。

Early exaggeration: この設定は、埋め込み空間における、元の空間の自然クラスターの気密度、およびクラスター間の間隔を制御します。デフォルトは **12**、範囲は **2** から **9999999** です。

学習率**:** 学習率が高すぎる場合、データは、すべてのポイントがその最近傍からほぼ等距離にある 1 つの「ボール」のように表示されることがあります。学習率が低すぎる場合、大部分のポイントは、圧縮された厚い雲のように表示されることがあります。外れ値はほとんどなくなります。誤った局所最小値でコスト関数が停止している場合は、学習率を高くすると改善することがあります。デフォルトは **200**、範囲は **0** から **9999999** です。

最大反復**:** 最適化の最大反復数。デフォルトは **1000**、範囲は **250** から **9999999** です。

角サイズ**:** あるポイントから測定した遠方ノードの角サイズ。**0** から **1** の間の値を入力します。デフォルトは **0.5** です。

ランダム シード

ランダム シードの設定**:** 乱数発生ルーチンによって使用されるシードを生成するには、このオプションを選択し、「生成」をクリックします。

最適化の中断条件

進捗のない最大反復**:** 最適化を中断するまでに実行する、進捗のない反復の最大数。Early exaggeration を伴う 250 回の初期反復の後に使用されます。進捗は 50 回の反復ごとにしか検査されないため、この値は次の 50 の倍数に丸められることに注意してください。デフォルトは **300**、範囲は **0** から **9999999** です。

最小勾配ノルム **(Min gradient norm):** 勾配ノルムがこの最小しきい値を下回る場合、最適化は中断されます。デフォルトは **1.0E-7** です。

メトリック。 機能配列内のインスタンス間の距離を計算するときに使用するメトリック。メトリックが文字列である場合、メトリックは、scipy.spatial.distance.pdist のメトリック パラメータとして許可されているオプションの 1 つであるか、pairwise.PAIRWISE_DISTANCE_FUNCTIONS にリストされているメトリックである必要があります。使用可能ないずれかのメトリック タイプを選択します。デフォルトは **euclidean** です。

レコード数が次の値より大きい場合**:** 大規模データ・セットの作図の手法を指定します。最大データ・セット・サイズを使用するか、またはデフォルト値 (2,000 ポイント) を使用することができます。「ビン」 または 「サンプル」 を選択すると、大きいデータ・セットに対するパフォーマンスが向上します。代わりに、「すべてのデータを使用」 を選択して、すべてのデータ・ポイントを作図することもできます。ただし、この場合ソフトウェアのパフォーマンスが大幅に低下する可能性があります。

- ビン**:** データ・セットに格納されているレコード数が、指定した数より大きい場合に、分割を有効にします。分割を行うと、グラフが細かいグリッドに分割されてから、作図や各グリッド・セルに現れる接続数のカウントが実際に行われます。最終的なグラフでは、ビン重心 (ビン中のすべての接続の位置の平均) でセルごとに 1 つの接続が作図されます。
- サンプリング**:** ここに指定した数のレコードまで、無作為にデータのサンプリングを行います。

次の表に、SPSS Modeler t-SNE ノードのダイアログの「エキスパート」タブの設定と、Python t-SNE ライブラリーのパラメータとの間の関係を示します。

表 *36.* ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**Python t-SNE** パラメータ
モード	mode_type	
視覚化タイプ	n_components	n_components
メソッド	method	method
埋め込みの初期化 (Initialization of embedding)	init	init
対象	target_field	target_field
Perplexity	perplexity	perplexity
Early exaggeration	early_exaggeration	early_exaggeration
学習率	learning_rate	learning_rate
最大反復	n_iter	n_iter
角サイズ	angle	angle
ランダム シードの設定	enable_random_seed	
ランダム シード	random_seed	random_state
進捗のない最大反復	n_iter_without_progress	n_iter_without_progress
最小勾配ノルム (Min gradient norm)	min_grad_norm	min_grad_norm
複数の Perplexity で t-SNE を実行	isGridSearch	

t-SNE ノードの出力オプション

「出力」タブで、t-SNE ノードの出力オプションを指定します。

出力名。ノードの実行時に生成される出力の名前を指定します。「自動」を選択すると、出力の名前が自動的に設定されます。

画面に出力。出力を生成し、新規ウィンドウに表示するには、このオプションを選択します。出力は、出力マネージャーにも追加されます。

ファイルに出力。出力をファイルに保存するには、このオプションを選択します。選択すると、「ファイル名」フィールドと「ファイルの種類」フィールドが有効になります。比較する目的で他のフィールドを使用するプロットを作成する場合、あるいは分類モデルまたは回帰モデルで出力を予測値として使用する場合、t-SNE ノードはこの出力ファイルにアクセスする必要があります。t-SNE モデルは、固定長ファイル入力ノードを使用すると最も容易にアクセスできる、x、y (および z) 座標フィールドから成る結果ファイルを作成します。

t-SNE モデル ナゲット

t-SNE モデル ナゲットには、t-SNE モデルが取得したすべての情報が含まれます。以下のタブがあります。

Graph

「グラフ」タブには、t-SNE ノードのグラフ出力が表示されます。pyplot 散布図には、低次元の結果が示されます。t-SNE ノードの「エキスパート」タブで「複数の **Perplexity** で **t-SNE** を実行」オプションを選択しなかった場合、異なる Perplexity を使用した 6 つのグラフではなく、1 つのグラフのみが含まれます。

テキスト出力

「テキスト出力」タブには、t-SNE アルゴリズムの結果が表示されます。t-SNE ノードの「エキスパート」タブで、視覚化タイプに「**2** 次元」を選択した場合、ここに表示される結果は 2 次元のポイント値です。「**3** 次元」を選択した場合、結果は 3 次元のポイント値になります。

ガウス混合ノード

ガウス混合©モデルは、未知パラメータを持つ有限個数のガウス分布の混合からすべてのデータ ポイントが生成されると仮定する確率モデルです。混合モデルは、データの共分散構造および潜在ガウス分布の中心に関する情報を取り込むための一般化 K-means クラスタリングと考えることができます。[1]

SPSS Modeler のガウス混合ノードは、ガウス混合ライブラリーのコア機能およびよく使用されるパラメータを公開します。このノードは Python で実装されています。

ガウス混合モデル アルゴリズムおよびパラメータについて詳しくは、http://scikit-learn.org/stable/modules/mixture.html および https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html にあるガウス混合に関する資料を参照してください。[2]

[1] "User Guide." *Gaussian mixture models*. Web. © 2007 - 2017. scikit-learn developers.

[2] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

ガウス混合ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

定義済みの役割を使用**:** このオプションでは、上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の入力設定を使用します。

カスタム・フィールド割り当ての使用: 入力を手動で割り当てる場合は、このオプションを選択します。

フィールド**:** このリストの項目を画面右側の「予測値」リストに手動で割り当てるには、矢印ボタンを使用します。アイコンは、各フィールドの有効な測定の尺度を示します。リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

予測値: 予測値として 1 つ以上のフィールドを選択してください。

ガウス混合ノードの作成オプション

ガウス混合ノードの作成オプションを指定するには、「作成オプション」タブを使用します。このタブには、基本オプションや拡張オプションが用意されています。このセクションで説明していないこれらのオプションの詳細については、以下のオンライン情報源を参照してください。

- ガウス混合パラメータのリファレンス[1]
- ガウス混合ノード ユーザー ガイド[2]

基本

共分散タイプ。以下の分散共分散行列のいずれかを選択します。

- 完全。各コンポーネントが独自の一般的な分散共分散行列を持ちます。
- **Tied**。すべてのコンポーネントが同じ一般的な分散共分散行列を共有します。
- **Diag**。各コンポーネントが独自の対角分散共分散行列を持ちます。
- **Spherical**。各コンポーネントが独自の分散を 1 つずつ持ちます。

コンポーネントの数。モデルの作成時に使用する混合コンポーネントの数を指定します。

クラスター ラベル: クラスター・ラベルが数値なのか文字列なのかを指定します。「文字列」を選択した場合、クラスター・ラベルの接頭辞を指定します (例えば、デフォルト接頭辞は `cluster` であり、**cluster-1**、**cluster-2** といったクラスター・ラベルが作成されることになります)。

ランダム シード**:** 乱数発生ルーチンによって使用されるシードを生成するには、このオプションを選択し、「生成」をクリックします。

詳細

許容度: 収束のしきい値を指定します。デフォルト値は **0.001** です。

反復数**:** これにより、指定された反復数の後にモデル評価の作成を停止できます。実行する反復の最大回数を指定します。デフォルト値は **100** です。

Init パラメータ。初期化パラメータ「**KMeans**」(KMeans を使用して負担率を初期化する) または「無作為」(ランダムに負担率を初期化する) を選択します。

ウォーム スタート。「**True**」を選択すると、最後の適合の解を次の適合の初期値として使用します。これにより、類似した問題に対して何度も適合を呼び出す場合に収束が早くなります。

次の表に、SPSS Modeler のガウス混合ノードのダイアログの設定と Python のガウス混合ライブラリーのパラメータとの間の関係を示します。

表 37. ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	ガウス混合パラメータ
定義済みの役割を使用/ユーザー設定フィールドの割り当てを使用	role_use	
入力	**予測値**	
区分されたデータを使用	use_partition	
共分散タイプ	covariance_type	covariance_type
コンポーネントの数	number_component	n_components
クラスター ラベル	component_lable	
ラベル接頭辞	label_prefix	
ランダム シードの設定	enable_random_seed	
ランダム シード	random_seed	random_state
許容度	tol	tol
反復数	max_iter	max_iter
Init パラメータ	init_params	init_params
ウォーム スタート	warm_start	warm_start

[1] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[2] "User Guide." *Gaussian mixture models*. Web. © 2007 - 2017. scikit-learn developers.

ガウス混合ノードのモデル オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

KDE ノード

カーネル密度推定 (KDE)© は、Ball Tree または KD Tree のアルゴリズムを使用してクエリを効率化し、教師なし学習、特徴量エンジニアリング、データのモデル化の概念を結合します。KDE などの近隣ベースの手法が、最もよく使用され、有用な密度推定手法です。KDE は任意の数の次元で実行できますが、実際のところは、高次元ではパフォーマンスが低下する可能性があります。SPSS Modeler の KDE モデル作成および KDE シミュレーションのノードは、KDE ライブラリーのコア機能およびよく使用されるパラメータを公開します。これらのノードは Python で実装されています。[1]

KDE ノードを使用するには、上流のデータ型ノードをセットアップする必要があります。KDE ノードは、データ型ノード (または上流の入力ノードの「データ型」タブ) の入力値を読み取ります。

「**KDE** モデル作成」ノードは、SPSS Modeler の「モデル作成」タブおよび「Python」タブで使用できます。KDE モデル作成ノードはモデル ナゲットを生成します。ナゲットのスコア値は、入力データからのカーネル密度の値になります。

「**KDE** シミュレーション」ノードは「出力」タブおよび「Python」タブで使用できます。「KDE シミュレーション」ノードは、入力データと同じ分布を持つレコードを作成できる KDE 生成入力ノードを生成

します。KDE 生成ノードには「設定」タブがあり、ノードが作成するレコードの数 (デフォルトでは 1) および生成するランダム シードの数を指定できます。

KDE およびその例について詳しくは、http://scikit-learn.org/stable/modules/density.html#kernel-density-estimation にある KDE の資料を参照してください。[1]

[1] "User Guide." *Kernel Density Estimation*. Web. © 2007-2018, scikit-learn developers.

KDE モデル作成ノードおよび KDE シミュレーション ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

定義済みの役割を使用: このオプションでは、上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の入力設定を使用します。

カスタム・フィールド割り当ての使用: 入力を手動で割り当てる場合は、このオプションを選択します。

フィールド: このリストの項目を画面右側の「入力」リストに手動で割り当てるには、矢印ボタンを使用します。アイコンは、各フィールドの有効な測定の尺度を示します。リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

入力: 1 つ以上のフィールドをクラスター化の入力として選択します。KDE が処理できるのは連続型フィールドのみです。

KDE ノードの作成オプション

KDE ノードの作成オプションを指定するには、「作成オプション」タブを使用します。このタブには、カーネル密度パラメータおよびクラスター ラベルのための基本オプションや、許容度、リーフ サイズ、幅優先の方法を使用するかどうかなどの拡張オプション があります。これらのオプションについて詳しくは、以下のオンライン情報源を参照してください。

- Kernel Density Estimation Python API Parameter Reference[1]
- Kernel Density Estimation User Guide[2]

基本

帯域幅。カーネルの帯域幅を指定します。

カーネル。使用するカーネルを選択します。KDE モデル作成ノードに使用可能なカーネルは「ガウス」、「**Tophat**」、「**Epanechnikov**」、「指数」、「線型」、「コサイン」です。KDE シミュレーション ノードに使用可能なカーネルは「ガウス」および「**Tophat**」です。使用可能なこれらのカーネルについて詳しくは、『Kernel Density Estimation User Guide』[2]を参照してください。

アルゴリズム: 使用するツリー アルゴリズムとして「自動」、「**Ball Tree**」または「**KD Tree**」を選択します。詳しくは、『Ball Tree』[3]および『KD Tree』[4]を参照してください。

メトリック。 距離メトリックを選択します。使用可能なメトリックは「ユークリッド」、「**Braycurtis**」、「チェビシェフ」、「**Canberra**」、「市区町村ブロック」、「**Dice**」、「**Hamming**」、「正の無限方向」、「**Jaccard**」、「**L1**」、「**L2**」、「**Matching**」、「**Manhattan**」、「**P**」、「**Rogerstanimoto**」、「**Russellrao**」、「**Sokalmichener**」、「**Sokalsneath**」、「**Kulsinski**」、「ミンコフスキー」です。「ミンコフスキー」を選択した場合は、必要に応じて「**P** 値」を設定してください。

このドロップダウンで使用できるメトリックは、選択したアルゴリズムに応じて変わります。また、密度出力の正規化は、ユークリッド距離メトリックの場合にのみ正確になることに注意してください。

詳細

絶対許容度。結果の必要な絶対許容度を指定します。許容度が大きいと、一般に実行が高速になります。デフォルトは **0.0** です。

相対許容度。結果の必要な相対許容度を指定します。許容度が大きいと、一般に実行が高速になります。デフォルトは **1E-8** です。

リーフ サイズ: 下位ツリーのリーフ サイズを指定します。デフォルトは **40** です。リーフ サイズを変更すると、パフォーマンスおよびメモリ所要量に大きな影響が及ぶ場合があります。Ball Tree および KD Tree のアルゴリズムについて詳しくは、『Ball Tree』[3]および『KD Tree』[4]を参照してください。

幅優先。幅優先の方法を使用する場合は「**True**」を選択します。深さ優先の方法を使用する場合は「**False**」を選択します。

次の表に、SPSS Modeler の KDE ノードのダイアログの設定と Python KDE ライブラリーのパラメータとの間の関係を示します。

表 *38.* ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**KDE** のパラメータ
入力	inputs	
帯域幅	bandwidth	bandwidth
カーネル	kernel	kernel
アルゴリズム	algorithm	algorithm
メトリック	metric	metric
P 値	pValue	pValue
絶対許容度	atol	atol
相対許容度	rtol	Rtol
リーフ サイズ	leafSize	leafSize
幅優先	breadthFirst	breadthFirst

[1] "API Reference." *sklearn.neighbors.KernelDensity*. Web. © 2007-2018, scikit-learn developers.

[2] "User Guide." *Kernel Density Estimation*. Web. © 2007-2018, scikit-learn developers.

[3] "Ball Tree." *Five balltree construction algorithms*. © 1989, Omohundro, S.M., International Computer Science Institute Technical Report.

[4] "K-D Tree." *Multidimensional binary search trees used for associative searching*. © 1975, Bentley, J.L., Communications of the ACM.

KDE モデル作成ノードおよび KDE シミュレーション ノードのモデル オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

ランダム・フォレスト・ノード

ランダム フォレスト© は、ツリー モデルを基本モデルとして使用するバギング アルゴリズムの高度な実装です。ランダム フォレストでは、置き換えを行って学習セットから抽出したサンプル (ブートストラップ サンプルなど) からアンサンブルの各ツリーを構築します。ツリーの構築中にノードを分割するとき、選択された分割がすべてのフィーチャーの間で最良の分割になるわけではありません。その代わり、選択された分割は、フィーチャーのランダムなサブセットの間で最良の分割になっています。このようにランダム性があるため、通常は (単一の非 Random Trees の偏りについて) フォレストの偏りがやや増えますが、平均化により分散も減少するため (通常は偏りの増加の補正より減少が大きくなります)、全体としてよりよいモデルが得られます。[1]

SPSS Modeler のランダム・フォレスト・ノードは Python で実装されています。「ノード パレット」の「Python」タブには、このノードおよびその他の Python ノードがあります。

ランダム フォレストのアルゴリズムについて詳しくは、https://scikit-learn.org/stable/modules/ensemble.html#forestを参照してください。

[1]L. Breiman, "Random Forests," Machine Learning, 45(1), 5-32, 2001.

ランダム・フォレスト・ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

定義済みの役割を使用**:** 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: 対象および予測値を手動で割り当てる場合は、このオプションを選択します。

フィールド**:** 矢印ボタンを使用して、このリストの項目を画面右側の「目標」役割フィールドおよび「予測値」役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

目標: この予測の目標として使用するフィールドを選択します。

予測値: 1 つ以上のフィールドを予測の入力として選択します。

ランダム・フォレスト・ノードの作成オプション

ランダム・フォレスト・ノードの作成オプションを指定するには、「作成オプション」タブを使用します。このタブには、基本オプションや拡張オプションが用意されています。これらのオプションの詳細については、https://scikit-learn.org/stable/modules/ensemble.html#forestを参照してください。

基本

作成するツリーの数**:** フォレスト内のツリー数を選択します。

最大の深さの指定**:** 選択されていない場合、リーフがすべて純粋なリーフになるまで、またはすべてのリーフのサンプル数が `min_samples_split` 未満になるまでノードが展開されます。

Max depth: ツリーの最大の深さ。

リーフ ノードの最小サイズ**:** リーフ ノードで必要なサンプルの最小数。

分割に使用するフィーチャー数**:** 最良の分割を求めるときに考慮するフィーチャーの数。

- auto の場合、分類に max_features=sqrt(n_features) を使用し、回帰に max_features=sqrt (n_features) を使用します。
- sqrt の場合は max_features=sqrt(n_features) です。
- log2 の場合は max_features=log2 (n_features) です。

詳細

ツリーの作成時にブートストラップ サンプルを使用**:** 選択すると、ツリーの作成時にブートストラップ サンプルを使用します。

一般化の精度を推定するために **Out of Bag** サンプルを使用**:** 選択すると、Out of Bag サンプルを使用して一般化の精度を推定します。

Extremely Randomized Trees を使用**:** 選択すると、一般的なランダム・フォレストではなく、Extremely Randomized Trees を使用します。Extremely Randomized Trees では、分割の計算におけるランダム性が一段階向上します。ランダム フォレストと同様に、候補フィーチャーのランダムなサブセットが使用されますが、最も明確に区分するしきい値が探索されるのではなく、各候補フィーチャーに対してランダムにしきい値が設けられ、それらのランダムに生成されたしきい値のうち最良のものが分割ルールとして選択されます。これにより、通常はモデルの分散を少し抑えることができますが、偏りはわずかに大きくなります。[1]

結果の再現: 選択すると、モデル作成処理が再現され、同じスコアリング結果が得られます。

ランダム シード。「生成」をクリックすると、乱数発生ルーチンによって使用されるシードを生成できます。

ハイパーパラメータ最適化 **(Rbfopt** に基づく**):** Rbfopt に基づくハイパーパラメータ最適化を有効にするには、このオプションを選択します。有効にすると、パラメータの最適な組み合わせが自動的に検出され、サンプルに対するモデルの誤差率が予測値以下になります。Rbfopt について詳しくは、http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.htmlを参照してください。

目標: 目標とする目的関数 (サンプルに対するモデルの誤差率) の値 (未知の最適条件の値)。許容できる値 (0.01 など) に設定してください。

最大反復**:** モデルを試行する最大反復数。デフォルトは 1000 です。

最大評価**:** 精度モードにおいてモデルを試行するための関数評価の最大回数。デフォルトは 300 です。

次の表に、SPSS Modeler のランダム・フォレスト・ノードのダイアログの設定と Python ランダム・フォレスト・ライブラリーのパラメータとの間の関係を示します。

表 *39.* ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	ランダム フォレストのパラメータ
対象	target	
予測値	inputs	
作成するツリーの数	n_estimators	n_estimators
最大の深さの指定	specify_max_depth	specify_max_depth

表 39. ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング (続き)

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	ランダム フォレストのパラメータ
Max depth	max_depth	max_depth
リーフ ノードの最小サイズ	min_samples_leaf	min_samples_leaf
分割に使用するフィーチャー数	max_features	max_features
ツリーの作成時にブートストラップ サンプルを使用	bootstrap	bootstrap
一般化の精度を推定するために Out of Bag サンプルを使用	oob_score	oob_score
Extremely Randomized Trees を使用	extreme	
結果を再現	use_random_seed	
ランダム シード	random_seed	random_seed
ハイパーパラメータ最適化 (Rbfopt に基づく)	enable_hpo	
目標 (HPO の場合)	target_objval	
最大反復 (HPO の場合)	max_iterations	
最大評価 (HPO の場合)	max_evaluations	

[1]L. Breiman, "Random Forests," Machine Learning, 45(1), 5-32, 2001.

ランダム・フォレスト・ノードのモデル・オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

ランダム フォレスト モデル ナゲット

ランダム フォレスト モデル ナゲットには、ランダム フォレスト モデルによって取り込まれたすべての情報が含まれます。以下のセクションがあります。

モデル情報

このビューには、入力フィールド、ワン ホット エンコード値、およびモデル パラメーターを含む、モデルについての重要な情報が提供されます。

予測変数の重要度

このビューには、モデルを推定する際の各予測値の相対重要度を示すグラフが表示されます。詳しくは、44 ページの『予測変数の重要度』を参照してください。

HDBSCAN ノード

Hierarchical Density-Based Spatial Clustering (HDBSCAN)© は、教師なし学習を使用してデータ・セットのクラスター (つまり、密度の高い領域) を検出します。SPSS Modeler の HDBSCAN ノードは、HDBSCAN ライブラリーのコア機能およびよく使用されるパラメーターを公開します。このノードは Python で実装されており、最初にグループの性質が分からない場合にデータ・セットを異なるグループにクラスター化するために使用できます。他の SPSS Modeler の学習方法とは異なり、HDBSCAN モデルは対象フィールドを使用しません。このタイプの学習は、対象フィールドがないことから、教師なし学習と呼

ばれます。HDBSCAN では、結果が予測されるのではなく、一連の入力フィールドのパターンが明らかにされます。レコードは、1 つのグループまたはクラスター内のレコード同士がよく似た特性を持ち、異なるグループのレコードが互いに類似しないように分類されます。HDBSCAN アルゴリズムは、クラスターを、密度が低い領域とは異なる密度が高い領域と見なします。このかなり汎用的な考え方のため、HDBSCAN によって検出されたクラスターは、任意の形状を取り得ます。これは、クラスターが凸状であると想定する K-Means とは異なります。低密度領域に単独で存在している外れ値点もマークされます。HDBSCAN では、新規サンプルのスコアリングもサポートされます。[1]

HDBSCAN ノードを使用するには、上流のデータ型ノードをセットアップする必要があります。HDBSCAN ノードは、データ型ノード (または上流の入力ノードの「データ型」タブ) の入力値を読み取ります。

HDBSCAN クラスター化アルゴリズムについて詳しくは、http://hdbscan.readthedocs.io/en/latest/ にある HDBSCAN の資料を参照してください。[1]

[1] "User Guide / Tutorial." *The hdbscan Clustering Library*. Web. © 2016, Leland McInnes, John Healy, Steve Astels.

HDBSCAN ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

重要**:** HDBSCAN モデルの学習には、役割が「入力」に設定された 1 つ以上のフィールドを使用する必要があります。出力、両方、またはなしが役割に設定されたフィールドは無視されます。

定義済みの役割を使用**:** このオプションでは、上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の入力設定を使用します。

カスタム・フィールド割り当ての使用: 入力を手動で割り当てる場合は、このオプションを選択します。

フィールド**:** このリストの項目を画面右側の「入力」リストに手動で割り当てるには、矢印ボタンを使用します。アイコンは、各フィールドの有効な測定の尺度を示します。リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

入力**:** 1 つ以上のフィールドをクラスター化の入力として選択します。

HDBSCAN ノードの作成オプション

HDBSCAN ノードの作成オプションを指定するには、「作成オプション」タブを使用します。このタブには、クラスター・パラメーターおよびクラスター・ラベルのための基本オプションや、拡張パラメーターおよびグラフ出力オプションのための拡張オプションがあります。これらのオプションについて詳しくは、以下のオンライン情報源を参照してください。

- HDBSCAN Python API のパラメーターの解説[1]
- HDBSCAN のホーム・ページ[2]

基本

ハイパーパラメータ最適化 **(Rbfopt** に基づく**):** Rbfopt に基づくハイパーパラメータ最適化を有効にするには、このオプションを選択します。有効にすると、パラメータの最適な組み合わせが自動的に検出され、サンプルに対するモデルの誤差率が予測値以下になります。Rbfopt について詳しくは、http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.htmlを参照してください。

最小クラスター サイズ: クラスターの最小サイズを指定します。ここで指定された値より少ないポイントが含まれた単結合分割が、クラスターの「範囲外の」ポイントと見なされます (クラスターを 2 つの新しいクラスターに分割するのではない)。

最小サンプル数: ポイントをコア・ポイントと見なす、隣接内のサンプルの最小数を指定します。**0** に設定した場合、デフォルト値は最小クラスター サイズの値になります。

アルゴリズム: 使用するアルゴリズムを選択します。HDBSCAN には、データの各種特性に特化したバリアントがあります。デフォルトでは、「最良」が使用され、データの性質に基づいて最適なアルゴリズムが自動的に選択されます。これらのアルゴリズム・タイプについて詳しくは、HDBSCAN の資料を参照してください。[1] 選択するアルゴリズムはパフォーマンスに影響を与えることに注意してください。例えば、大規模なデータの場合、Boruvka KDTree または Boruvka BallTree を試すことをお勧めします。

距離のメトリック: 機能配列内のインスタンス間の距離を計算するときに使用するメトリックを選択します。

クラスター ラベル: クラスター・ラベルが数値なのか文字列なのかを指定します。「文字列」を選択した場合、クラスター・ラベルの接頭辞を指定します (例えば、デフォルト接頭辞は `cluster` であり、**cluster-1**、**cluster-2** といったクラスター・ラベルが作成されることになります)。

詳細

近似最小スパンニング ツリー: 近似最小スパンニング・ツリーを受け入れる場合、「**True**」を選択します。一部のアルゴリズムでは、これにより、パフォーマンスを改善させることができますが、結果のクラスター化の品質が若干低下することがあります。速度を犠牲にし、正確さを重視する場合は、「**False**」オプションを試すことができます。ほとんどの場合、「**True**」にすることをお勧めします。

クラスターの選択方式: 圧縮ツリーからクラスターを選択するために使用する方法を選択します。HDBSCAN での標準の方法では、Excess of Mass (**EOM**) アルゴリズムを使用して、最も持続的なクラスターを検出します。あるいは、ツリーのリーフでクラスターを選択することもできます。これにより、最も微細化された均一なクラスターが得られます。

単一クラスターの受け入れ: この設定を「**True**」に変更すると、単一のクラスターの結果が許可されます (データ・セットで有効な結果である場合のみ)。

P 値: ミンコフスキー距離メトリックを使用する場合 (「基本」作成オプション下)、必要であれば、この p 値を変更できます。

リーフ サイズ: スペース・ツリー・アルゴリズム (Boruvka KDTree または Boruvka BallTree) を使用する場合、これが、ツリーのリーフ・ノードのポイント数になります。この設定によって結果のクラスター化は変更されませんが、アルゴリズムの実行時間が影響を受けることがあります。

妥当性インデックス: このオプションを選択すると、妥当性インデックス・グラフがモデル・ナゲット出力に含まれます。

圧縮ツリー: このオプションを選択すると、圧縮ツリー・グラフがモデル・ナゲット出力に含まれます。

単結合ツリー: このオプションを選択すると、単結合ツリー・グラフがモデル・ナゲット出力に含まれます。

最小スパンニング ツリー: このオプションを選択すると、最小スパンニング・ツリー・グラフがモデル・ナゲット出力に含まれます。

次の表に、SPSS Modeler の HDBSCAN ノードのダイアログの設定と Python HDBSCAN ライブラリーのパラメーターとの間の関係を示します。

表 40. ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**HDBSCAN** パラメーター
入力	inputs	inputs
ハイパーパラメータ最適化	useHPO	
最小クラスター サイズ	min_cluster_size	min_cluster_size
最小サンプル数	min_samples	min_samples
アルゴリズム	algorithm	algorithm
距離のメトリック	metric	metric
クラスター ラベル	useStringLabel	
ラベル接頭辞	stringLabelPrefix	
近似最小スパンニング ツリー	approx_min_span_tree	approx_min_span_tree
クラスターの選択方式	cluster_selection_method	cluster_selection_method
単一クラスターの受け入れ	allow_single_cluster	allow_single_cluster
P 値	p_value	p_value
リーフ サイズ	leaf_size	leaf_size
妥当性インデックス	outputValidity	
圧縮ツリー	outputCondensed	
単結合ツリー	outputSingleLinkage	
最小スパンニング ツリー	outputMinSpan	

[1] "API Reference." *The hdbscan Clustering Library*. Web. © 2016, Leland McInnes, John Healy, Steve Astels.

[2] "User Guide / Tutorial." *The hdbscan Clustering Library*. Web. © 2016, Leland McInnes, John Healy, Steve Astels.

HDBSCAN ノードのモデル・オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

One-Class SVM ノード

One-Class SVM© ノードでは、教師なし学習アルゴリズムを使用します。このノードは、新規性検知の目的で使用できます。このノードは、与えられたサンプル・セットのソフト境界を検知し、新規ポイントがこのセットに属するか、属さないかを分類します。One-Class SVM モデル作成ノードは Python で実装されており、scikit-learn© Python ライブラリーを必要とします。scikit-learn ライブラリーについて詳しくは、http://contrib.scikit-learn.org/imbalanced-learn/about.html[1] を参照してください。

「ノード パレット」の「Python」タブには、One-Class SVM ノードおよびその他の Python ノードがあります。

注**:** One-Class SVM は、教師なし学習外れ値検知および新規性検知に使用します。多くの場合は、アルゴリズムによって所与のサンプルの正しい境界を設定できるように、既知の「標準」のデータ・セットを使用

してモデルを作成することをお勧めします。モデルのパラメータ (nu、gamma、kernel など) が、結果に大きく影響します。したがって、状況に最適の設定が見つかるまで、これらのオプションを試してみることが必要になる可能性があります。

[1]Smola, Schölkopf. "A Tutorial on Support Vector Regression." *Statistics and Computing Archive*, vol. 14, no. 3, August 2004, pp. 199-222. (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.114.4288)

One-Class SVM ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

定義済みの役割を使用**:** 入力の定義済みの役割が設定されたすべてのフィールドを選択するには、このオプションを選択します。

カスタム・フィールド割り当ての使用: フィールドを手動で選択するには、このオプションを選択し、入力フィールドおよび分割フィールドを選択します。

入力**:** 分析で使用する入力フィールドを選択します。データ型不明および不明以外のすべてのストレージ タイプおよび尺度タイプがサポートされます。フィールドのストレージ タイプが文字列の場合、このフィールドの値はワン・ホット・エンコーディング・アルゴリズムを使用して 1 対その他 (one-vs-all) の形式で二値化されます。

分割**:** 分割フィールドとして使用するフィールド (複数可) を選択します。すべての尺度タイプ (フラグ型、名義型、順序型、および離散型) がサポートされます。

データ区分データを使用**:** データ区分フィールドが定義されている場合、このオプションを指定すると、学習用データ区分のデータのみを使用してモデルが構築されます。

One-Class SVM ノードの「エキスパート」

One-Class SVM ノードの「エキスパート」タブで、「シンプル」モードと「エキスパート」モードの中から選択できます。「シンプル」を選択した場合、下に示すデフォルト値を使用してすべてのパラメータが設定されます。「エキスパート」を選択した場合、これらのパラメータのカスタム値を指定できます。これらのオプションについて詳しくは、http://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html#sklearn.svm.OneClassSVMを参照してください。

停止基準**:** 停止基準の許容度を指定します。デフォルトは **1.0E-3** (0.001) です。

回帰精度 **(**ニュー**):** 学習誤差およびサポート・ベクターの小数部の範囲です。デフォルトは **0.1** です。

カーネル タイプ**:** アルゴリズムで使用するカーネル タイプ。オプションは 「**RBF**」、「多項式」、「**Sigmoid**」、「線型」、または 「事前計算済み」 です。デフォルトは「**RBF**」です。

ガンマの指定**:** ガンマを指定するにはこのオプションを選択します。それ以外の場合は、自動ガンマが適用されます。

ガンマ**:** ガンマの設定は、カーネル タイプ「RBF」、「多項式」、および「Sigmoid」の場合のみ使用可能です。

Coef0: Coef0 は、カーネル タイプ「多項式」および「Sigmoid」の場合のみ使用可能です。

次数**:** 次数は、カーネル タイプ「多項式」の場合のみ使用可能です。

収縮ヒューリスティックを使用**:** 収縮ヒューリスティックを使用するには、このオプションを選択します。このオプションは、デフォルトで選択解除されています。

ランダム シードの設定**:** 確率推定のためにデータをシャッフルする際に使用する乱数シードを設定するには、このオプションを選択します。このオプションは、デフォルトで選択解除されています。

カーネル キャッシュのサイズを指定 **(**単位**: MB):** カーネル キャッシュのサイズを指定するには、このオプションを選択します。このオプションは、デフォルトで選択解除されています。選択した場合、デフォルト値は **200** MB です。

ハイパーパラメータ最適化 **(Rbfopt** に基づく**):** Rbfopt に基づくハイパーパラメータ最適化を有効にするには、このオプションを選択します。有効にすると、パラメータの最適な組み合わせが自動的に検出され、サンプルに対するモデルの誤差率が予測値以下になります。Rbfopt について詳しくは、http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.htmlを参照してください。

目標: 目標とする目的関数 (サンプルに対するモデルの誤差率) の値 (例えば、未知の最適条件の値)。許容できる値 (**0.01** など) に設定してください。

最大反復**:** モデルを試行する最大反復数。デフォルトは **1000** です。

最大評価**:** 速度より精度を重視する場合の、モデルを試行するための関数評価の最大回数。デフォルトは **300** です。

One-Class SVM ノードは、scikit-learn© Python ライブラリーを必要とします。次の表に、SPSS Modeler の SMOTE ノードのダイアログの設定と Python アルゴリズムとの間の関係を示します。

表 *41.* ノードのプロパティーと *Python* ライブラリーのパラメータのマッピング

パラメータ名	スクリプト名 **(**プロパティー名**)**	**Python API** パラメータ名
停止基準	stopping_criteria	tol
回帰精度	precision	nu
カーネル タイプ	kernel	kernel
ガンマ	gamma	gamma
Coef0	coef0	coef0
次数	degree	degree
収縮ヒューリスティックを使用	shrinking	shrinking
カーネル キャッシュのサイズを指定 (数値の入力ボックス)	cache_size	cache_size
ランダム シード	random_seed	random_state

One-Class SVM ノードのオプション

One-Class SVM ノードの「オプション」タブで、以下のオプションを設定できます。

並行座標グラフィックスの種類 SPSS Modeler は、構築したモデルを表現するために並行座標グラフィックスを描画します。場合によっては、一部のデータ列が他と比べてかなり大きく表示され、グラフの一部が見えにくくなることがあります。このような場合は、「独立垂直軸」オプションを選択してすべての垂直軸に独立した軸スケールを指定することも、「一般垂直軸」を選択して、すべての垂直軸が同じ軸スケールを共有するように強制することもできます。

グラフィック上の最大行数**:** グラフ出力に表示するデータ行の最大数を指定します。デフォルトは 100 です。パフォーマンス上の理由から、最大で 20 個のフィールドが表示されます。

グラフィック上にすべての入力フィールドを描画**:** すべての入力フィールドをグラフ出力に表示するには、このオプションを選択します。デフォルトでは、各データ・フィールドは垂直軸として描画されます。パフォーマンス上の理由から、最大で 30 個のフィールドが表示されます。

グラフィック上に描画するカスタム フィールド**:** グラフ出力にすべての入力フィールドを表示する代わりに、このオプションを選択して、表示するフィールドのサブセットを選択できます。これによりパフォーマンスを向上させることができます。パフォーマンス上の理由から、最大で 20 個のフィールドが表示されます。

第 18 章 Spark ノード

SPSS Modeler には、Spark 固有のアルゴリズムを使用するためのノードが用意されています。「ノードパレット」の「**Spark**」タブには次のノードがあり、これらを使用して Spark のアルゴリズムを実行できます。これらのノードは、Windows 64、Mac 64、および Linux 64 でサポートされます。なお、これらのノードでは、モデルを構築するためにフラグ/名義型として整数/倍精度の列を指定することはできません。これを行うには、列値を 0/1 または 0、1、2、3、4... に変換する必要があります。

Isotonic 回帰は、回帰アルゴリズムのファミリーに属します。SPSS Modeler の Isotonic-AS ノードは Spark で実装されています。Isotonic 回帰アルゴリズムについて詳しくは、https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html を参照してください。

XGBoost© は、勾配ブースティング・アルゴリズムの高度な実装です。ブースティング・アルゴリズムでは、弱い分類子に繰り返し学習させ、それを最終的な強い分類子に追加します。XGBoost は柔軟性が極めて高く、多くのユーザーを圧倒するほどの多数のパラメータが用意されています。このため、SPSS Modeler の XGBoost-AS ノードでは、コア・フィーチャーおよびよく使用されるパラメータが公開されています。XGBoost-AS ノードは Spark で実装されています。

K-Means は、最も一般的に使用されるクラスタリング アルゴリズムの 1 つです。このアルゴリズムは、データ ポイントをクラスタリングして、事前定義された数のクラスタを作成します。SPSS Modeler の K-Means-AS ノードは Spark で実装されています。 K-Means アルゴリズムについて詳しくは、https://spark.apache.org/docs/2.2.0/ml-clustering.html を参照してください。K-Means-AS ノードでは、カテゴリ変数の場合にワン ホット エンコーディングが自動的に実行されることに留意してください。

多層パーセプトロンはフィード フォワード人工ニューラル ネットワークに基づく分類器であり、複数の層から構成されます。各層は、ネットワーク内の次の層に全結合されます。SPSS Modeler の MultiLayerPerceptron-AS ノードは Spark で実装されています。多層パーセプトロン分類器 (MLPC) について詳しくは、https://spark.apache.org/docs/latest/ml-classification-regression.html#multilayer-perceptron-classifierを参照してください。

Isotonic-AS ノード

Isotonic 回帰は、回帰アルゴリズムのファミリーに属します。SPSS Modeler の Isotonic-AS ノードは Spark で実装されています。

Isotonic 回帰アルゴリズムについて詳しくは、https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html を参照してください。[1]

[1] "Regression - RDD-based API." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

Isotonic-AS ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

フィールド**:** データ ソースのすべてのフィールドをリストします。矢印ボタンを使用して、このリストの項目を画面右側の「目標」フィールド、「入力」フィールド、および「重み」フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

目標: 目標として使用するフィールドを選択します。

入力: 1 つ以上の入力フィールドを選択します。

重み**:** 指数重みの重みフィールドを選択します。設定しない場合は、デフォルト重み値の **1** が使用されます。

Isotonic-AS ノードの作成オプション

Isotonic-AS ノードの作成オプションを指定するには、「作成オプション」タブを使用します。このタブには、機能インデックスや Isotonic タイプが用意されています。詳しくは、http://spark.apache.org/docs/latest/api/java/org/apache/spark/ml/regression/IsotonicRegression.htmlを参照してください。[1]

入力フィールド インデックス**:** 入力フィールドのインデックスを指定します。デフォルトは **0** です。

Isotonic タイプ**:** この設定は、出力シーケンスを isotonic/増加または antitonic/減少のどちらにするかを決定します。デフォルトは **Isotonic** です。

[1] "Class IsotonicRegression." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

Isotonic-AS モデル ナゲット

Isotonic-AS モデル ナゲットには、Isotonic 回帰モデルが取得したすべての情報が含まれます。以下のセクションがあります。

モデル要約

このビューには、入力フィールド、目標フィールド、およびモデル作成オプションを含む、モデルについての重要な情報が提供されます。

モデル グラフ

このビューには、散布図ダイアグラムが表示されます。

XGBoost-AS ノード

XGBoost© は、勾配ブースティング・アルゴリズムの高度な実装です。ブースティング・アルゴリズムでは、弱い分類子に繰り返し学習させ、それを最終的な強い分類子に追加します。XGBoost は柔軟性が極めて高く、多くのユーザーを圧倒するほどの多数のパラメータが用意されています。このため、SPSS Modeler の XGBoost-AS ノードでは、コア・フィーチャーおよびよく使用されるパラメータが公開されています。XGBoost-AS ノードは Spark で実装されています。

ブースティング・アルゴリズムについて詳しくは、XGBoost のチュートリアル (http://xgboost.readthedocs.io/en/latest/tutorials/index.html) を参照してください。[1]

XGBoost の交差検証機能は、SPSS Modeler ではサポートされていません。この機能の代わりに、SPSS Modeler のデータ区分ノードを使用できます。SPSS Modeler の XGBoost では、カテゴリー変数の場合にワン・ホット・エンコーディングが自動的に実行されることにも留意してください。

注: Mac では、XGBoost-AS モデルを構築するには、バージョン 10.12.3 以上が必要です。

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

XGBoost-AS ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

定義済みの役割を使用: 上流のデータ型ノード (または上流の入力ノードの「データ型」タブ) の役割設定 (対象、予測など) を使用します。

カスタム・フィールド割り当ての使用: 対象および予測値を手動で割り当てる場合は、このオプションを選択します。

フィールド: 矢印ボタンを使用して、このリストの項目を画面右側の「目標」役割フィールドおよび「予測値」役割フィールドに手動で割り当てます。アイコンは、各役割フィールドの有効な測定の尺度を示します。リスト内のすべてのフィールドを選択する場合は「すべて」ボタンをクリックし、特定の尺度のすべてのフィールドを選択する場合は各尺度のボタンをクリックします。

目標: この予測の目標として使用するフィールドを選択します。

予測値: 1 つ以上のフィールドを予測の入力として選択します。

XGBoost-AS ノードの作成オプション

XGBoost-AS ノードの作成オプションを指定するには、「作成オプション」タブを使用します。このタブには、モデルの構築や不均衡データセットの処理のための全般オプション、目的や評価メトリックのための学習タスク オプション、および特定のブースティングのためのブースティング パラメータが含まれています。これらのオプションについて詳しくは、以下のオンライン情報源を参照してください。

- XGBoost ホーム ページ[1]
- XGBoost のパラメータの解説[2]
- XGBoost Spark API[3]

一般

ワーカーの数: XGBoost モデルの学習に使用するワーカーの数。

スレッドの数: ワーカーごとに使用するスレッドの数。

外部メモリーの使用: キャッシュとして外部メモリーを使用するかどうか。

ブースティング タイプ: 使用するブースティング (「**gbtree**」、「**gblinear**」、または「**dart**」)。

ブースティングのラウンド数 **(Booster Rounds Number):** ブースティングのラウンド数。

Scale pos weight: この設定は、正の重みと負の重みのバランスを制御します。不均衡クラスの場合に有用です。

ランダム シード**:** 「生成」をクリックして、乱数発生ルーチンによって使用されるシードを生成します。

学習タスク

目的**:** 学習タスクの目的タイプ **reg:linear**、**reg:logistic**、**reg:gamma**、**reg:tweedie**、**rank:pairwise**、**binary:logistic**、または **multi** から選択します。

評価メトリック。 検証データの評価メトリック。デフォルトのメトリックは、目的に応じて割り当てられます (回帰の場合は「**rmse**」、分類の場合は「**error**」、ランク付けの場合は「**mean average precision**」)。選択可能なオプションは「**rmse**」、「**mae**」、「**logloss**」、「**error**」、「**merror**」、「**mlogloss**」、「**uac**」、「**ndcg**」、「**map**」、または「**gamma-deviance**」です。デフォルトは「**rmse**」です。

ブースティング パラメータ

ラムダ**:** 重みに関する L2 正規化項。この値を大きくすると、モデルがより保守的になります。

アルファ**:** 重みに関する L1 正規化項。この値を大きくすると、モデルがより保守的になります。

ラムダ バイアス**:** 偏りに関する L2 正規化項。(偏りに関する L1 正規化項は重要ではないため、存在しません。)

Tree method: 使用する XGBoost ツリー構築アルゴリズムを選択します。

Max depth: ツリーの最大深度を指定します。この値を大きくすると、モデルがより複雑になり、オーバーフィッティングになりやすくなります。

Min child weight: 子で必要なインスタンスの重み (ヘシアン) の最小合計を指定します。ツリーの分割ステップで生じた葉ノードのインスタンスの重みの合計が、この「**Min child weight**」より小さい場合、構築プロセスでそれ以上の分割は行いません。線型モードの場合は、単純に、各ノードに必要なインスタンスの最小値に対応しています。重みを大きくするほど、アルゴリズムがより保守的になります。

Max delta step: 各ツリーの重みを推定できるようにするには、Max delta step を指定します。**0** に設定した場合、何も制約はありません。正の値に設定した場合、更新ステップがより保守的になります。通常はこのパラメータは必要ありませんが、クラスのバランスが著しく悪いときに、ロジスティック回帰で役立つ可能性があります。

サブサンプル**:** サブサンプルは、学習インスタンスの比率を示します。例えば、**0.5** に設定した場合、データ・インスタンスの半数をランダムに収集してツリーを成長させることで、オーバーフィッティングを防ぎます。

イータ**:** オーバーフィッティングを防ぐために更新ステップ中に使用するステップ サイズの収縮。各ブースティング ステップの後で、新規フィーチャーの重みを直接取得できます。イータはフィーチャーの重みも収縮させるため、ブースティングのプロセスがより保守的になります。

ガンマ**:** ツリーの葉ノードをさらに分割するために必要な最小の損失低減。ガンマの設定を大きくするほど、アルゴリズムがより保守的になります。

Colsample by tree: 各ツリーを構築する際の、列のサブサンプルの比率。

Colsample by level: 各分割における、各レベルでの列のサブサンプルの比率。

正規化アルゴリズム**:** 全般オプションで dart のブースティング タイプを選択した場合に使用する正規化アルゴリズム。選択可能なオプションは「ツリー」または「フォレスト」です。デフォルトは「ツリー」です。

サンプル アルゴリズム**:** 全般オプションで dart のブースティング タイプを選択した場合に使用するサンプル アルゴリズム。「一様」アルゴリズムでは、除去対象のツリーを一様に選択します。「重み付き」アルゴリズムでは、重みに比例して除去対象のツリーを選択します。デフォルトは「一様」です。

ドロップアウト率**:** 全般オプションで dart のブースティング タイプを選択した場合に使用するドロップアウト率。

スキップ ドロップアウトの確率**:** 全般オプションで dart のブースティング タイプを選択した場合に使用するスキップ ドロップアウトの確率。ドロップアウトがスキップされる場合、新しいツリーは「**gbtree**」と同様に追加されます。

次の表に、SPSS Modeler の XGBoost-AS ノードのダイアログの設定と XGBoost Spark パラメータとの間の関係を示します。

表 *42.* ノードのプロパティと *Spark* パラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**XGBoost Spark** パラメータ
対象	target_fields	
予測値	input_fields	
ラムダ	lambda	lambda
ワーカーの数	nWorkers	nWorkers
スレッドの数	numThreadPerTask	numThreadPerTask
外部メモリーの使用	useExternalMemory	useExternalMemory
ブースティング タイプ	boosterType	boosterType
ブースティングのラウンド数 (Boosting Round Number)	numBoostRound	round
Scale Pos Weight	scalePosWeight	scalePosWeight
目的	objectiveType	objective
評価メトリック	evalMetric	evalMetric
ラムダ	lambda	lambda
アルファ	alpha	alpha
ラムダ バイアス	lambdaBias	lambdaBias
Tree Method	treeMethod	treeMethod
Max Depth	maxDepth	maxDepth
Min child weight	minChildWeight	minChildWeight
Max delta step	maxDeltaStep	maxDeltaStep
サブサンプル	sampleSize	sampleSize
イータ	eta	eta
ガンマ	gamma	gamma
Colsample by tree:	colsSampleRation	colSampleByTree
Colsample by level	colsSampleLevel	colsSampleLevel

表 42. ノードのプロパティと *Spark* パラメータのマッピング (続き)

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**XGBoost Spark** パラメータ
正規化アルゴリズム	normalizeType	normalizeType
サンプル アルゴリズム	sampleType	sampleType
ドロップアウト率	rateDrop	rateDrop
スキップ ドロップアウトの確率	skipDrop	skipDrop

[1] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

[2] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "ml.dmlc.xgboost4j.scala.spark Params." *DMLC for Scalableand Reliable Machine Learning*. Web. 3 Oct 2017.

XGBoost-AS ノードのモデル オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

K-Means-AS ノード

K-Means は、最も一般的に使用されるクラスタリング アルゴリズムの 1 つです。このアルゴリズムは、データ ポイントをクラスタリングして、事前定義された数のクラスタを作成します。[1] SPSS Modeler の K-Means-AS ノードは Spark で実装されています。

K-Means アルゴリズムについて詳しくは、https://spark.apache.org/docs/2.2.0/ml-clustering.html を参照してください。

K-Means-AS ノードでは、カテゴリ変数の場合にワン ホット エンコーディングが自動的に実行されることに留意してください。

[1] "Clustering." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

K-Means-AS ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

定義済みの役割を使用**:** このオプションを選択すると、上流のデータ型ノードからのフィールド情報がこのノードで使用されます。これは、デフォルトで選択されます。

カスタム・フィールド割り当ての使用: 入力フィールドを手動で割り当てる場合は、このオプションを選択し、1 つ以上の入力フィールドを選択します。このオプションの使用は、データ型ノードのフィールドの役割を「入力」に設定するのと似ています。

K-Means-AS ノードの作成オプション

K-Means-AS ノードの作成オプションを指定するには、「作成オプション」タブを使用します。このタブには、モデル作成のための通常オプション、クラスタ中心の初期化のための初期化オプション、および計算反復とランダム シードのための詳細オプションが含まれています。詳しくは、JavaDoc for K-Means on SparkML を参照してください。[1]

通常

「モデル名」。特定のクラスタへのスコアリングの後に生成されるフィールドの名前。「自動」(デフォルト) を選択するか、「カスタム」を選択してから名前を入力します。

クラスター数: 生成するクラスタの数を指定します。デフォルト値は **5**、最小値は **2** です。

初期化

初期化モード: クラスタ中心の初期化の方法を指定します。**K-Means|| **がデフォルトです。これらの 2 つの方法について詳しくは、ScalableK-Means++ を参照してください。[2]

初期化ステップ: **K-Means|| **初期化モードが選択されている場合は、初期化ステップの数を指定します。**2** がデフォルトです。

詳細

詳細設定: 詳細オプションを以下のように設定する場合は、このオプションを選択します。

最大反復: クラスタ中心を検索するときに実行する最大反復数を指定します。**20** がデフォルトです。

許容度: 反復アルゴリズムの収束許容度を指定します。**1.0E-4** がデフォルトです。

ランダム シードの設定: 乱数発生ルーチンによって使用されるシードを生成するには、このオプションを選択し、「生成」をクリックします。

表示

グラフの表示: 出力にグラフを含める場合は、このオプションを選択します。

次の表に、SPSS Modeler の K-Means-AS ノードの設定と K-Means Spark パラメータとの間の関係を示します。

表 *43.* ノードのプロパティと *Spark* パラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**K-Means SparkML** パラメータ
入力フィールド	`features`	
クラスター数	`clustersNum`	`k`
初期化モード	`initMode`	`initMode`
初期化ステップ	`initSteps`	`initSteps`
最大反復	`maxIter`	`maxIter`
許容度	`toleration`	`tol`
ランダム シード	`randomSeed`	`seed`

[1] "Class KMeans." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

[2] Bahmani, Moseley, et al. "Scalable K-Means++." Feb 28, 2012. http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf.

MultiLayerPerceptron-AS ノード

多層パーセプトロンはフィード フォワード人工ニューラル ネットワークに基づく分類器であり、複数の層から構成されます。各層は、ネットワーク内の次の層に全結合されます。多層パーセプトロン分類器 (MLPC) について詳しくは、https://spark.apache.org/docs/latest/ml-classification-regression.html#multilayer-perceptron-classifierを参照してください。[1]

SPSS Modeler の MultiLayerPerceptron-AS ノードは Spark で実装されています。このノードを使用するには、上流のデータ型ノードをセットアップする必要があります。MultiLayerPerceptron-AS ノードは、データ型ノード (または上流の入力ノードの「データ型」タブ) の入力値を読み取ります。

[1] "Multilayer perceptron classifier." *Apache Spark*. MLlib: Main Guide. Web. 5 Oct 2018.

MultiLayerPerceptron-AS ノードのフィールド

「フィールド」タブでは、分析で使用するフィールドを指定します。

定義済みの役割を使用**:** このオプションを選択すると、上流のデータ型ノードからのフィールド情報がこのノードで使用されます。これがデフォルトです。

カスタム・フィールド割り当ての使用: 目標および予測値を手動で割り当てる場合は、このオプションを選択します。

目標: この予測の目標として使用するフィールドを選択します。

予測値: 予測の入力として使用する 1 つ以上のフィールドを選択します。

MultiLayerPerceptron-AS ノードの作成オプション

MultiLayerPerceptron-AS ノードの作成オプションを指定するには、「作成オプション」タブを使用します。このタブには、パフォーマンス、モデル作成、およびエキスパートのオプションが用意されています。これらのオプションについて詳しくは、http://spark.apache.org/docs/latest/api/java/org/apache/spark/ml/classification/MultilayerPerceptronClassifier.html[1]を参照してください。

パフォーマンス

パーセプトロン層。この設定は、含めるパーセプトロン層の数を定義するために使用します。この値は、パーセプトロン フィールドの数より大きくなければなりません。デフォルト値は **1** です。

隠れ層。隠れ層の数を指定します。複数の隠れ層は間をコンマで区切ってください。デフォルト値は **1** です。

出力層。出力層の数を指定します。デフォルト値は **1** です。

ランダム シード**:** 乱数発生ルーチンで使用するシードを生成する場合は、「生成」をクリックします。

モデルの構築

最大反復**:** 実行する反復の最大回数を指定します。デフォルト値は **10** です。

エキスパートのみ

ブロック サイズ。入力データを行列にスタックするためのブロック サイズを指定する場合は、「モデルの構築」セクションにある「エキスパート モード」オプションを選択します。これによって計算を高速化で

きます。デフォルトのブロック サイズは **128** です。

次の表に、SPSS Modeler の MultiLayerPerceptron-AS ノードのダイアログの設定と Spark KDE ライブラリーのパラメータとの間の関係を示します。

表 *44.* ノードのプロパティと *Spark* パラメータのマッピング

SPSS Modeler の設定	スクリプト名 **(**プロパティー名**)**	**Spark** のパラメータ
予測値	features	
対象	label	
パーセプトロン層	layers[0]	layers[0]
隠れ層	layers[1...<latest-1>]	layers[1...<latest-1>]
出力層	layers[<latest>]	layers[<latest>]
ランダム シード	seed	seed
最大反復	maxiter	maxiter

[1] "Class MultilayerPerceptronClassifier." *Apache Spark*. JavaDoc. Web. 5 Oct 2018.

MultiLayerPerceptron ノードのモデル オプション

モデル名**:** ターゲットまたは ID フィールド (その指定がない場合はモデル タイプ) に基づいてモデル名を生成、またはカスタム名を指定することができます。

特記事項

本書は米国 IBM が提供する製品およびサービスについて作成したものです。この資料は、IBM から他の言語でも提供されている可能性があります。ただし、これを入手するには、本製品または当該言語版製品を所有している必要がある場合があります。

本書に記載の製品、サービス、または機能が日本においては提供されていない場合があります。日本で利用可能な製品、サービス、および機能については、日本 IBM の営業担当員にお尋ねください。本書で IBM 製品、プログラム、またはサービスに言及していても、その IBM 製品、プログラム、またはサービスのみが使用可能であることを意味するものではありません。これらに代えて、IBM の知的所有権を侵害することのない、機能的に同等の製品、プログラム、またはサービスを使用することができます。ただし、IBM 以外の製品とプログラムの操作またはサービスの評価および検証は、お客様の責任で行っていただきます。

IBM は、本書に記載されている内容に関して特許権 (特許出願中のものを含む) を保有している場合があります。本書の提供は、お客様にこれらの特許権について実施権を許諾することを意味するものではありません。実施権についてのお問い合わせは、書面にて下記宛先にお送りください。

〒*103-8510*
東京都中央区日本橋箱崎町*19*番*21*号
日本アイ・ビー・エム株式会社
法務・知的財産
知的財産権ライセンス渉外

IBM およびその直接または間接の子会社は、本書を特定物として現存するままの状態で提供し、商品性の保証、特定目的適合性の保証および法律上の瑕疵担保責任を含むすべての明示もしくは黙示の保証責任を負わないものとします。国または地域によっては、法律の強行規定により、保証責任の制限が禁じられる場合、強行規定の制限を受けるものとします。

この情報には、技術的に不適切な記述や誤植を含む場合があります。本書は定期的に見直され、必要な変更は本書の次版に組み込まれます。IBM は予告なしに、随時、この文書に記載されている製品またはプログラムに対して、改良または変更を行うことがあります。

本書において IBM 以外の Web サイトに言及している場合がありますが、便宜のため記載しただけであり、決してそれらの Web サイトを推奨するものではありません。それらの Web サイトにある資料は、この IBM 製品の資料の一部ではありません。それらの Web サイトは、お客様の責任でご使用ください。

IBM は、お客様が提供するいかなる情報も、お客様に対してなんら義務も負うことのない、自ら適切と信ずる方法で、使用もしくは配布することができるものとします。

本プログラムのライセンス保持者で、(i) 独自に作成したプログラムとその他のプログラム (本プログラムを含む) との間での情報交換、および (ii) 交換された情報の相互利用を可能にすることを目的として、本プログラムに関する情報を必要とする方は、下記に連絡してください。

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

本プログラムに関する上記の情報は、適切な使用条件の下で使用することができますが、有償の場合もあります。

本書で説明されているライセンス・プログラムまたはその他のライセンス資料は、IBM 所定のプログラム契約の契約条項、IBM プログラムのご使用条件、またはそれと同等の条項に基づいて、IBM より提供されます。

記載されている性能データとお客様事例は、例として示す目的でのみ提供されています。実際の結果は特定の構成や稼働条件によって異なります。

IBM 以外の製品に関する情報は、その製品の供給者、出版物、もしくはその他の公に利用可能なソースから入手したものです。IBM は、それらの製品のテストは行っておりません。したがって、他社製品に関する実行性、互換性、またはその他の要求については確証できません。 IBM 以外の製品の性能に関する質問は、それらの製品の供給者にお願いします。

IBM の将来の方向性および指針に関する記述は、予告なく変更または撤回される場合があります。これらは目標および目的を提示するものにすぎません。

本書には、日常の業務処理で用いられるデータや報告書の例が含まれています。より具体性を与えるために、それらの例には、個人、企業、ブランド、あるいは製品などの名前が含まれている場合があります。これらの名称はすべて架空のものであり、類似する個人や企業が実在しているとしても、それは偶然にすぎません。

商標

IBM、IBM ロゴおよび ibm.com は、世界の多くの国で登録された International Business Machines Corporation の商標です。他の製品名およびサービス名等は、それぞれ IBM または各社の商標である場合があります。現時点での IBM の商標リストについては、http://www.ibm.com/legal/copytrade.shtml をご覧ください。

Adobe、Adobe ロゴ、PostScript、PostScript ロゴは、Adobe Systems Incorporated の米国およびその他の国における登録商標または商標です。

インテル、Intel、Intel ロゴ、Intel Inside、Intel Inside ロゴ、Centrino、Intel Centrino ロゴ、Celeron、Xeon、Intel SpeedStep、Itanium、および Pentium は、Intel Corporation または子会社の米国およびその他の国における商標または登録商標です。

Linux は、Linus Torvalds の米国およびその他の国における登録商標です。

Microsoft、Windows、Windows NT および Windows ロゴは、Microsoft Corporation の米国およびその他の国における商標です。

UNIX は The Open Group の米国およびその他の国における登録商標です。

Java およびすべての Java 関連の商標およびロゴは Oracle やその関連会社の米国およびその他の国における商標または登録商標です。

製品資料に関するご使用条件

これらの資料は、以下のご使用条件に同意していただける場合に限りご使用いただけます。

適用条件

IBM Web サイトの「ご利用条件」に加えて、以下のご使用条件が適用されます。

個人的使用

これらの資料は、すべての著作権表示その他の所有権表示をしていただくことを条件に、非商業的な個人による使用目的に限り複製することができます。ただし、IBM の明示的な承諾をえずに、これらの資料またはその一部について、二次的著作物を作成したり、配布 (頒布、送信を含む) または表示 (上映を含む) することはできません。

商業的使用

これらの資料は、すべての著作権表示その他の所有権表示をしていただくことを条件に、お客様の企業内に限り、複製、配布、および表示することができます。ただし、IBM の明示的な承諾をえずにこれらの資料の二次的著作物を作成したり、お客様の企業外で資料またはその一部を複製、配布、または表示することはできません。

権利

ここで明示的に許可されているもの以外に、資料や資料内に含まれる情報、データ、ソフトウェア、またはその他の知的所有権に対するいかなる許可、ライセンス、または権利を明示的にも黙示的にも付与するものではありません。

資料の使用が IBM の利益を損なうと判断された場合や、上記の条件が適切に守られていないと判断された場合、IBM はいつでも自らの判断により、ここで与えた許可を撤回できるものとさせていただきます。

お客様がこの情報をダウンロード、輸出、または再輸出する際には、米国のすべての輸出入 関連法規を含む、すべての関連法規を遵守するものとします。

IBM は、これらの資料の内容についていかなる保証もしません。これらの資料は、特定物として現存するままの状態で提供され、商品性の保証、特定目的適合性の保証および法律上の瑕疵担保責任を含むすべての明示もしくは黙示の保証責任なしで提供されます。

用語集

A

AICC． -2 (制限) 対数尤度に基づいて混合モデルを選択し、比較するための指標。値が小さいほどモデルが良好であることを示します。AICC は、小さな標本サイズに適応するように AIC を「修正」します。標本サイズが大きくなるに従い、AICC は AIC に収束します。

B

Bayesian Information Criterion (BIC) (ベイジアン情報基準 *(BIC))*． -2 対数尤度に基づいてモデルを選択し、比較するための指標。値が小さいほどモデルが良好であることを示します。BIC もパラメータが過多のモデル (例えば、大量の入力がある複雑なモデル) にペナルティを科しますが、AIC よりも厳密にそれを行います。

Box's M test (*Box* の *M* 検定)． グループの共分散行列の等質性を調べる検定。サンプル数が十分に多いが p 値が有意でない場合は、行列が異なるという証拠が不十分であることを意味します。この検定は、多変量正規性からの逸脱に対して敏感です。

C

Cases (ケース)． 実際のグループ、予測グループ、事後確率、および判別得点のコードをケースごとに表示します。

Classification Results (分類結果 (距離と近接度))． 判別分析に基づいて各グループに正しくまたは誤って割り当てられたケースの数。「混同行列」と呼ぶこともあります。

Combined-Groups Plots (結合されたグループの散布図 (判別分析))． 最初の 2 つの判別関数の値を使用して全グループ散布図を作成します。関数が 1 つしかない場合は、代わりにヒストグラムが表示されます。

Covariance (共分散)． 2 つの変数の間の、標準化されていない関連度。偏差の積和を N-1 で割った値に等しくなります。

F

Fisher's (*Fisher* の分類関数の係数)． 分類に直接使用できる、Fisher の分類関数の係数を表示します。分類関数の一連の係数をグループごとに個別に求め、最大判別得点 (分類関数の値) を持つグループにケースを割り当てます。

H

Hazard Plot (累積ハザード関数プロット (生命表/*Kaplan-Meier*/*Cox* 回帰))． 累積ハザード関数を線型スケールで表示します。

K

Kurtosis (尖度)． 外れ値が存在する度合いの指標。正規分布の場合、尖度の統計値は 0 です。尖度が正の場合、そのデータの極端な外れ値は正規分布よりも多いことを示します。尖度が負の場合、そのデータの極端な外れ値は正規分布よりも少ないことを示します。

L

Leave-one-out Classification (Leave-one-out 分類法)． 分析における各ケースを、そのケース以外のすべてのケースから派生した関数で分類します。「U 手法」とも呼びます。

M

MAE ． 平均絶対誤差。モデルによって予測されるレベルから系列がどの程度外れているかを測定します。MAE は、元の系列単位で報告されます。

Mahalanobis Distance (Mahalanobis の距離)． 独立変数のケースの値と全ケースの平均との差異の程度を示す指標。マハラノビスの距離が大きい場合は、ケースにおいて 1 つ以上の独立変数に極値が存在することを示します。

MAPE ． 平均絶対パーセント誤差。モデルによって予測されるレベルから従属系列がどの程度外れているかの指標。使用する単位に依存しないので、異なる単位の系列との比較に使用することができます。

MaxAE ． 最大絶対誤差。最大予測誤差であり、従属系列と同じ単位で表します。MaxAPE と同様に、予測に対する最悪のシナリオを想定する場合に有用です。最大絶対値誤差と最大絶対パーセント誤差は、異なる系列ポイントで生じる場合があります。例えば、大きな系列値の絶対誤差が小さな系列値の絶対誤差よりわずかに大きい場合が挙げられます。その場合、最大絶対誤差は大きい側の系列値で発生し、最大絶対パーセント誤差は小さい側の系列値で発生します。

MaxAPE ． 最大絶対パーセント誤差。最大予測誤差であり、パーセント単位で表します。この指標は、予測に対する最悪のシナリオを想定する場合に有用です。

*Maximizing the Smallest F Ratio Method of Entry (*最小 *F* 比最大化投入法)． グループ間のマハラノビスの距離から計算した F 比の最大化に基づく、ステップワイズ分析での変数選択法。

*Maximum (*最大)． 数値変数の最大値。

*Mean (*平均)． 中心傾向の指標。算術平均 (合計をケース数で割った値) です。

*Means (*平均値 *(*信頼性分析*))*． 独立変数の合計、グループ平均値、および標準偏差を表示します。

*Median (*中央値)． この値より上と下それぞれにケースの半数ずつが該当することになる値。50 パーセンタイル。ケース数が偶数の場合の中央値は、昇順または降順にソートしたときに中央に来る 2 つのケースの平均です。中央値は、外れ値に対して敏感でない、中心傾向の指標です。それに対して平均値は、少数の極端に大きいまたは小さい値に影響されることがあります。

Minimize Wilks' Lambda (Wilks のラムダ最小化による変数選択)． ステップワイズ判別分析における変数選択法の 1 つ。変数が Wilks のラムダを低下させる程度に基づいて式に投入する変数を選択します。各ステップでは、Wilks のラムダが最小になる変数を投入します。

*Minimum (*最小値)． 数値変数の最小値。

*Mode (*最頻値)． 最も多く出現する値。複数の値が最高の頻度で出現し、その頻度が同じである場合は、それぞれが最頻値となります。

N

*Normalized BIC (*正規化 *BIC)*． 正規化ベイズ情報量基準。モデルの複雑さを説明しようとする、モデルの全体的適合度の一般的な指標。平方平均誤差に基づいたスコアであり、モデルおよび系列の長さのパラメーターの数に応じたペナルティーを含みます。ペナルティーにより、パラメーターが多いモデルの利点が減殺されますが、同じ系列の各種のモデルにわたる統計量の比較が容易になります。

O

One Minus Survival (1 マイナス累積生存確率*)*. 線型スケールで 1 マイナス累積生存関数を作図します。

R

*Range (OK (*ファイルオープン時のオプション*))*. 数値変数の最大値と最小値の差。最大値から最小値を引いた値。

Rao's V (Discriminant Analysis) (Rao の *V (*判別分析*))*. グループ平均値の差の指標。Lawley-Hotelling のトレースとも呼びます。各ステップで、Rao の V における増加を最大化する変数を投入します。このオプションを選択した後、分析に投入する変数が持つべき最小値を入力してください。

RMSE. 平方平均誤差平方根。平均平方誤差の平方根。モデルによって予測されるレベルから従属系列がどの程度外れているかを、従属系列と同じ単位を使用して表した指標。

R-Squared (R2 乗*)*. 線型モデルの適合度。決定係数とも呼びます。従属変数の変動のうち、回帰モデルによって説明される割合です。値の範囲は 0 から 1 までです。値が小さい場合は、モデルが十分にデータに適合していないことを示します。

S

*Separate-Groups (*グループ別 *(*判別分析の分類*))*. グループ別共分散行列は分類に使用します。分類は (元の変数ではなく) 判別関数に基づいて行うため、このオプションは必ずしも 2 次の判別と等価ではありません。

*Separate-Groups Covariance (*グループ別分散共分散行列 *(*判別分析*))*. 各グループの個別の共分散行列を表示します。

*Separate-Groups Plots (*グループ別散布図 *(*判別分析*))*. 最初の 2 つの判別関数の値のグループ別散布図を作成します。関数が 1 つしかない場合は、代わりにヒストグラムを表示します。

*Sequential Bonferroni (*逐次 *Bonferroni)*. 個々の仮説を棄却する点であまり保守的ではないが、同じ全体の有意水準を維持する逐次ステップダウン棄却 Bonferroni 手続き。

*Sequential Sidak (*逐次 *Sidak)*. 個々の仮説を棄却する点であまり保守的ではないが、同じ全体の有意水準を維持する逐次ステップダウン棄却 Sidak 手続き。

*Skewness (*歪度*)*. 分布の非対称性の指標。正規分布は対称であり、歪度の値は 0 です。歪度が正の大きな値である分布は、右側の裾が長くなります。歪度が負で絶対値が大きい分布は、左側の裾が長くなります。目安として、歪度が標準誤差の 2 倍より大きい場合は、対称分布からずれていると解釈します。

*standard deviation (*標準偏差*)*. 平均の周りの散らばりの指標。分散の平方根に等しくなります。標準偏差は元の変数と同じ単位で表します。

*Standard Deviation (*標準偏差*)*. 平均値の周りの散らばりの指標。正規分布では、平均から 1 標準偏差以内にケースの 68% が含まれ、2 標準偏差以内にケースの 95% が含まれます。例えば平均年齢が 45 で標準偏差が 10 である場合、正規分布ではケースの 95% が 25 と 65 の間に含まれます。

*Standard Error (*標準誤差*)*. サンプル間で検定統計量の値がどの程度ばらついているかの指標。統計量のサンプル分布の標準偏差です。例えば、平均値の標準誤差はサンプル平均の標準偏差です。

*Standard Error of Kurtosis (*尖度の標準誤差*)*. 標準誤差に対する尖度の比率は、正規性の検定として使用することができます (比率が -2 より小さいか +2 より大きい場合は、正規性を棄却することができます)。尖度が大きな正の値である場合は、分布の裾が正規分布の裾より長いことを示します。尖度が負の値である場合は、裾が短いことを示します (箱形の一様分布に似た形になります)。

*Standard Error of Mean (*平均値の標準誤差*)*. 同じ分布から抽出したサンプルの間で平均値がどの程度異なるかを示す指標。観測した平均と仮説による値をおおまかに比較するために使用することができます (差と標準誤差の比率が -2 より小さいか +2 より大きい場合は、2 つの値が異なっていると結論付けることができます)。

*Standard Error of Skewness (*歪度の標準誤差*)*. 標準誤差に対する歪度の比率は、正規性の検定として使用することができます (比率が -2 より小さいか +2 より大きい場合は、正規性を棄却することができます)。歪度が大きな正の値である場合は、右側の裾が長いことを示します。極端な負の値の場合は、左側の裾が長いことを示します。

*Stationary R-squared (*定常 *R-2* 乗*)*. モデルの定常部分を単純平均モデルと比較する指標。トレンド・パターンまたは季節パターンがある場合は、通常の R 2 乗よりもこの指標を推奨します。定常 R 2 乗は負になる場合があり、範囲は負の無限大から 1 までです。負の値は、検討中のモデルがベースライン・モデルより悪いことを意味します。正の値は、検討中のモデルがベースライン・モデルより良いことを意味します。

*Sum (*合計*)*. 欠損値でない値を持つすべてのケースにわたる値の和 (合計)。

*Survival Plot (*累積生存関数プロット *(*生命表*/Kaplan-Meier/Cox* 回帰*))*. 累積生存関数を線型スケールで表示します。

T

*Territorial Map (*領域マップ *(*判別分析*))*. 関数の値に基づいてケースをグループに分類するために使用する境界のプロット。これらの数値は、ケースの分類先グループに対応します。各グループの平均値は、その境界の内側にアスタリスクで示します。判別関数が 1 つしかない場合は、このマップを表示しません。

*Total Covariance (*全分散共分散行列 *(*判別分析*))*. すべてのケースから得た共分散行列を、1 つのサンプルから取り出したかのように表示します。

U

*Unexplained Variance (*説明されない分散 *(*判別分析*))*. 各ステップで、グループ間の説明されない分散の合計を最小化する変数を投入します。

*Unique (*固有*)*. あらゆる種類の他のすべての効果に適合するように各効果を調整することによって、すべての効果を同時に評価します。

Univariate ANOVAs (1 変量の分散分析 *(*判別分析*))*. 一元配置分散分析を実行して、独立変数ごとにグループ平均値の等質性を検定します。

*Unstandardized (*標準化されていない *(*判別分析*))*. 標準化していない判別関数の係数を表示します。

*Use F Value (*ステップワイズのための *F* 値*)*. F 値が「投入」の値より大きい場合に変数をモデルに投入し、「削除」の値より小さい場合に変数を除去します。「投入」は「削除」より大きくなければならず、いずれの値も正でなければなりません。さらに多くの変数をモデルに投入するには、「投入」の値を下げてください。さらに多くの変数をモデルから除去するには、「除去」の値を上げてください。

*Use Probability of F (*ステップワイズのための *F* 値確率*)*. F 値の有意水準が「投入」の値より小さい場合に変数をモデルに投入し、有意水準が「削除」の値より大きい場合に変数を除去します。「投入」は「削除」より小さくなければならず、いずれの値も正でなければなりません。さらに多くの変数をモデルに投入するには、「投入」の値を上げてください。さらに多くの変数をモデルから除去するには、「除去」の値を下げてください。

V

*Valid (*有効*)*. ユーザー欠損として定義された値もシステム欠損値も持たない有効なケース。

Variance (分散 (信頼性分析)). 平均値の周りの値の散らばりの指標。平均値からの偏差の平方和を、ケース数より 1 少ない値で割ったものに等しくなります。分散の測定単位は、変数自体の単位の 2 乗です。

W

Within-Groups (グループ内 (判別分析の分類)). プールされたグループ内共分散行列は、ケースの分類に使用します。

Within-Groups Correlation (グループ内相関行列 (判別分析)). 相関を計算する前にすべてのグループの個別の共分散行列を平均化することによって得られるプールされたグループ内相関行列を表示します。

Within-Groups Covariance (グループ内共分散). プールされたグループ内共分散行列を表示します。全共分散行列とは異なる場合があります。この行列は、すべてのグループの個別の共分散行列を平均化することによって得られます。

索引

日本語, 数字, 英字, 特殊文字の順に配列されています。なお, 濁音と半濁音は清音と同等に扱われています。

［ア行］

［カ行］

[サ行]

[タ行]

[ナ行]

[ハ行]

［マ行］

［ヤ行］

[ラ行]

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

W

X

Printed in Japan

日本アイ・ビー・エム株式会社
〒103-8510 東京都中央区日本橋箱崎町19-21