IBM SPSS Modeler 16
アプリケーション・ガイド

お願い

本書および本書で紹介する製品をご使用になる前に、 353 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(r) SPSS(r) Modeler バージョン 16、リリース 0、モディフィケーション 0、および新しい版で明記されていない限り、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 16
Applications Guide

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

第 1 章 IBM SPSS Modeler について

IBM® SPSS® Modeler は、ビジネス専門知識を活かして予測モデルを素早く開発し、それをビジネス・オペレーションに展開して意思決定を改善できる、データ・マイニング・ツールのセットです。 IBM SPSS Modeler は、業界標準の CRISP-DM モデルに沿って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優れたビジネスの成果を達成できるようにします。

IBM SPSS Modeler には、機械学習、人工知能、および統計によって実現されるさまざまなモデル作成手法が用意されています。「モデル作成」パレットで利用可能な手法を使用し、データから新しい情報を派生させ、予測モデルを作成できます。各手法によって、その長所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するか、SPSS Modeler Server と連携するクライアントとして使用することができます。以降のセクションで説明するように、さまざまな追加オプションを使用することもできます。 詳細については、http://www.ibm.com/software/analytics/spss/products/modeler/を参照してください。

IBM SPSS Modeler 製品

IBM SPSS Modeler 製品ファミリーおよび関連ソフトウェアには、次のものがあります。

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Collaboration and Deployment Services 用の IBM SPSS Modeler Server アダプター

IBM SPSS Modeler

SPSS Modeler は、パーソナル・コンピューターにインストールして実行する、すべての機能が搭載された製品バージョンです。SPSS Modeler は、スタンドアロン製品としてローカル・モードで実行することも、IBM SPSS Modeler Server とともに分散モードで使用して大規模データ・セットでのパフォーマンスを向上させることもできます。

SPSS Modeler では、正確な予測モデルを、プログラミングせずに素早く直感的に構築できます。独自のビジュアル・インターフェースを使用して、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。この製品に組み込まれている高度な分析機能の支援を受けて、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、それに影響を与える要因を理解することで、ビジネス・チャンスを活用して、リスクを軽減することができます。

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。詳細については、 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 では、クライアント/サーバー方式を使用して、リソースに負荷がかかる操作の要求を強力なサーバー・ソフトウェアに分散させることによって、大規模データ・セットでのパフォーマンスが高速化されます。

SPSS Modeler Server は、個別にライセンス付与される製品であり、1 つ以上の IBM SPSS Modeler インストール環境と連携して、1 つのサーバー・ホスト上で分散分析モードで絶えず動作し続けます。このように、SPSS Modeler Server は大規模データ・セットでのパフォーマンスに優れています。これはメモリーの負荷が高い操作はサーバー上で実行され、クライアント・コンピューターにはデータがダウンロードされないためです。また、IBM SPSS Modeler Server は SQL 最適化機能およびデータベース内モデル作成機能もサポートしており、パフォーマンスおよび自動化においてさらなるメリットがあります。

IBM SPSS Modeler Administration Console

Modeler Administration Console は、多くの SPSS Modeler Server 構成オプションを管理するグラフィカル・アプリケーションであり、これはオプション・ファイルを使用して構成することもできます。このアプリケーションでは、SPSS Modeler Server インストール済み環境をモニターして構成するためのコンソール・ユーザー・インターフェースが提供されており、現在の SPSS Modeler Server をご使用のお客様は無償でご使用いただけます。この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、サポートされるプラットフォームにインストールされているサーバーを管理することができます。

IBM SPSS Modeler Batch

通常、データ・マイニングは対話的なプロセスですが、コマンド・ラインから SPSS Modeler を実行することも可能で、グラフィカル・ユーザー・インターフェースを使用する必要はありません。例えば、ユーザーの介入なしで実行する必要がある長時間のタスクまたは反復的なタスクがあるとします。SPSS Modeler Batch は、通常のユーザー・インターフェースへのアクセスなしで SPSS Modeler のすべての分析機能をサポートする特別なバージョンの製品です。 SPSS Modeler Batch を使用するには、SPSS Modeler Server ライセンスが必要です。

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher は、外部のランタイム・エンジンで実行するか、外部アプリケーションに組み込むことが可能なパッケージ・バージョンの SPSS Modeler ストリームを作成できるツールです。これにより、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 ストリームを発行して、展開することができます。IBM SPSS Collaboration and Deployment Services - Scoring サービスの一部として SPSS Modeler Solution Publisher が配布されますが、これには別途ライセンスが必要です。ランセンスが付与されると、SPSS Modeler Solution Publisher Runtime を受け取り、発行済みストリームを実行できるようになります。

IBM SPSS Collaboration and Deployment Services 用の IBM SPSS Modeler Server アダプター

SPSS Modeler および SPSS Modeler Server が IBM SPSS Collaboration and Deployment Services リポジトリーと対話可能になる IBM SPSS Collaboration and Deployment Services 用のアダプターが用意されています。この方法では、リポジトリーに展開された SPSS Modeler ストリームを複数のユーザーが共有することも、シン・クライアント・アプリケーション IBM SPSS Modeler Advantage からアクセスすることもできます。 このアダプターは、リポジトリーをホストするシステムにインストールします。

IBM SPSS Modeler のエディション

SPSS Modeler には、次のエディションがあります。

SPSS Modeler Professional

SPSS Modeler Professional では、CRM システムで追跡される行動および対話、人口統計、購買行動、販売データなど、ほとんどの種類の構造化データの処理に必要なツールがすべて提供されています。

SPSS Modeler Premium

SPSS Modeler Premium は、別途ライセンス付与される製品であり、SPSS Modeler Professional を拡張して、エンティティー分析やソーシャル・ネットワーキングなどに使用される特殊なデータ、および構造化されていないテキスト・データを処理できるようにします。SPSS Modeler Premium には、次のコンポーネントが含まれています。

IBM SPSS Modeler Entity Analytics により、IBM SPSS Modeler の予測分析に新たな次元が追加されます。予測分析は過去のデータから将来の行動を予測しようとするのに対し、エンティティー分析ではレコードの中で Identity の競合を解決することで現在のデータの一貫性と整合性を改善することに焦点を当てます。Identity は、個人、会社、オブジェクト、またはあいまいさの存在する他のエンティティーとなります。Identity の解決は、カスタマー・リレーションシップ・マネジメント、不正行為の検出、マネーロンダリング防止、国内および国際的なセキュリティーを含むさまざまなフィールドにおいて重要になります。

IBM SPSS Modeler Social Network Analysis は、関係についての情報を、個人およびグループの社会的行動を特徴づけるフィールドに変換します。IBM SPSS Modeler Social Network Analysis はソーシャル・ネットワークの基礎となる関係を説明するデータを使用し、ネットワーク内の他の人物の行動に影響を与える社会的リーダーを特定します。また、他のネットワーク参加者に最も影響を受ける人々を確認できます。これらの結果を他の指標と組み合わせることによって、予測モデルの基となる個人の包括的プロファイルを作成できます。この社会的情報を含むモデルは、含まないモデルよりもパフォーマンスが高くなります。

IBM SPSS Modeler Text Analytics は、高度な言語技術と自然言語処理 (NLP) を使用して、多様な構造化されていないテキスト・データを迅速に処理し、重要な概念を抽出し編成し、その概念をカテゴリーにグループ化します。抽出された概念およびカテゴリーを、デモグラフィックなど、既存の構造化データと組み合わせて、IBM SPSS Modeler データ・マイニング・ツール一式をすべて使用してモデル作成に適用することで、より的確に焦点の合った意思決定につなげることができます。

IBM SPSS Modeler の資料

オンライン・ヘルプ形式の資料は、SPSS Modeler の「ヘルプ」メニューから参照できます。これには、SPSS Modeler、SPSS Modeler Server および SPSS Modeler Solution Publisher のほか、このアプリケーション・ガイドなどのサポート資料が含まれています。

インストール手順を含む各製品のすべての資料は PDF 形式で用意されており、各製品 DVD の「*¥Documentation*」フォルダーにあります。インストール資料は Web サイト (http://www-01.ibm.com/support/docview.wss?uid=swg27038316) からダウンロードすることもできます。

どちらの形式の資料も、SPSS Modeler インフォメーション・センター (http://publib.boulder.ibm.com/infocenter/spssmodl/v16r0m0/) で参照できます。

SPSS Modeler Professional の資料

SPSS Modeler Professional の資料一式を次に示します (インストール手順は除きます)。

- **IBM SPSS Modeler ユーザーズ・ガイド:** SPSS Modeler の使用方法に関する入門ガイドです。データ・ストリームの構築方法、欠損値の処理方法、CLEM の式の作成方法、プロジェクトおよびレポートの使用方法、および IBM SPSS Collaboration and Deployment Services や、予測アプリケーション、IBM SPSS Modeler Advantage に展開するストリームのパッケージ化方法について説明しています。
- 「**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。つまり、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、機械学習、人工知能、および統計によって実現されるさまざまなモデル作成手法が用意されています。
- **IBM SPSS Modeler アルゴリズム・ガイド:** IBM SPSS Modeler で使用するモデル作成手法の数学的な基礎についての説明です。このガイドは、PDF 形式のみ提供されています。
- 「**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。このガイドのオンライン・バージョンは、「ヘルプ」メニューからも参照できます。詳細については、5 ページの『アプリケーションの例(E)』を参照してください。
- 「**IBM SPSS Modeler スクリプトとオートメーション**」。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 Collaboration and Deployment Services 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 のストリームおよびシナリオ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、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・パーティー製プログラムを IBM SPSS Modeler にノードとして統合する機能が用意されています。
- 「**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 の構成および管理方法に関する情報を示します。
- **IBM SPSS Modeler Administration Console ユーザー・ガイド:** IBM SPSS Modeler Server のモニターおよび構成用のコンソール・ユーザー・インターフェースのインストールおよび使用方法に関する情報を示します。このコンソールは、Deployment Manager アプリケーションへのプラグインとして実装されます。
- 「**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 形式のみ提供されています。

SPSS Modeler Premium の資料

SPSS Modeler Premium の資料一式を次に示します (インストール手順は除きます)。

- 「**IBM SPSS Modeler Entity Analytics ユーザー・ガイド**」。SPSS Modeler でエンティティー分析を使用する場合の情報。リポジトリーのインストールと構成、エンティティー分析ノード、および管理用タスクについて説明します。

- 「**IBM SPSS Modeler Social Network Analysis ユーザー・ガイド**」。SPSS Modeler でソーシャル・ネットワーク分析を行うためのガイド。グループ分析、拡散分析などについて説明します。
- 「**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。
- **IBM SPSS Modeler Text Analytics Administration Console ユーザー・ガイド:** SPSS Modeler Text Analytics と併用する IBM SPSS Modeler Server のモニターおよび構成用のコンソール・ユーザー・インターフェースのインストールおよび使用に関する情報。このコンソールは、Deployment Manager アプリケーションへのプラグインとして実装されます。

アプリケーションの例(E)

SPSS Modeler のデータ・マイニング・ツールは多種多様なビジネスおよび組織上の問題の解決に役立てることができますが、この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法について、簡潔に対象を絞って紹介します。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。

SPSS Modeler の「ヘルプ」メニューで「**アプリケーションの例**」をクリックすると、サンプルにアクセスできます。データ・ファイルおよびサンプル・ストリームは、製品のインストール・ディレクトリーの下の「*Demos*」フォルダーにインストールされます。詳細については、『「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。「*IBM SPSS Modeler データベース内 マイニング・ガイド*」の例を参照してください。

スクリプトの例。「*IBM SPSS Modeler スクリプトとオートメーション ガイド*」の例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するデータ・ファイルおよびサンプル・ストリームは、製品のインストール・ディレクトリーの下の「*Demos*」フォルダーにインストールされます。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・グループからアクセスするか、または「ファイルを開く」ダイアログ・ボックスの最近使用したディレクトリーの一覧で「*Demos*」をクリックすることでアクセスできます。

第 2 章 IBM SPSS Modeler概要

はじめに

IBM SPSS Modeler では、データ・マイニング・アプリケーションとして、大規模データ・セットにおける有用な関係を見つけるための戦略的アプローチが用意されています。従来の統計的方式とは対照的に、開始時点で自分が求めるものを把握している必要はありません。有用な情報が見つかるまで、色々なモデルを当てはめ、さまざまな関係を調査してデータを探索することができます。

IBM SPSS Modeler の開始

アプリケーションを開始するには、次の項目をクリックします。

「スタート」 > 「すべてのプログラム」 > 「IBM SPSS Modeler 16」 > 「IBM SPSS Modeler 16」

数秒後にメイン・ウィンドウが表示されます。

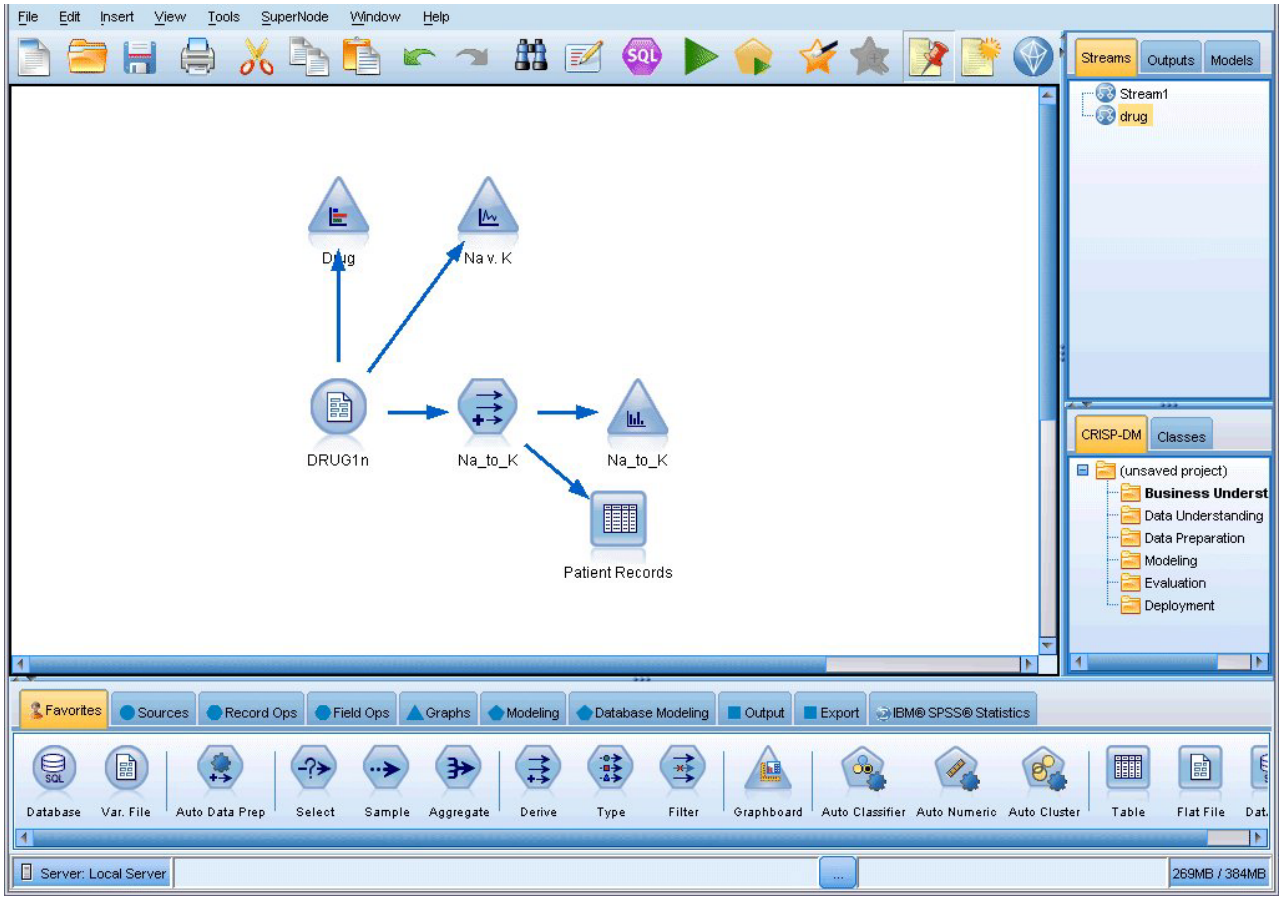


図 1. IBM SPSS Modeler のメイン・アプリケーション・ウィンドウ

コマンド・ラインからの起動

オペレーティング・システムのコマンド・ラインを使用して、IBM SPSS Modeler を起動できます。次のように行います。

1. IBM SPSS Modeler がインストールされているコンピューターで、DOS (コマンド・プロンプト) ウィンドウを開きます。
2. IBM SPSS Modeler インターフェースを対話モードで起動するには、`modelerclient` コマンドに続けて必要な引数を入力します。次に例を示します。

```
modelerclient -stream report.str -execute
```

使用可能な引数 (フラグ) により、サーバーへの接続、ストリームの読み込み、スクリプトの実行、または必要に応じて他のパラメーターの指定を行うことができます。

IBM SPSS Modeler Server への接続

IBM SPSS Modeler はスタンドアロン・アプリケーションとして、あるいは IBM SPSS Modeler Server に直接接続されているクライアントまたはIBM SPSS Collaboration and Deployment Services から Coordinator of Processes プラグインを使用して IBM SPSS Modeler Server かサーバー・クラスターに接続されているクライアントとして実行できます。現在の接続状況は、IBM SPSS Modeler ウィンドウの左下に表示されます。

サーバーに接続するたびに、接続先のサーバー名を手動で入力することも、事前に定義した名前を選択することもできます。ただし、IBM SPSS Collaboration and Deployment Services がある場合は、「サーバーへのログイン」ダイアログ・ボックスからサーバーまたはサーバー・クラスターのリストを検索できます。ネットワーク上で実行されている Statistics サービスを通じて参照する機能が、Coordinator of Processes によって使用可能になります。

サーバーに接続するには

1. 「ツール」メニューで、「**サーバーへのログイン**」をクリックします。「サーバーへのログイン」ダイアログ・ボックスが開きます。または、IBM SPSS Modeler ウィンドウの接続状況の領域をダブルクリックします。
2. このダイアログ・ボックスを使用して、ローカル・サーバー・コンピューターに接続するオプションを指定するか、テーブルから接続を選択します。
   - 「**追加**」または「**編集**」をクリックして、接続を追加または編集します。詳細については、 9 ページの『IBM SPSS Modeler Server 接続の追加および編集』を参照してください。
   - 「**検索**」をクリックして、Coordinator of Processes 内のサーバーまたはサーバー・クラスターにアクセスします。詳細については、 10 ページの『IBM SPSS Collaboration and Deployment Services でのサーバーの検索』を参照してください。

   「**サーバー・テーブル**」。このテーブルには、一連の定義済みのサーバー接続が含まれています。テーブルには、デフォルトの接続、サーバー名、説明、およびポート番号が表示されます。既存の接続を選択または検索できるほか、新しい接続を手動で追加することができます。特定のサーバーをデフォルトの接続として設定するには、接続のテーブルの「デフォルト」列のチェック・ボックスを選択します。

   「**デフォルト・データ・パス**」。サーバー・コンピューター上のデータに使用されるパスを指定します。省略符号ボタン 「**...**」をクリックして、目的の場所を参照します。

   「**資格情報の設定**」。 このボックスのチェック・マークを外した状態にして、**シングル・サインオン**機能を有効にします。これにより、ローカル・コンピューターのユーザー名およびパスワードの詳細を使

用してサーバーへのログインを試みます。シングル・サインオンが不可能な場合、またはこのボックスにチェック・マークを付けてシングル・サインオンを無効にした場合 (例えば、管理者アカウントにログインする場合)、次のフィールドで資格情報を入力できます。

「**ユーザー ID**」。サーバーへのログオンに使用するユーザー名を入力します。

「**パスワード**」。指定したユーザー名に関連付けられたパスワードを入力します。

「**ドメイン**」。サーバーにログオンするために使用するドメインを指定します。ドメイン名は、サーバー・コンピューターがクライアント・コンピューターとは異なる Windows ドメインにある場合にのみ必要です。

3. 「**OK**」をクリックして接続を完了します。

サーバーとの接続を切断するには

1. 「ツール」メニューで、「**サーバーへのログイン**」をクリックします。「サーバーへのログイン」ダイアログ・ボックスが開きます。または、IBM SPSS Modeler ウィンドウの接続状況の領域をダブルクリックします。
2. ダイアログ・ボックスで、「ローカル・サーバー」を選択し、「**OK**」をクリックします。

IBM SPSS Modeler Server 接続の追加および編集

「サーバーへのログイン」ダイアログ・ボックスでサーバー接続を手動で編集または追加できます。「追加」をクリックすると、空の「サーバーの追加/編集」ダイアログ・ボックスが表示され、そこでサーバー接続の詳細を入力することができます。「サーバーへのログイン」ダイアログ・ボックスで既存の接続を選択して「編集」をクリックすると、「サーバーの追加/編集」ダイアログ・ボックスが開いて接続の詳細が表示され、その接続を変更できるようになります。

注: IBM SPSS Collaboration and Deployment Services から追加されたサーバー接続は編集できません。これは、名前、ポート、およびその他の詳細が IBM SPSS Collaboration and Deployment Services で定義されているためです。

サーバー接続を追加するには

1. 「ツール」メニューで、「**サーバーへのログイン**」をクリックします。「サーバーへのログイン」ダイアログ・ボックスが開きます。
2. このダイアログ・ボックスで、「**追加**」をクリックします。「サーバーへのログイン: サーバーの追加/編集」ダイアログ・ボックスが開きます。
3. サーバー接続の詳細を入力し、「**OK**」をクリックして接続を保存し、「サーバーへのログイン」ダイアログ・ボックスに戻ります。

- 「**サーバー**」。使用可能なサーバーを指定するか、リストから 1 つ選択します。サーバー・コンピューターは、英数字名 (例えば、*myserver*) またはサーバー・コンピューターに割り当てられた IP アドレス (例えば、202.123.456.78) で識別されます。
- 「**ポート**」。サーバーが listen するポート番号を指定します。デフォルト設定では機能しない場合は、システム管理者に正しいポート番号を問い合わせてください。
- 「**説明**」。必要に応じて、このサーバー接続の説明を入力します。
- 「**セキュア接続を確保 (SSL の使用)**」。SSL (**Secure Sockets Layer**) 接続を使用するかどうかを指定します。SSL は、ネットワーク経由でのデータ送信を保護するために一般的に使用されるプロトコルです。この機能を使用するには、SSL をサーバーがホスティングする IBM SPSS Modeler Server で有効にする必要があります。必要な場合は、詳細をローカル管理者に問い合わせてください。

サーバー接続を編集するには

1. 「ツール」メニューで、「**サーバーへのログイン**」をクリックします。「サーバーへのログイン」ダイアログ・ボックスが開きます。
2. このダイアログ・ボックスで、編集する接続を選択し、「**編集**」をクリックします。「サーバーへのログイン: サーバーの追加/編集」ダイアログ・ボックスが開きます。
3. サーバー接続の詳細を変更し、「**OK**」をクリックして変更を保存し、「サーバーへのログイン」ダイアログ・ボックスに戻ります。

IBM SPSS Collaboration and Deployment Services でのサーバーの検索

サーバー接続を手動で入力する代わりに、IBM SPSS Collaboration and Deployment Services で使用できる Coordinator of Processes を介して、ネットワークで使用可能なサーバーまたはサーバー・クラスターを選択できます。サーバー・クラスターとはサーバーのグループであり、その中から Coordinator of Processes が処理要求への応答に最適なサーバーを決定します。

「サーバーへのログイン」ダイアログ・ボックスで手動でサーバーを追加することもできますが、使用可能なサーバーの検索では、正しいサーバー名およびポート番号を把握していなくてもサーバーに接続できます。この情報は自動的に提供されます。ただし、ユーザー名、ドメイン、パスワードなどの、正しいログオン情報は必要です。

注: Coordinator of Processes 機能にアクセスできない場合でも、接続先のサーバー名を手動で入力するか、以前定義した名前を選択することができます。詳細については、9 ページの『IBM SPSS Modeler Server 接続の追加および編集』を参照してください。

サーバーおよびクラスターを検索するには

1. 「ツール」メニューで、「**サーバーへのログイン**」をクリックします。「サーバーへのログイン」ダイアログ・ボックスが開きます。
2. このダイアログ・ボックスで、「**検索**」をクリックして「サーバーの検索」ダイアログ・ボックスを開きます。Coordinator of Processes を参照しようとしたときに、IBM SPSS Collaboration and Deployment Services にログオンしていない場合は、ログオンするように求めるプロンプトが表示されます。
3. リストからサーバーまたはサーバー・クラスターを選択します。
4. 「**OK**」をクリックしてダイアログ・ボックスを閉じ、この接続を「サーバーへのログイン」ダイアログ・ボックスのテーブルに追加します。

一時ディレクトリーの変更

IBM SPSS Modeler Server で実行する操作の中には、一時ファイルを作成する必要があるものもあります。デフォルトでは、IBM SPSS Modeler は一時ファイルを作成する際に、システムの一時ディレクトリーを使用します。一時ディレクトリーの場所を変更するには、次の手順に従ってください。

1. 新規ディレクトリー *spss* およびそのサブディレクトリー *servertemp* を作成します。
2. IBM SPSS Modeler のインストール・ディレクトリーの */config* ディレクトリーにある *options.cfg* を編集します。このファイルの中の `temp_directory` パラメーターを、次のように編集します。
   `temp_directory, "C:/spss/servertemp"`
3. これを行った後は、IBM SPSS Modeler Server サービスを再始動する必要があります。再始動するには、Windows の「コントロール パネル」で「**サービス**」タブをクリックします。サービスを停止した後、再び開始すると、変更内容が有効になります。また、マシンを再始動しても、サービスが再始動されます。

これで、すべての一時ファイルがこの新しいディレクトリーに書き込まれるようになります。

注: この作業を行う際によくあるエラーは、誤ったタイプのスラッシュの使用によるものです。普通のスラッシュを使用します。

複数の IBM SPSS Modeler セッションの開始

一度に複数の IBM SPSS Modeler セッションを起動する必要がある場合、IBM SPSS Modeler および Windows の設定を一部変更しなければなりません。例えば、2 つの個別のサーバー・ライセンスがあるので、同じクライアント・マシンから 2 つのストリームを 2 つの異なるサーバーに対して実行したい場合に、変更が必要になります。

複数の IBM SPSS Modeler セッションを有効化するには、次の操作を行います。

1. 以下の項目をクリックします。

   「スタート」 > 「すべてのプログラム」 > 「IBM SPSS Modeler 16」

2. IBM SPSS Modeler 16 のショートカット (アイコンで表示) を右クリックし、**「プロパティー」**を選択します。
3. **「対象」**テキスト・ボックスで、文字列の最後に `-noshare` を追加します。
4. Windows エクスプローラーで、次の項目を選択します。

   「ツール」 > 「フォルダー・オプション...」

5. 「ファイルの種類」タブで、「IBM SPSS Modeler ストリーム」オプションを選択し、**「詳細設定」**をクリックします。
6. 「ファイルの種類の編集」ダイアログ・ボックスで、「IBM SPSS Modeler で開く」を選択し、**「編集」**をクリックします。
7. **「アクションを実行するアプリケーション」**テキスト・ボックスで、**-stream** 引数の前に `-noshare` を追加します。

IBM SPSS Modeler のインターフェースの概要

IBM SPSS Modeler の使いやすいインターフェースでは、データ・マイニング・プロセスの各ポイントで、特定のビジネス専門知識を活用できます。予測、分類、セグメント化、関連性検出などのモデル作成アルゴリズムによって、高機能で正確なモデルが作成されます。作成されたモデルは簡単に展開してデータベース、IBM SPSS Statistics、およびその他の多種多様なアプリケーションで読み取ることが可能です。

IBM SPSS Modeler では、3 つの手順のプロセスでデータを処理します。

- 最初に、データを IBM SPSS Modeler に読み取ります。
- 次に、一連の操作でデータを実行します。
- 最後に、そのデータを宛先に送信します。

この一連の操作は、データが入力元からレコード単位で各操作を通り、最終的には宛先に辿り着くため (モデルまたはある種のデータ出力) 、**データ・ストリーム**と呼ばれます。

図 2. 単純なストリーム

IBM SPSS Modeler のストリーム領域

ストリーム領域は、IBM SPSS Modeler ウィンドウの最大の領域であり、ここでデータ・ストリームを構築し、操作します。

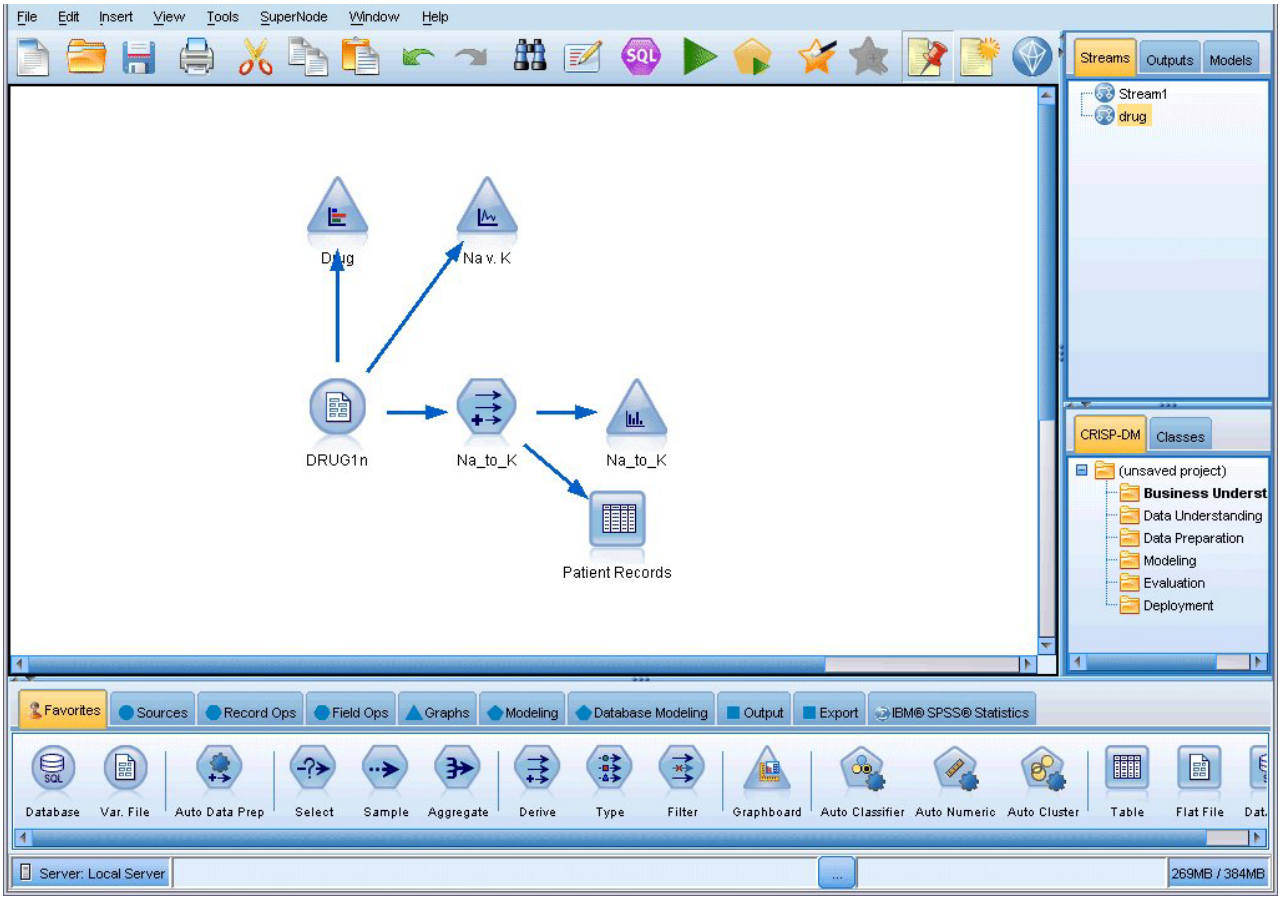


図 3. IBM SPSS Modeler ワークスペース (デフォルト・ビュー)

ストリームは、インターフェースのメイン領域でビジネスに関係するデータ操作のダイアグラムを描画して作成します。各操作は、アイコンまたは**ノード**で表され、そのノードは**ストリーム**内で相互に結び付けられて、各操作を経由するデータの流れを表します。

IBM SPSS Modeler では、同じストリーム領域内で、または新しいストリーム・キャンバスを開くことで、一度に複数のストリームを処理することが可能です。セッション中、ストリームは IBM SPSS Modeler ウィンドウの右上にあるストリーム・マネージャーに格納されます。

ノード・パレット(N)

IBM SPSS Modeler のデータおよびモデル作成ツールのほとんどは、ストリーム領域の下のウィンドウ最下部の「**ノード・パレット**」にあります。

例えば、「レコード操作」パレット・タブには、選択、結合、追加など、データ・**レコード**の操作を実行するために使用できるノードが含まれています。

この領域にノードを追加するには、ノード・パレットのアイコンをダブルクリックするか、アイコンを領域にドラッグ・アンド・ドロップします。次にそれらを接続して、データの流れを示す**ストリーム**を作成します。

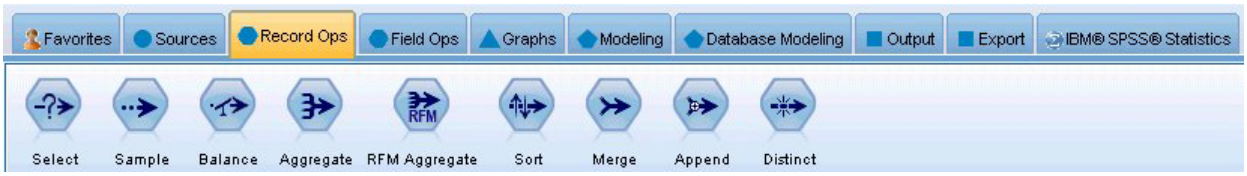


図 4. ノード・パレットの「レコード操作」タブ

各パレット・タブには、次のようなストリーム操作のさまざまなフェーズで使用される関連ノードが集められています。

- 「**ソース**」。IBM SPSS Modeler にデータを入力するノード。
- 「**レコード操作**」。選択、結合、追加など、データ・**レコード**の操作を実行するノード。
- 「**フィールド操作**」。フィルタリング、新規フィールドの派生、特定フィールドの測定の尺度の特定など、データ・**フィールド**の操作を実行するノード。
- 「**グラフ**」。モデルの作成前後にデータをグラフィカル表示するノード。グラフには、プロット、ヒストグラム、Web グラフ・ノード、および評価グラフなどがあります。
- 「**モデル作成**」。ニューラル・ネット、デシジョン・ツリー、クラスタリング・アルゴリズム、データの順序付けなど、IBM SPSS Modeler で使用可能なモデル作成アルゴリズムを使用するノード。
- 「**データベース・モデリング**」。Microsoft SQL Server、IBM DB2、および Oracle データベースと Netezza データベースで利用できるモデル作成アルゴリズムを表すノードです。
- 「**出力**」。IBM SPSS Modeler で表示可能なデータ、グラフ、モデル用にさまざまな出力を作成するノード。
- 「**エクスポート**」。IBM SPSS Data Collection または Excel などの外部アプリケーションで表示できるさまざまな出力を作成するノード。
- **IBM SPSS Statistics.** IBM SPSS Statistics 手続きを実行するほか、IBM SPSS Statistics との間でデータのインポートまたはエクスポートを行います。

IBM SPSS Modeler に慣れてきたら、パレットの内容を使いやすくカスタマイズすることができます。

「ノード・パレット」の下にある「レポート」ペインには、データ・ストリームにデータを読み取り中など、さまざまな操作の進捗に関するフィードバックが表示されます。「ノード・パレット」の下には「ステータス」ペインもあり、ここにはアプリケーションの現在の処理状況や、ユーザーへのフィードバックが必要なときの指示などが表示されます。

IBM SPSS Modeler マネージャー

ウィンドウの右上はマネージャー・ペインです。3 つのタブがあり、ストリーム、出力、およびモデルの管理に使用します。

「ストリーム」タブを使用して、セッション中に作成されたストリームを開くほか、名前の変更、保存、および削除を実行できます。

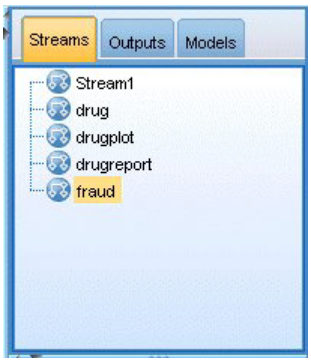


図 5. 「ストリーム」タブ

図 6. 「出力」タブ

「出力」タブには、IBM SPSS Modeler でのストリーム操作で作成された各種のファイル (グラフ、表など) が表示されます。このタブにリストされる表、グラフ、およびレポートを表示、保存、名前変更、閉じることができます。

図 7. モデル・ナゲットが表示された「モデル」タブ

「モデル」タブは、「マネージャー」タブの最も高機能なタブです。このタブには、現在のセッションの**ナゲット**が表示されます。ここには、ナゲットには、IBM SPSS Modeler で生成されたモデルが含まれます。これらのモデルは、「モデル」タブから直接参照することも、領域内のストリームに追加することもできます。

IBM SPSS Modeler のプロジェクト

ウィンドウの右下はプロジェクト・ペインで、データ・マイニングの**プロジェクト** (データ・マイニング・タスクに関連するファイルのグループ) の作成および管理に使用します。 IBM SPSS Modeler で作成したプロジェクトを表示する方法には、クラス・ビューと CRISP-DM ビューの 2 とおりの方法があります。

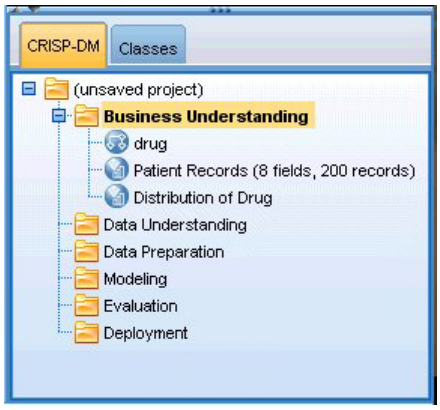


図 8. CRISP-DM ビュー

「CRISP-DM」タブでは、業界で認められた一般的方法論である CRISP-DM (Cross-Industry Standard Process for Data Mining) に従って、プロジェクトを編成できます。データ・マイニングの熟練者も初心者も、CRISP-DM ツールを使用すると、編成も、成果のやり取りも実行しやすくなります。

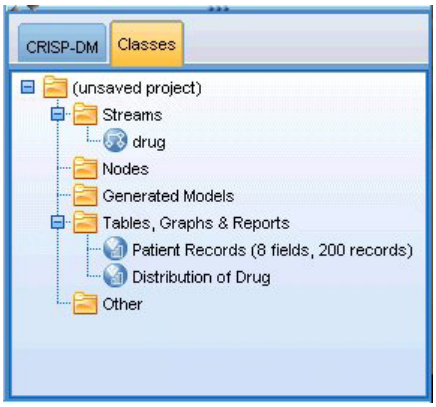


図 9. クラス・ビュー

「クラス」タブでは、IBM SPSS Modeler での作業内容を、作成したオブジェクトの種類別に分類できます。このビューは、データ、ストリーム、およびモデルの現状を把握するときに便利です。

IBM SPSS Modeler ツールバー

IBM SPSS Modeler ウィンドウの最上部には、多くの有用な機能を提供するアイコンを表示するツールバーがあります。ツールバー・ボタンとその機能を次に説明します。


新規ストリームの作成


ストリームを開く


ストリームを保存


現在のストリームを印刷


切り取ってクリップボードに移動


クリップボードにコピー


選択項目の貼り付け


最後の操作を元に戻す


やり直し


ノードの検索


ストリームのプロパティーを編集

SQL 生成をプレビュー


現在のストリームを実行


選択したストリームを実行


ストリームの中止 (ストリームの実行中のみアクティブ)


スーパーノードの追加

ズームイン (スーパーノード専用)

ズームアウト (スーパーノード専用)

ストリームにマークアップがありません

コメントの挿入

ストリーム・マークアップを非表示 (ある場合)

非表示のストリーム・マークアップを表示

IBM SPSS Modeler Advantage でストリームを開く

ストリーム・マークアップは、ストリームのコメント、モデル・リンク、およびスコアリング・ブランチの標識で構成されています。

モデル・リンクの説明は、「*IBM SPSS モデル作成ノード ガイド*」を参照してください。

ツールバーのカスタマイズ

ツールバーは、次のようにさまざまな観点から変更できます。

- 表示するかどうか
- アイコンのツールチップを使用可能にするかどうか
- 大きいアイコンと小さいアイコンのどちらを使用するか

ツールバーの表示をオンまたはオフにするには

1. メイン・メニューで次の各項目をクリックします。

   「表示」 > 「ツールバー」 > 「表示」

ツールチップまたはアイコン・サイズの設定を変更するには

1. メインメニューで、次の項目をクリックします。

   「表示」 > 「ツールバー」 > 「カスタマイズ」

必要に応じて、**「ツールチップの表示」**または**「大きいボタン」**をクリックします。

IBM SPSS Modeler ウィンドウのカスタマイズ

IBM SPSS Modeler インターフェースのさまざまな構成要素間にある仕切りを使用すると、好みに合わせてサイズを変更したり、ツールを閉じることができます。例えば、大きいストリームを使用して作業を行う場合は、各仕切りにある小さい矢印を使用して、「ノード・パレット」、「マネージャー」ペイン、および「プロジェクト」ペインを閉じることができます。これによってストリーム領域が最大になり、大きいストリームや複数のストリーム用に十分な作業スペースを確保することができます。

あるいは、「表示」メニューで、**「ノード・パレット」**、**「マネージャー」**、または**「プロジェクト」**をクリックして、これらの項目の表示をオンまたはオフにします。

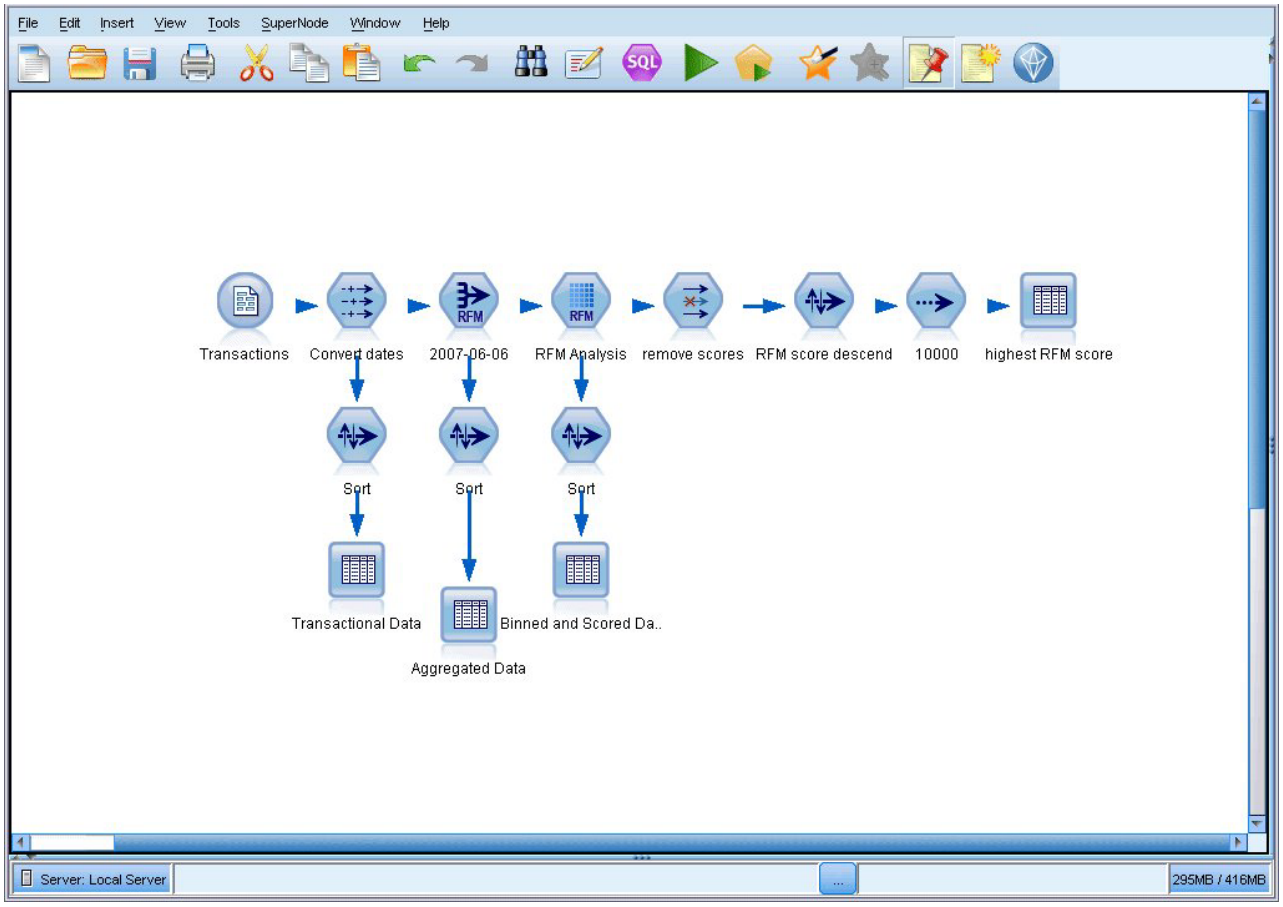


図 10. 最大化されたストリーム領域

「ノード・パレット」、「マネージャー」ペイン、「プロジェクト」ペインを閉じる代わりに、IBM SPSS Modeler ウィンドウの横と下にあるスクロール・バーを使用して垂直方向と水平方向に移動できるスクロール可能なページとしてストリーム領域を使用することもできます。

また、画面マークアップの表示を制御することも可能です。これは、ストリームのコメント、モデル・リンク、およびスコアリング・ブランチの表示で構成されています。この表示をオンまたはオフにするには、次をクリックします。

「表示」 > 「ストリーム・マークアップ」

ストリームのアイコン・サイズの変更

ストリーム・アイコンのサイズは、次の方法で変更できます。

- ストリーム・プロパティーの設定
- ストリーム内のポップアップ・メニュー
- キーボードの使用

ストリーム・ビュー全体を、8% から 200% の間の多くの標準アイコン・サイズのいずれかに拡大縮小できます

ストリーム全体を拡大縮小するには (ストリーム・プロパティーを使用する方法)

1. メインメニューから、次の項目を選択します。

「**ツール**」 > 「**ストリーム・プロパティー**」 > 「**オプション**」 > 「**レイアウト**」。

2. 「アイコン・サイズ」メニューから目的のサイズを選択します。
3. 「**適用**」をクリックして結果を確認します。
4. 「**OK**」をクリックして変更を保存します。

ストリーム全体を拡大縮小するには (メニューを使用する方法)

1. 領域でストリームの背景を右クリックします。
2. 「**アイコン・サイズ**」を選択し、目的のサイズを選択します。

ストリーム全体を拡大縮小するには (キーボードを使用する方法)

1. メイン・キーボードで Ctrl キーを押しながら [-] キーを押して、次に小さいサイズにズームアウトします。
2. メイン・キーボードで Ctrl および Shift キーを押しながら [+] キーを押して、次に大きいサイズにズームインします。

この機能は、複雑なストリームの全体ビューの表示に特に便利です。ストリームの印刷で必要となるページ数を最小限に抑える場合にも使用できます。

IBM SPSS Modeler でのマウスの使用

IBM SPSS Modeler でよく使われるマウス操作を次に示します。

- **シングルクリック**。マウスの右または左ボタンを使用して、メニューからオプションを選択したり、ポップアップ・メニューを開いたり、さまざまな標準のコントロールやオプションにアクセスすることができます。ノードをクリックし、ボタンを押したままマウスを動かして、ノードをドラッグします。
- **ダブルクリック**。マウスの左ボタンを使用してダブルクリックすると、ノードをストリーム領域に置いて、既存のノードを編集できるようになります。
- **中央ボタンのクリック**。マウスの中央ボタンをクリックして、カーソルをドラッグすることにより、ストリーム領域のノードを接続します。マウスの中央ボタンをダブルクリックすると、ノードの接続が解除されます。マウスに中央ボタンがない場合は、代わりに Alt キーを押しながらマウスでクリックしたり、ドラッグしたりします。

ショートカット・キーの使用

IBM SPSS Modeler の多くのビジュアル・プログラミング操作には、ショートカット・キーが割り当てられています。例えば、ノードをクリックして、キーボードの Delete キーを押すと、ノードを削除することができます。同様に、Ctrl キーを押しながら S キーを押すと、ストリームを素早く保存できます。このようなコントロール・コマンドは、Ctrl+S のように、Ctrl キーと他のキーの組み合わせで示されます。

Ctrl+X (切り取り) のように、標準の Windows 操作でも、さまざまなショートカット・キーが使用されています。IBM SPSS Modeler では、後述するアプリケーション固有のショートカット・キーだけでなく、このような標準のショートカットも使用することができます。

注: 場合によっては、IBM SPSS Modeler で使用されていた古いショートカット・キーが Windows 標準のショートカット・キーと競合することがあります。これらの古いショートカット・キーは、Alt キーも一緒に押すと、使用することができます。例えば、Ctrl+Alt+C キーを使用すると、キャッシュのオンとオフを切り替えることができます。

表 1. サポートしているショートカット・キー

ショートカット・キー	関数
Ctrl+A	すべて選択
Ctrl+X	切り取り
Ctrl+N	新規ストリーム
Ctrl+O	ストリームを開く
Ctrl+P	印刷
Ctrl+C	コピー
Ctrl+V	貼り付け
Ctrl+Z	元に戻す
Ctrl+Q	選択したノードの下流にあるすべてのノードを選択
Ctrl+W	下流のすべてのノードの選択を解除 (Ctrl+Q で切り替え)
Ctrl+E	選択したノードから実行
Ctrl+S	現在のストリームを保存
Alt+矢印キー	ストリーム領域上で選択したノードを矢印の方向に移動
Shift+F10	選択したノードのポップアップ・メニューを表示

表 2. 古いホット・キーに対応するショートカット

ショートカット・キー	関数
Ctrl+Alt+D	ノードの複製
Ctrl+Alt+L	ノードの読み込み
Ctrl+Alt+R	ノード名の変更
Ctrl+Alt+U	ユーザー入力ノードの作成
Ctrl+Alt+C	キャッシュのオン/オフの切り替え
Ctrl+Alt+F	キャッシュの取消
Ctrl+Alt+X	スーパーノードの展開
Ctrl+Alt+Z	ズームイン/ズームアウト
Delete	ノードまたは接続の削除

印刷

IBM SPSS Modeler では、次のオブジェクトを印刷できます。

- ストリーム・ダイアグラム
- グラフ作成(G)
- テーブル
- レポート (レポート・ノードおよびプロジェクト・レポートから取得)
- スクリプト (「ストリーム・プロパティー」、「スタンドアロン・スクリプト」、または「スーパーノード・スクリプト」ダイアログ・ボックスから取得)
- モデル (モデル・ブラウザー、現在フォーカスのあるダイアログ・ボックスのタブ、ツリー・ビューアー)

- 注釈 (出力の「注釈」タブを使用)

オブジェクトを印刷するには、次の操作を行います。

- プレビューを行わずにオブジェクトを印刷するには、ツールバーの「印刷」ボタンをクリックします。
- 印刷前にページ設定を行うには、「ファイル」メニューの「**ページ設定**」を選択します。
- 印刷前にプレビューを表示するには、「ファイル」メニューから「**印刷プレビュー**」を選択します。
- 標準の「印刷」ダイアログ・ボックスに、選択しているプリンターのオプションを表示して、外観オプションを指定するには、「ファイル」メニューから「**印刷**」を選択します。

IBM SPSS Modeler の自動化

高度なデータ・マイニング作業は、複雑で時間がかかる処理になる可能性もあるため、IBM SPSS Modeler には、さまざまなタイプのコーディングおよび自動化のサポート機能が用意されています。

- **Control Language for Expression Manipulation** (CLEM) は、IBM SPSS Modeler ストリーム中を流れるデータの分析と操作を行うための言語です。データ・マイニング作業者は、CLEM を使用して、経費および収益データから利益を算出するような簡単な作業から、Web ログ・データを有益な情報を含むフィールドやレコードに変換するような複雑な作業まで、さまざまなストリーム操作を行うことができます。
- **スクリプト**は、ユーザー・インターフェースのプロセスを自動化するための強力なツールです。スクリプトは、マウスやキーボードを使用して実行するのと同様の操作を実行できます。また、出力を指定して、生成されたモデルを操作することができます。

第 3 章 モデル作成の概要

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、過去の申請者に関する既知の情報に基づき、融資申請者のリスクが低いか高いかを予測することができます。

結果を予測できるようになることが、予測分析の主な目標であり、モデル作成プロセスを理解することは、IBM SPSS Modeler を使用する上での鍵となります。

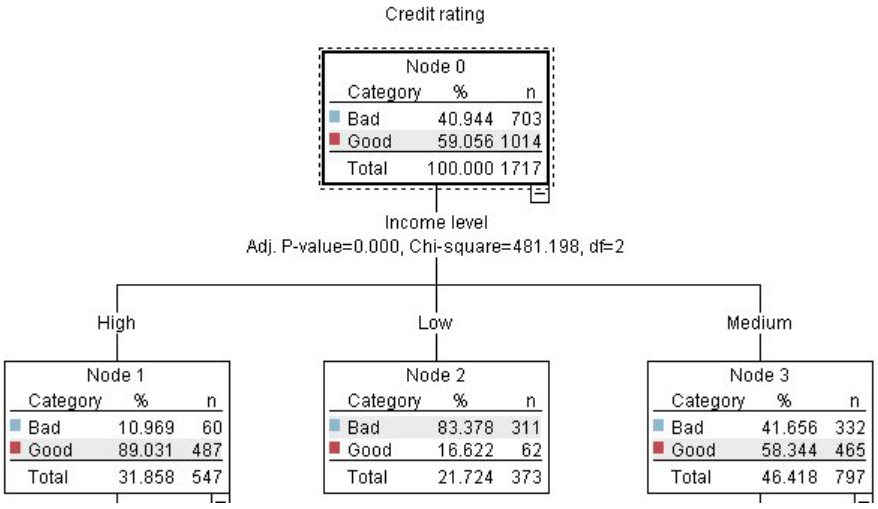


図 11. 簡単なディシジョン・ツリー・モデル

この例では、**ディシジョン・ツリー**・モデルを使用します。このモデルは、一連のディシジョン・ルールを使用してレコードを分類します (また、レスポンスを予測します)。以下に例を示します。

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

この例では、CHAID (カイ 2 乗自動相互検出) モデルを使用しますが、一般的な概要を説明するためにこれを使用しており、ほとんどの概念は IBM SPSS Modeler の他のモデル・タイプにも広く適用されます。

モデルを理解するには、まずそれに入力す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	信用格付け: 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 件未満、2=2 件以上

銀行は、ローンを返済したか (信用格付け = 良い) 返済していないか (信用格付け = 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する必要があります。

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* フォルダーの *streams* サブフォルダー内にある *modelingintro.str* というストリームを使用します。データ・ファイルは *tree_credit.sav* です。詳細については、5 ページの『「Demos」フォルダー』を参照してください。

では、ストリームを見ていきます。

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* というファイルをダブルクリックします。

ストリームの構築

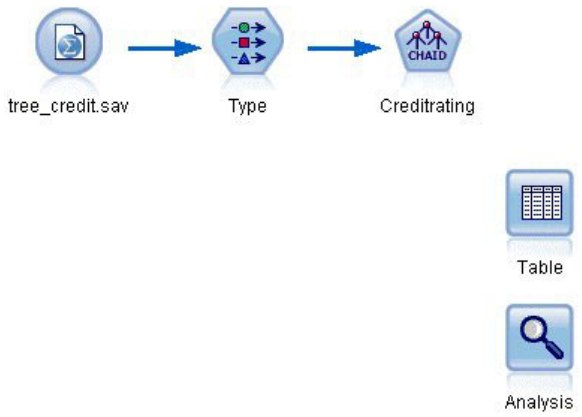


図 12. モデル作成ストリーム

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 (ここでは、IBM SPSS Statistics データ・ファイル) からデータを読み込む入力ノード。
- 測定の尺度 (フィールドが含んでいるデータの種類) などのフィールド・プロパティー、および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を指定する入力ノードまたはデータ型ノード。
- 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ます。 CHAID (Chi-squared Automatic Interaction Detection) は、カイ 2 乗統計として知られる特定のタイプの統計を使用してディシジョン・ツリー内で分割するための最適な場所を特定することでディシジョン・ツリーを構築する分類方法です。

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削除できます。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後にスコアリング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、*Demos* フォルダーにインストールされている *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。(現在の IBM SPSS Modeler インストール環境のこのフォルダーを参照するために、*$CLEO_DEMOS* という特殊変数が使用されます。これにより、現在のインストール・フォルダーやバージョンに関係なく、パスが確実に有効になります)。

図 13. Statistics ファイル入力ノードを使用したデータの読み取り

データ型ノードは、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。使用するソース・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*」フィールドなど) には「*悪い*」、「*良い*」、「*クレジット履歴なし*」などの複数の異なる値があります。**順序型**フィールド (「*収入レベル*」フィールドなど) は、特有の順序を持つ複数の値 (この場合は、「*低*」、「*中*」、および「*高*」) を含むデータについて説明します。

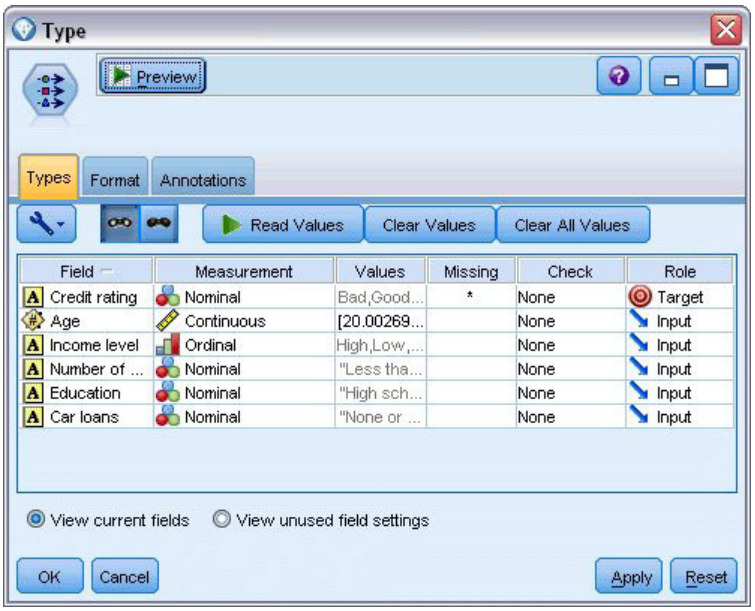


図 14. データ型ノードによる対象フィールドおよび入力フィールドの設定

フィールドごとに、データ型ノードは**役割**を指定し、モデル作成で各フィールドが果たす役割も示します。「*信用度*」フィールドの役割は「*対象*」に設定されています。これは、特定の顧客が債務不履行になったかどうかを示すフィールドです。これは、**対象** (値を予測する対象のフィールド) です。

他のフィールドの役割は、「入力」に設定されています。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。モデル作成アルゴリズムは、このフィールドの値を使用して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。

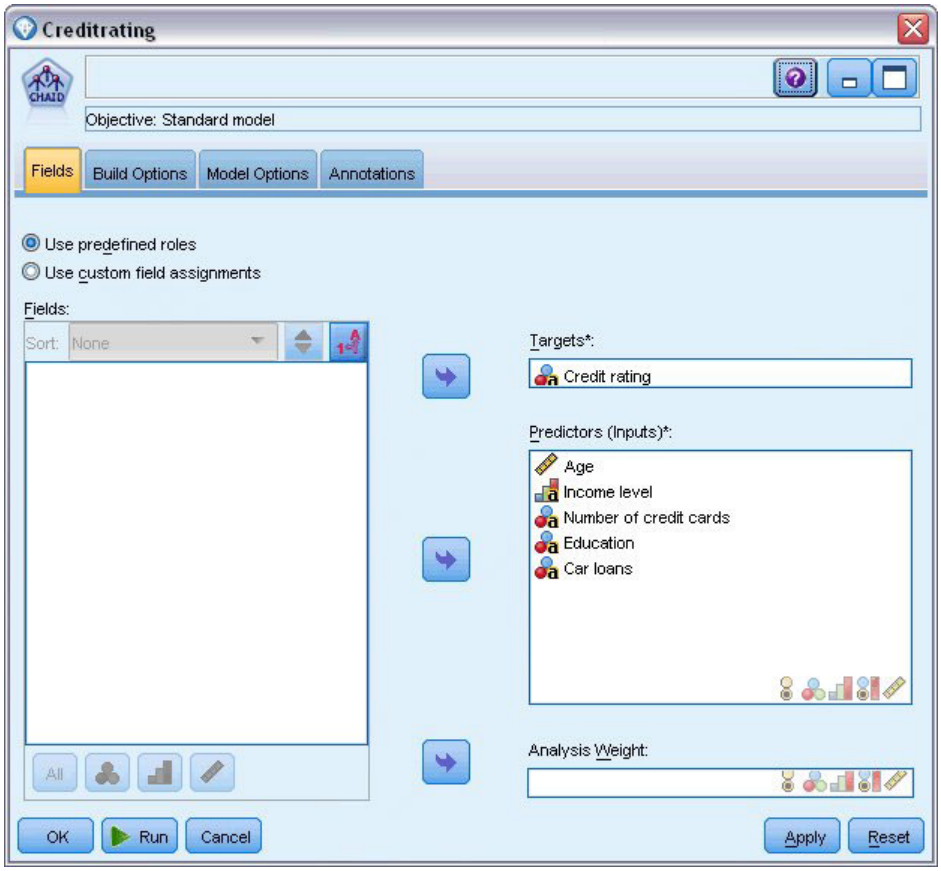


図 15. CHAID モデル作成ノードの「フィールド」タブ

ここでは、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。

新規モデルが必要なので、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・モデルが必要なので、デフォルトの目的オプション「**単一ツリーを作成**」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、モデルを微調整することも可能ですが、この例ではデフォルトのモード設定「**モデルの生成**」を使用して簡単にモデルを生成します。

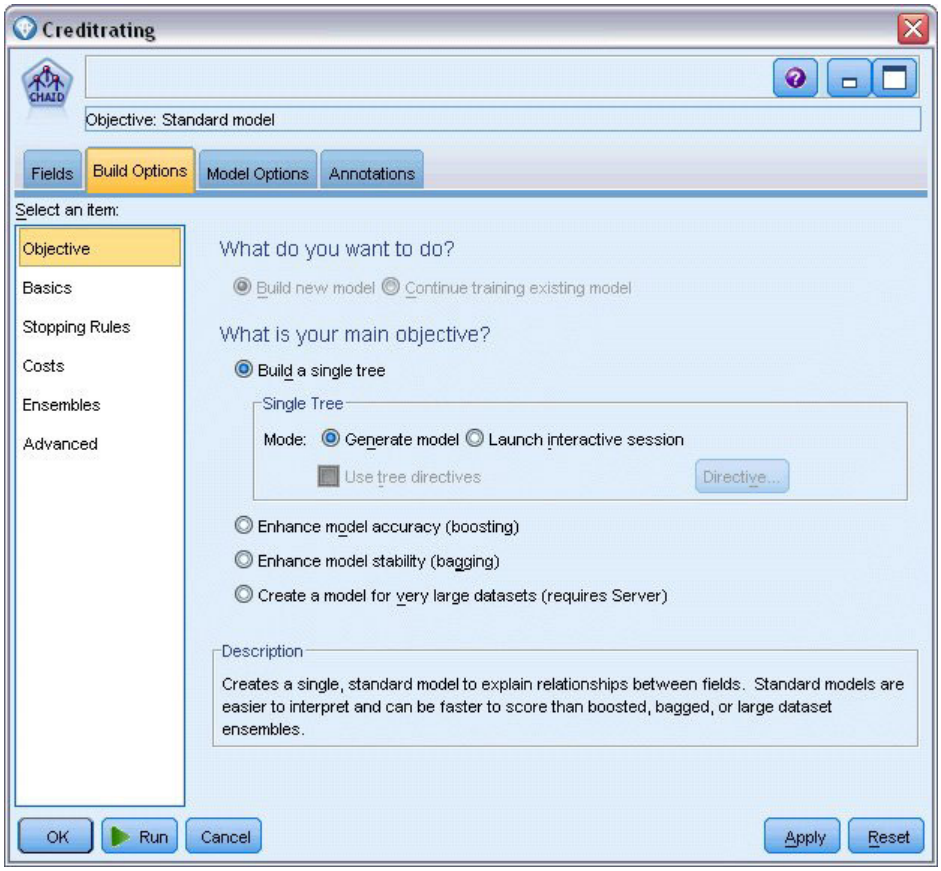


図 16. CHAID モデル作成ノードの「作成オプション」タブ

この例では、ツリーを非常にシンプルな状態に保つため、親ノードおよび子ノードのケースの最小数を大きくして、ツリーの成長を制限します。

2. 「作成オプション」タブで、左側のナビゲーター・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。

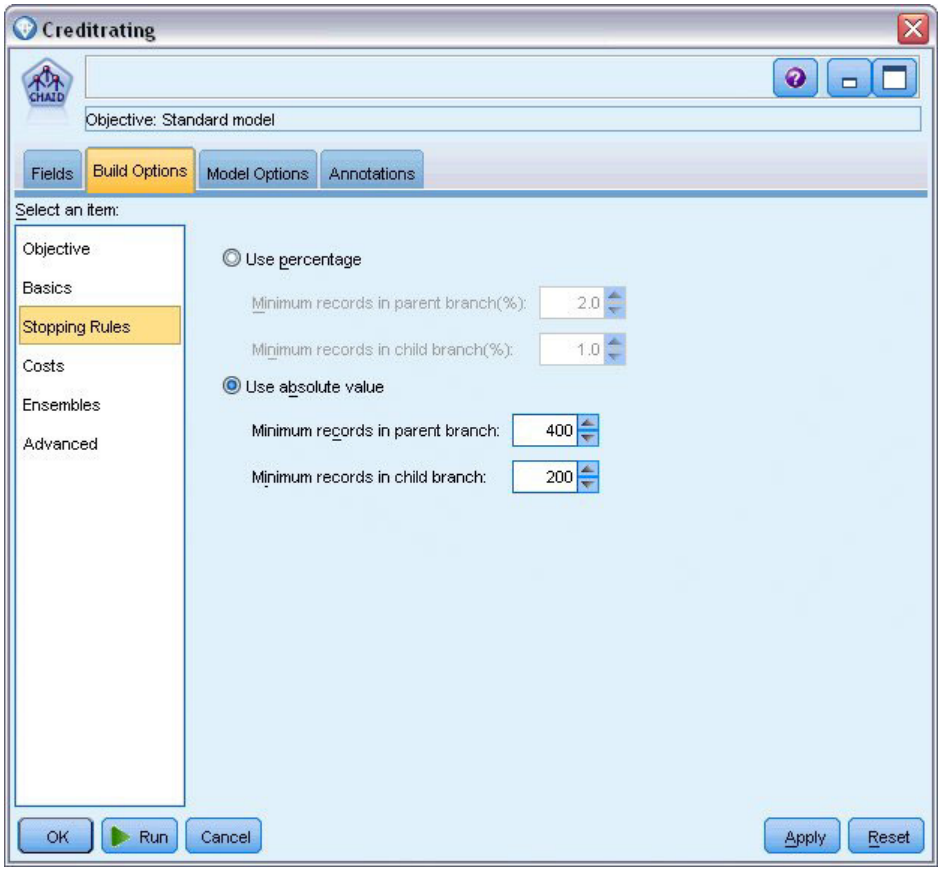


図 17. ディシジョン・ツリー構築の停止基準の設定

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・ウィンドウの右上隅の「モデル」パレットに追加されます。また、ストリーム領域内にも配置され、モデルが作成されたモデル作成ノードへのリンクもあります。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、モデル・パレットの「**ブラウズ**」、または領域の「**編集**」を選択します。

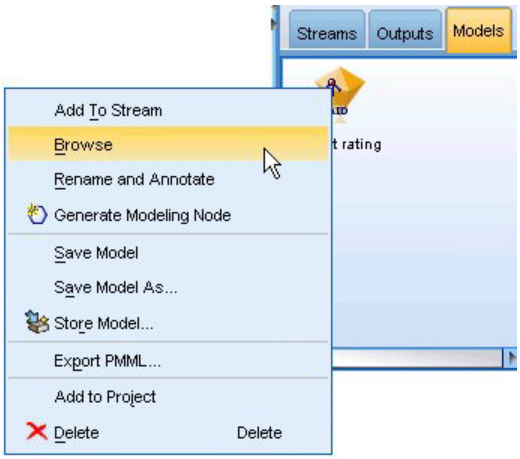


図 18. モデル・パレット

CHAID ナゲットの場合、「モデル」タブには、ルール・セットの形式で詳細が表示されます。これは基本的に、さまざまな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のに使用できる一連のルールです。

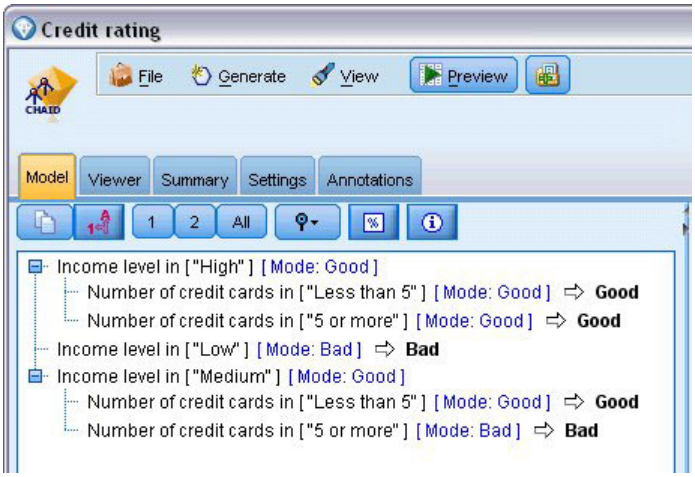


図 19. CHAID モデル・ナゲット、ルール・セット

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) の場合、「良い」または「悪い」という予測が返されます。どちらの場合でも、予測はそのノード内に収まるレコードの**モード** (最も一般的な応答) によって決定されます。

ルール・セットの右側の「モデル」タブには、予測値の重要度のグラフが表示されます。このグラフには、モデル推定時の各予測値の相対的な重要度が示されます。このことから、「収入レベル」がこの場合最も有意であり、その他では唯一「クレジット・カード数」が有意であることが分かります。

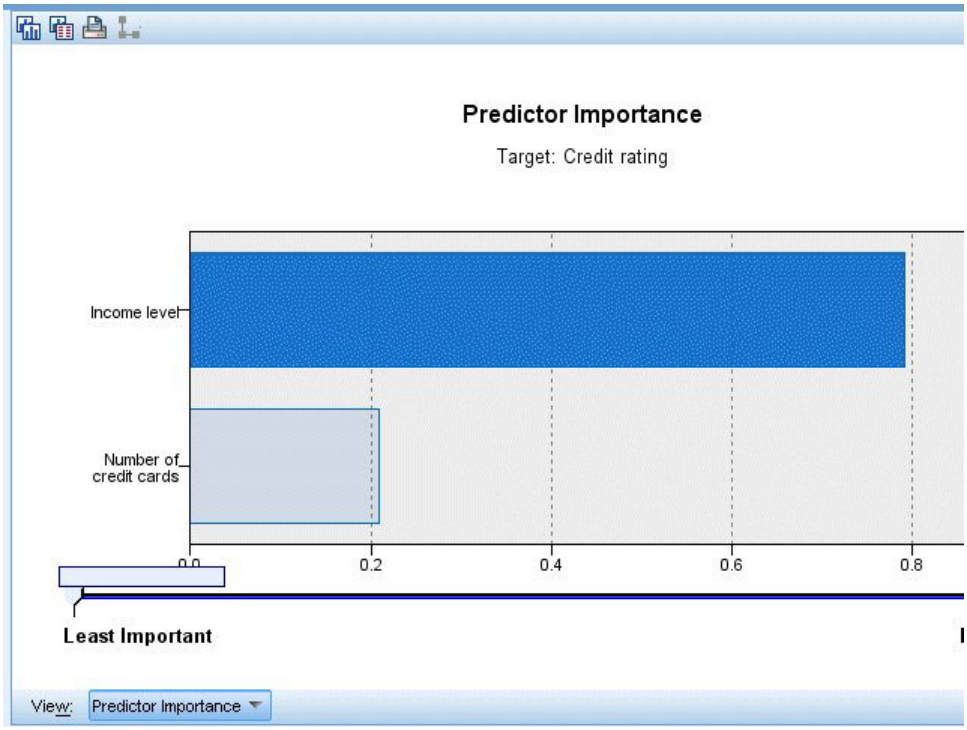


図 20. 予測値の重要度グラフ

モデル・ナゲットの「ビューアー」タブでは、同じモデルが、各ディシジョン・ポイントにノードを配したツリーの形式で表示されます。ツールバーの「ズーム」コントロールを使用すると、特定のノードにズームインしたり、ズームアウトしてツリーのより広い範囲を表示したりすることができます。

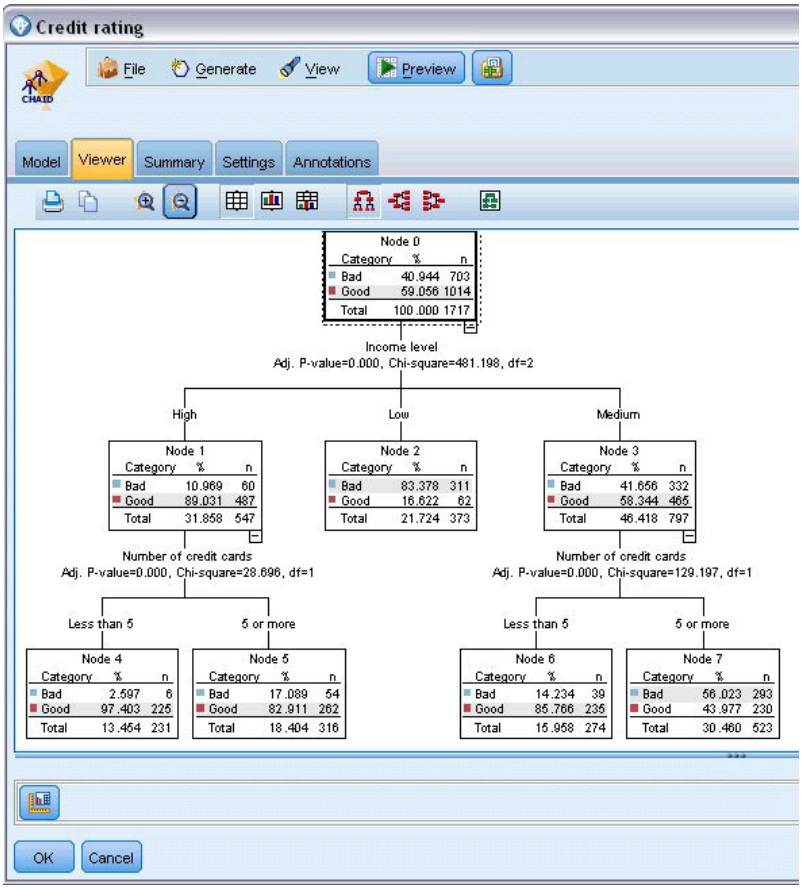


図 21. ズームアウトが選択されたモデル・ナゲットの「ビューアー」タブ

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ています。データ・セット内の 40% を少し超えるケースが、高リスクと分類されています。これは相当高い比率であるため、原因となっている可能性がある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* によるものであることが分かります。収入レベルが「低」のカテゴリーのレコードは、ノード 2 に割り当てられています。このカテゴリーに高い割合の債務不履行者が含まれているのが分かりますが、驚くことではありません。当然、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、このカテゴリーの顧客の 16% は、実際には債務不履行になっておらず、予測が常に正しいとは限りません。すべてのレスポンス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、各レコードの*最も可能性が高い* レスポンスを予測することを可能にする必要があります。

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、一方で、これらの顧客の 10 人に 1 人を上回る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。

保有しているクレジット・カードの数に基づいて、モデルがこれらの顧客を 2 つのサブカテゴリー (ノード 4 および 5) に分類した方法に注目してください。高収入の顧客について、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% に上昇し、さらに満足のいく結果になりま

す。

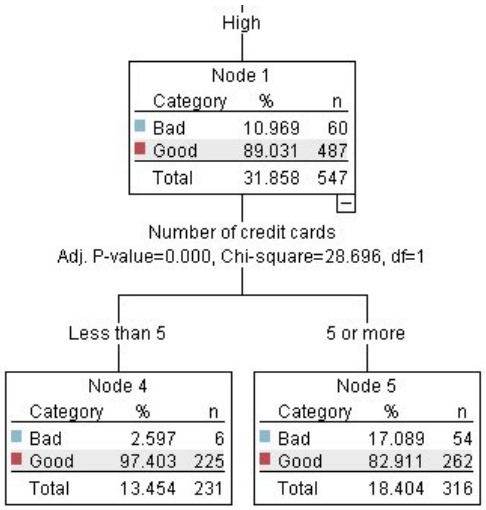


図 22. 高収入の顧客のツリー・ビュー

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。

また、サブカテゴリー (この場合ノード 6 および 7) も役立つことがあります。今度は、クレジット・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」の評価のパーセンテージが 58% から 85% に上昇し、大幅な改善が示されます。

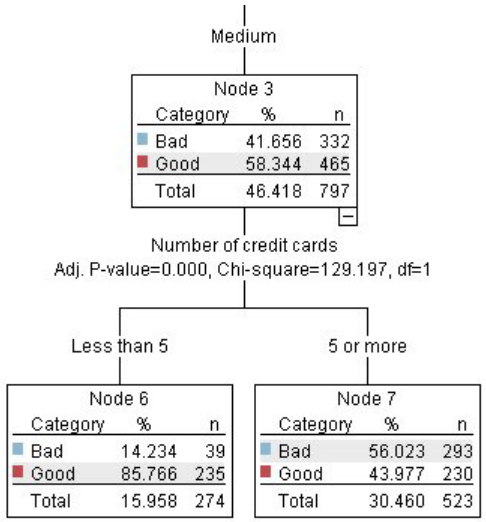


図 23. 中程度の収入の顧客のツリー・ビュー

ここまでで、このモデルに入力されたすべてのレコードは特定のノードに割り当てられ、各ノードの最も一般的な応答に基づいて、「良い」または「悪い」の予測が割り当てられると説明しました。

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**と呼ばれます。モデルを推定するの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 (結果を知っているデータ) に対してどれだけ正確に実行できるかを評価できます。スコアリングを行う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レスポンスと実際の結果とを比較する必要があります。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、観測レスポンスと予測レスポンスとを比較することができます。

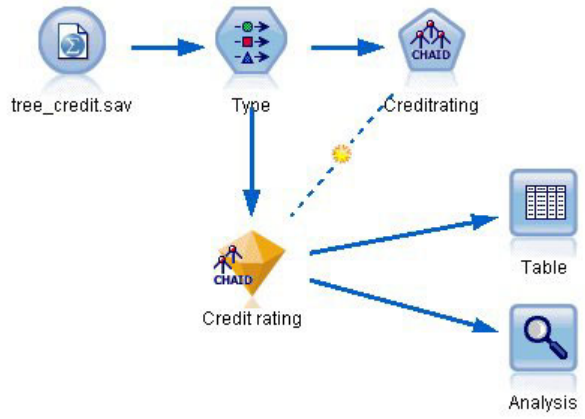
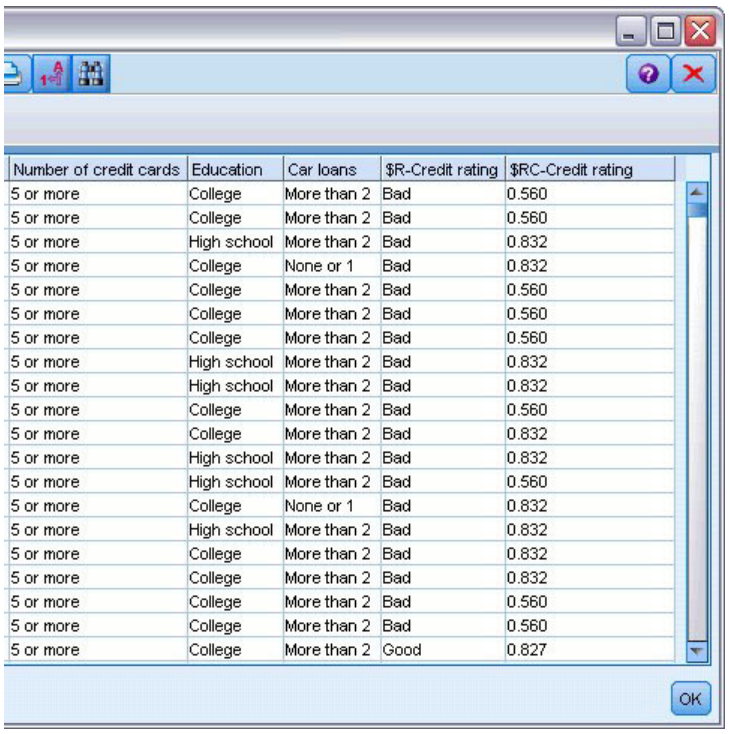


図 24. モデル評価のための出力ノードへのモデル・ナゲットの接続

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。

   テーブルに、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、実際のレスポンスが含まれている元の「*信用度*」フィールドと比較できます。

   規則により、スコアリングの間に生成されるフィールドの名前は対象フィールドを基にしていますが、予測値には *$R-*、信頼度値には *$RC-* といった標準の接頭辞が付きます。それぞれのモデル・タイプで異なる接頭辞が使用されます。**信頼度値**は、各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あり、スケールは 0.0 から 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 25. 生成されたスコアおよび信頼度値を示すテーブル

期待どおり、多くのレコードについては予測値と実際のレスポンスが一致していますが、すべてが一致しているわけではありません。各 CHAID ターミナル・ノードにレスポンスが混在しているのが理由です。予測は、*最も一般的* なレスポンスと一致していますが、そのノードの他のすべてのレスポンスは一致していません。(低収入の顧客の、債務不履行に陥っていない 16% の少数派を思い出してください)。

これを回避するには、すべてのノードが 100% ピュア (つまり、すべてが「*良い*」または「*悪い*」でレスポンスで、レスポンスが混在していない) になるまで、ツリーを小さい枝に分割し続けます。ただし、そのようなモデルは非常に複雑であり、恐らく他のデータ・セットにうまく一般化できません。

正しい予測の数を正確に確認するには、テーブル全体を読み、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数えます。幸いなことに、はるかに簡単な方法が用意されています。精度分析ノードを使用すれば、これを自動的に行うことができます。

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。

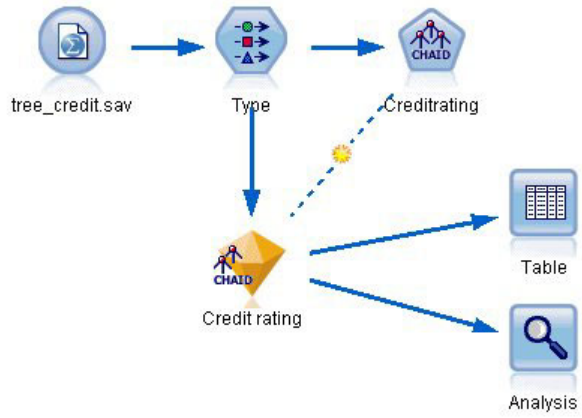


図 26. 精度分析ノードの接続

分析により、2464 個のレコード中 1899 個 (77% 超) で、モデルによって予測された値と実際のレスポンスが一致したことが分かります。

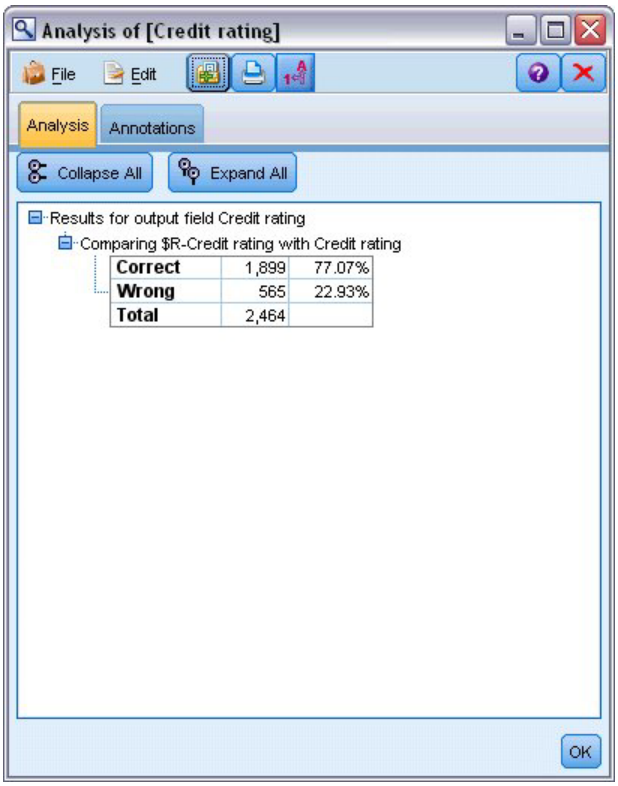


図 27. 観測レスポンスと予測レスポンスの比較の分析結果

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うことができます。

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モデルのテストに使用することにより、モデルがどの程度うまく他のデータ・セットに一般化できるかについてのより優れた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、既に実際の結果が分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、結果の分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、現在銀行の顧客ではないが、販促メールで見込み客対象となる人々が含めることができます。

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。ここでは、モデルの作成に使用される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。

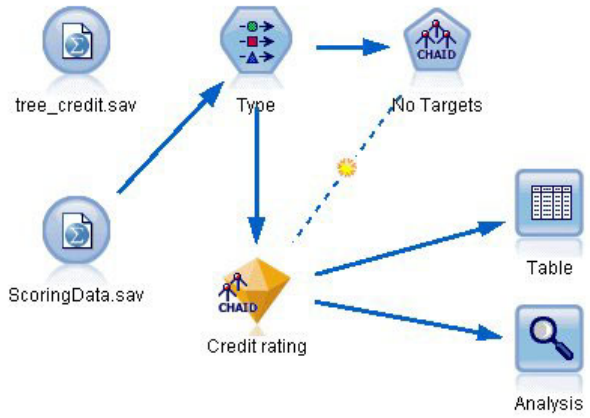


図 28. スコアリング用の新規データの接続

Statistics ファイル入力ノードを更新して別のデータ・ファイルを指すようにするか、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できます。どちらの場合でも、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る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、期待される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ことができます。フィールド名とタイプがモデルによって使用されるものと同じである限り、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ーに関係なく、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、結果が分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。既に結果が分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、そのパフォーマンスを評価できます。
- モデルのパフォーマンスが十分であると満足したら、新しいデータ (見込み客など) のスコアリングを行って、そのレスポンス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 4 章 フラグ型対象の自動化モデル作成

顧客のレスポンスのモデル作成 (自動分類)

自動分類ノードでは、フラグ型 (特定の顧客が債務不履行になる可能性が高いかどうか、特定の提案に反応するかどうかなど) または名義型 (セット型) 対象のさまざまなモデルを自動的に作成して比較できます。この例では、フラグ型 (「はい」または「いいえ」) の結果を検索します。比較的単純なストリームで、ノードは候補モデル・セットを生成およびランク付けし、最善のモデルを選択して、単一の集計済み (アンサンブル) モデルに結合します。この方法は自動化の容易さと複数モデルの結合の利点を組み合わせるため、多くの場合、単一のモデルから取得するよりも精度が高い予測が得られます。

この例は、それぞれの顧客に合った適切な提案を行うことで、さらに収益の高い結果を実現することを望んでいる架空の会社に基づいています。

この方法では、自動化の利点を強調します。連続型 (数値範囲型) の対象を使用する同様の例については、51 ページの『プロパティー値 (自動数値)』を参照してください。

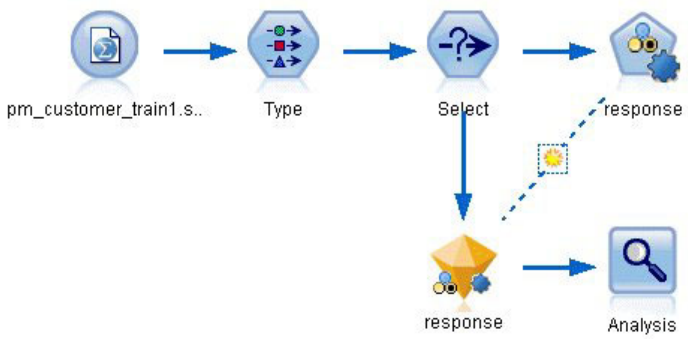


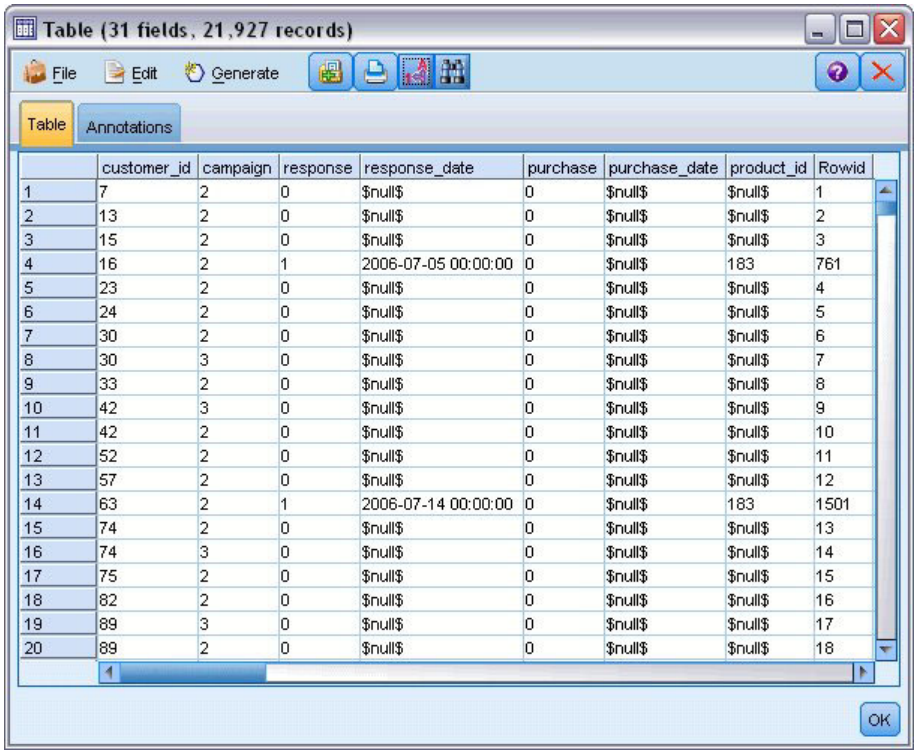
図 29. 自動分類のサンプル・ストリーム

この例では、Demo フォルダーの *streams* にインストールされているストリーム *pm_binaryclassifier.str* を使用します。使用するデータ・ファイルは *pm_customer_train1.sav* です。詳細については、『履歴データ』を参照してください。

履歴データ

ファイル *pm_customer_train1.sav* には、過去のキャンペーンでの特定のお客様に対する提案を追跡する履歴データが含まれており、「*campaign*」フィールドの値で示されています。最大多数のレコードが *Premium account* キャンペーンに分類されています。

campaign フィールドの値は、実際には、データ内で整数としてコード化されます (例えば、*2 = Premium account*)。後でより分かりやすい出力を作成するために使用するラベルを、これらの値に対して定義します。



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

図 30. 以前の販売促進活動に関するデータ

また、ファイルには、提案が受け入れられたかどうか (0 = 「いいえ」、1 = 「はい」) を示す *response* フィールドも含まれています。これは、予測する**対象フィールド** (または値) になります。各顧客についての人口統計情報および財務情報を含むさまざまなフィールドも含まれています。これらを使用して、収入、年齢、月ごとの取引数などの特性に基づいて、個人またはグループのレスポンス率を予測するモデルを構築 (「学習」) することができます。

ストリームの構築

1. IBM SPSS Modeler インストール環境の *Demos* フォルダーにある *pm_customer_train1.sav* を指し示す Statistics ファイル入力ノードを追加します。(このフォルダーを参照するショートカットとして、ファイル・パスに `$CLEO_DEMOS/` を指定できます。表示されているとおり、パスにはバックスラッシュではなく普通のスラッシュを使用する必要があります)。

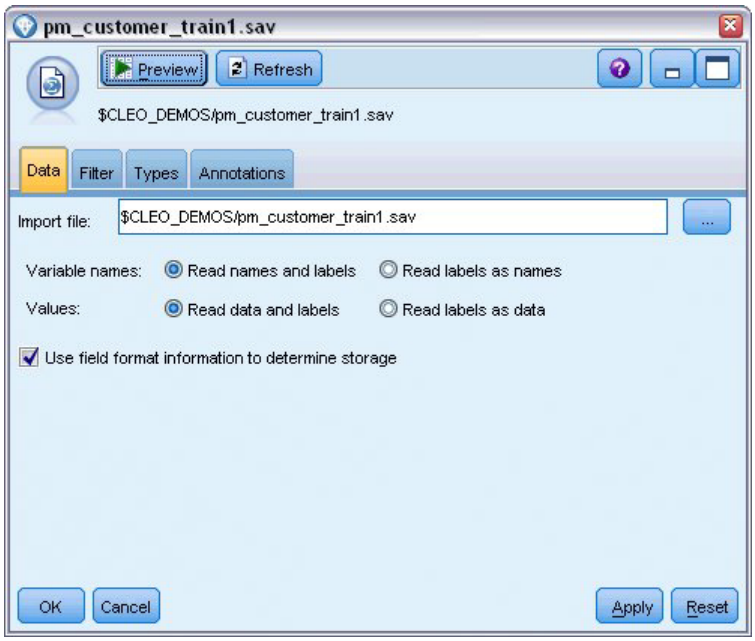


図 31. データの読み取り

2. データ型ノードを追加し、*response* を対象フィールドとして選択します (役割は「**対象**」)。このフィールドの測定を「**フラグ**」に設定します。

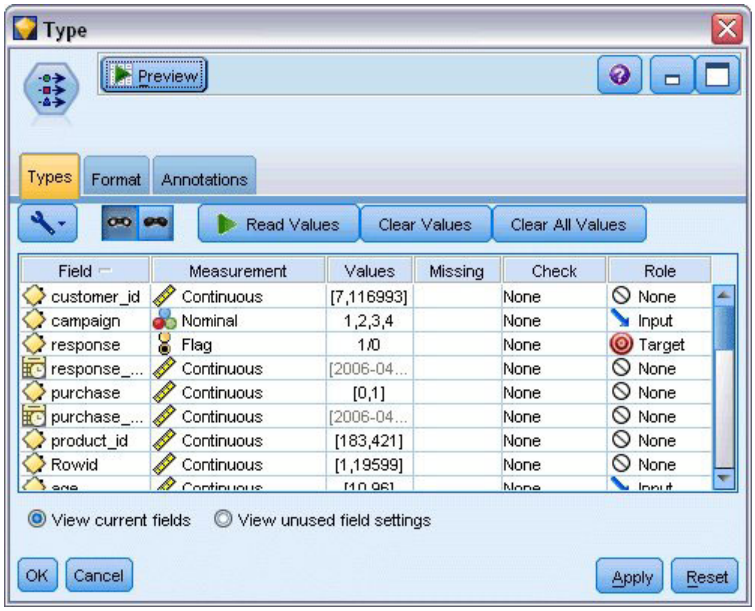


図 32. 測定の尺度および役割の設定

3. フィールド *customer_id*、*campaign*、*response_date*、*purchase*、*purchase_date*、*product_id*、*Rowid*、および *X_random* の役割を「**なし**」に設定します。これらのフィールドは、モデルの構築時には無視されます。
4. データ型ノードで「**値の読み込み**」ボタンをクリックして、値がインスタンス化されていることを確認します。

前述のとおり、ソース・データには、異なる種類の顧客アカウントを対象とした、4 つの異なるキャンペーンに関する情報が含まれています。これらのキャンペーンは、データ内で整数でコード化されるため、各整数が示すアカウント・タイプを分かりやすくするために、それぞれにラベルを定義します。

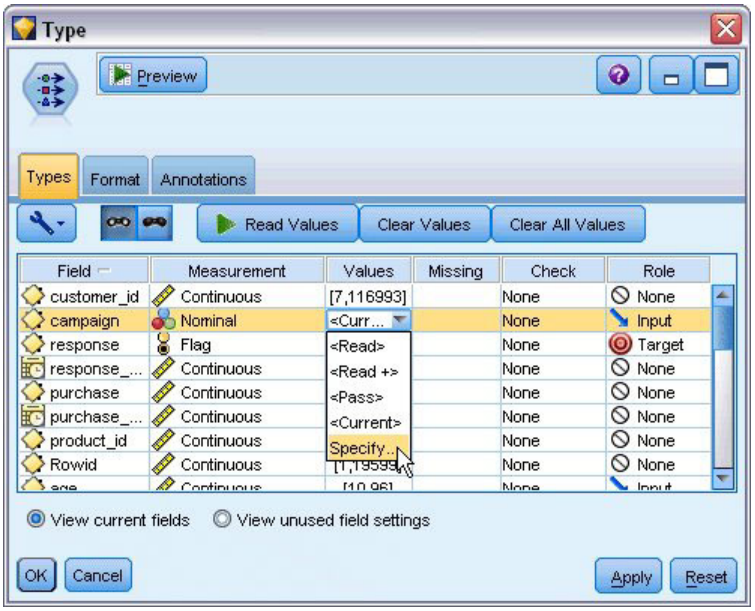


図 33. フィールドの値の指定

5. 「**キャンペーン**」 フィールドの行で、「**値**」列の項目をクリックします。
6. ドロップダウン・リストから「**指定**」を選択します。

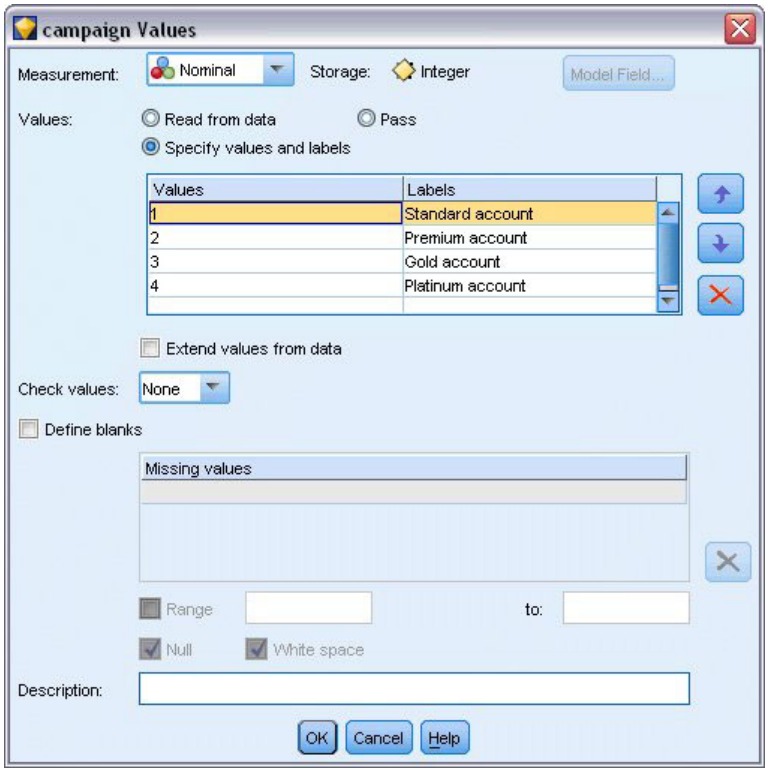
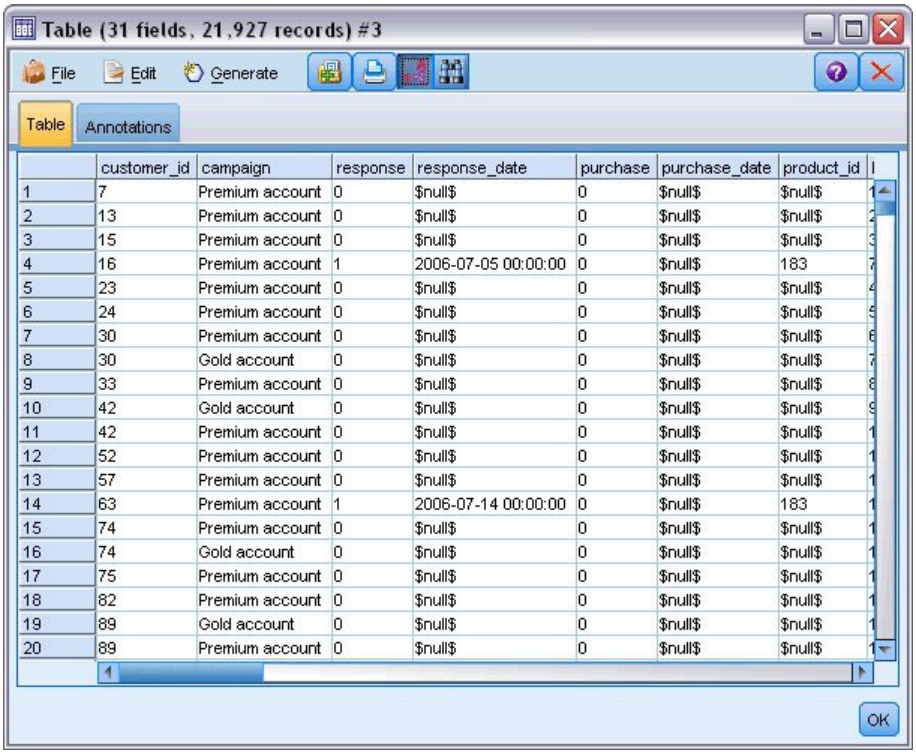


図 34. フィールド値のラベルの定義

7. 「**ラベル**」列に、「**キャンペーン**」 フィールドの 4 つの値それぞれに表示されるラベルを入力します。
8. 「**OK**」をクリックします。

これで、出力ウィンドウに、整数ではなくラベルを表示できるようになりました。



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id
1	7	Premium account	0	$null$	0	$null$	$null$
2	13	Premium account	0	$null$	0	$null$	$null$
3	15	Premium account	0	$null$	0	$null$	$null$
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183
5	23	Premium account	0	$null$	0	$null$	$null$
6	24	Premium account	0	$null$	0	$null$	$null$
7	30	Premium account	0	$null$	0	$null$	$null$
8	30	Gold account	0	$null$	0	$null$	$null$
9	33	Premium account	0	$null$	0	$null$	$null$
10	42	Gold account	0	$null$	0	$null$	$null$
11	42	Premium account	0	$null$	0	$null$	$null$
12	52	Premium account	0	$null$	0	$null$	$null$
13	57	Premium account	0	$null$	0	$null$	$null$
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183
15	74	Premium account	0	$null$	0	$null$	$null$
16	74	Gold account	0	$null$	0	$null$	$null$
17	75	Premium account	0	$null$	0	$null$	$null$
18	82	Premium account	0	$null$	0	$null$	$null$
19	89	Gold account	0	$null$	0	$null$	$null$
20	89	Premium account	0	$null$	0	$null$	$null$

図 35. フィールド値ラベルの表示

9. テーブル・ノードをデータ型ノードに接続します。
10. テーブル・ノードを開いて、「**実行**」をクリックします。
11. 出力ウィンドウで、「**出力にフィールドと値ラベルを表示**」ツールバー・ボタンをクリックしてラベルを表示します。
12. 「**OK**」をクリックして、出力ウィンドウを閉じます。

データには 4 つの異なるキャンペーンに関する情報が含まれますが、分析は一度に 1 つのキャンペーンに集中して行います。最も多くのレコードは Premium アカウント・キャンペーンに分類される (データでは *campaign = 2* にコード化されている) ため、条件抽出ノードを使用してこれらのレコードのみがストリームに含まれるようにできます。

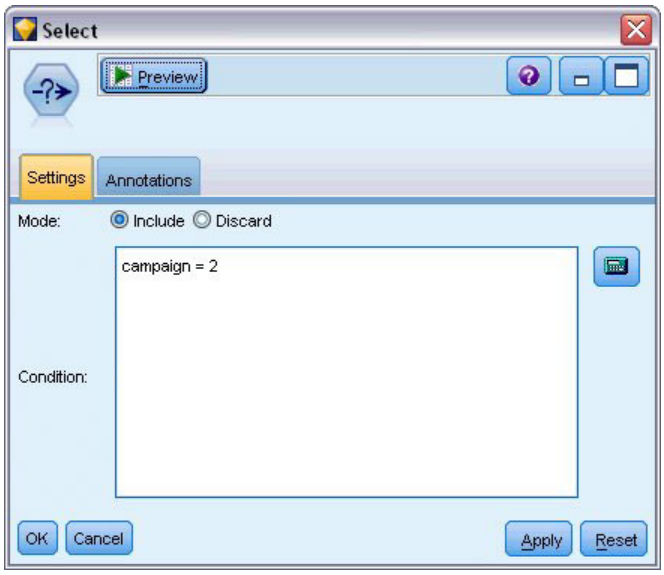


図 36. 単一キャンペーンのレコードの選択

モデルの生成およびキャンペーン

1. 自動分類ノードを接続して、「**全体の精度**」をモデルのランク付けに使用する測定基準として選択します。
2. 「**使用するモデルの数**」を 3 に設定します。 つまり、ノードを実行すると 3 つの最適モデルが作成されます。

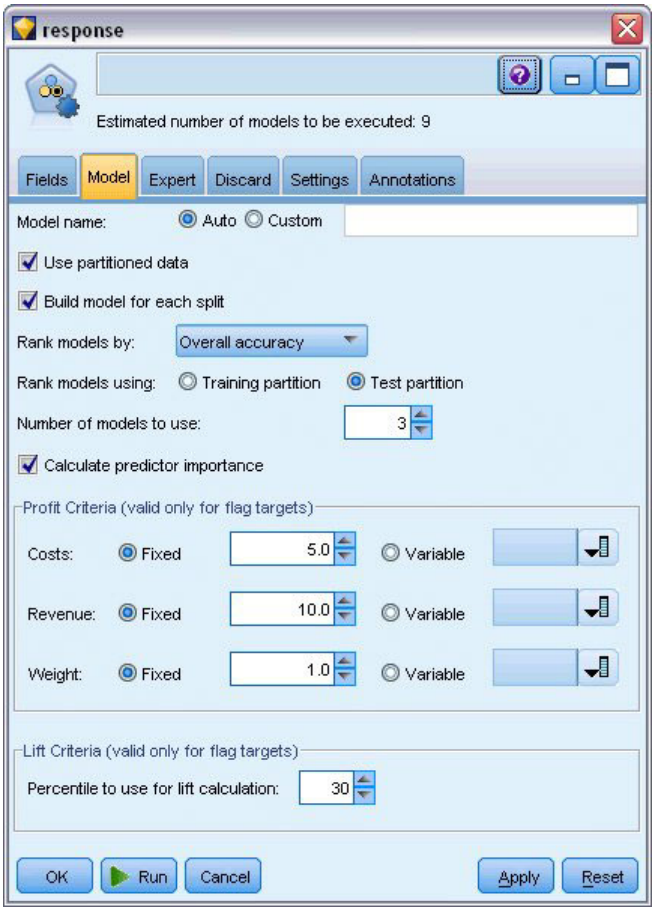


図 37. 自動分類ノードの「モデル」タブ

「エキスパート」タブで、最大 11 のモデル・アルゴリズムから選択できます。

3. 「**判別分析**」および「**SVM**」モデル・タイプの選択を解除します。(これらのモデルはこれらのデータの学習により時間がかかるため、選択を解除して例の時間を短縮します。時間がかかってもかまわない場合、選択したままにしてもよいです)。

「モデル」タブで 「**使用するモデル数**」を 3 に設定しているため、ノードは残りの 9 つのアルゴリズムの精度を計算し、3 つの最も正確なアルゴリズムを含む単一のモデル・ナゲットを構築します。

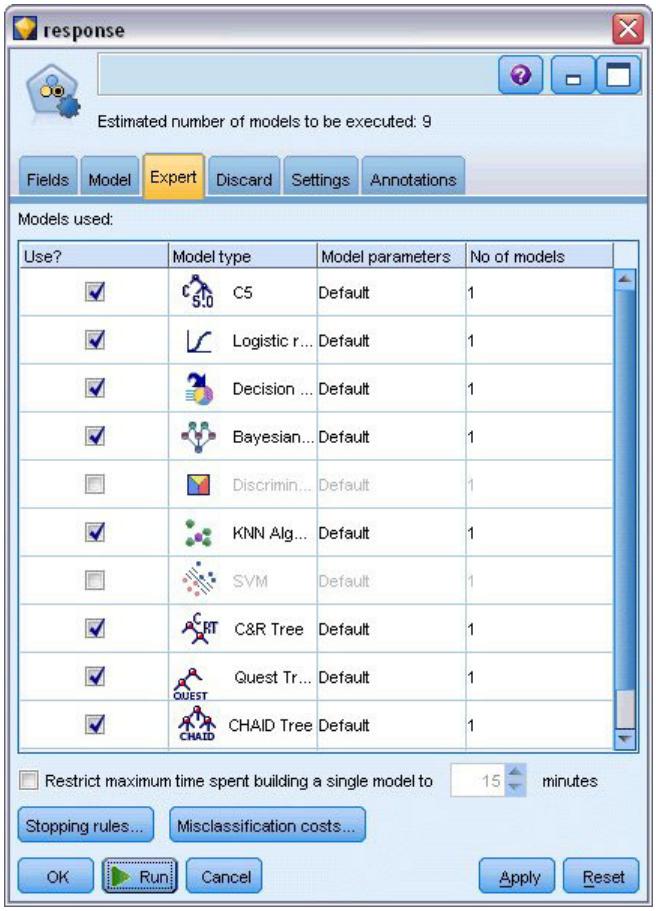


図 38. 自動分類ノードの「エキスパート」タブ

4. 「設定」タブで、アンサンブル方法について、「**信頼度-重み付き票決**」を選択します。これにより、単一の集計済みスコアが各レコードに作成される方法が決まります。

   単純な票決では、3 つのモデルのうち 2 つが「はい」と予測した場合、「はい」が 2 対 1 の票決で勝ちます。信頼度-重み付き票決では、各予測の信頼度値に基づいて、票決が重み付けされます。そのため、2 つの「はい」の予測を組み合わせたものより高い信頼度で 1 つのモデルが「いいえ」と予測する場合、「いいえ」が勝利します。

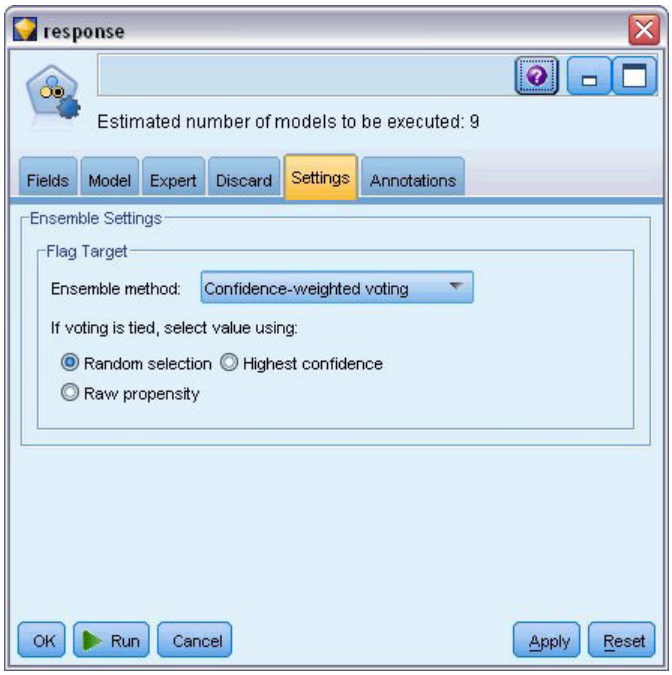


図 39. 自動分類ノード: 「設定」タブ

5. 「**実行**」をクリックします。

数分後、生成されたモデル・ナゲットが作成され、領域内、およびウィンドウの右上隅の「モデル」パレットに配置されます。モデル・ナゲットをさまざまな方法で参照、保存または展開することができます。

モデル・ナゲットを開きます。実行時に作成された各モデルの詳細がリストされます。(大規模なデータ・セットで何百ものモデルが作成されることがある現実の状況では、これには何時間もかかることがあります)。 39 ページの図 29を参照してください。

個々のモデルをより詳細に探る場合、「**モデル**」列のモデル・ナゲットのアイコンをダブルクリックし、個別のモデルの結果をドリルダウンして参照することができます。そこから、モデル作成ノード、モデル・ナゲット、または評価グラフを生成することができます。「**グラフ**」列で、サムネー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。

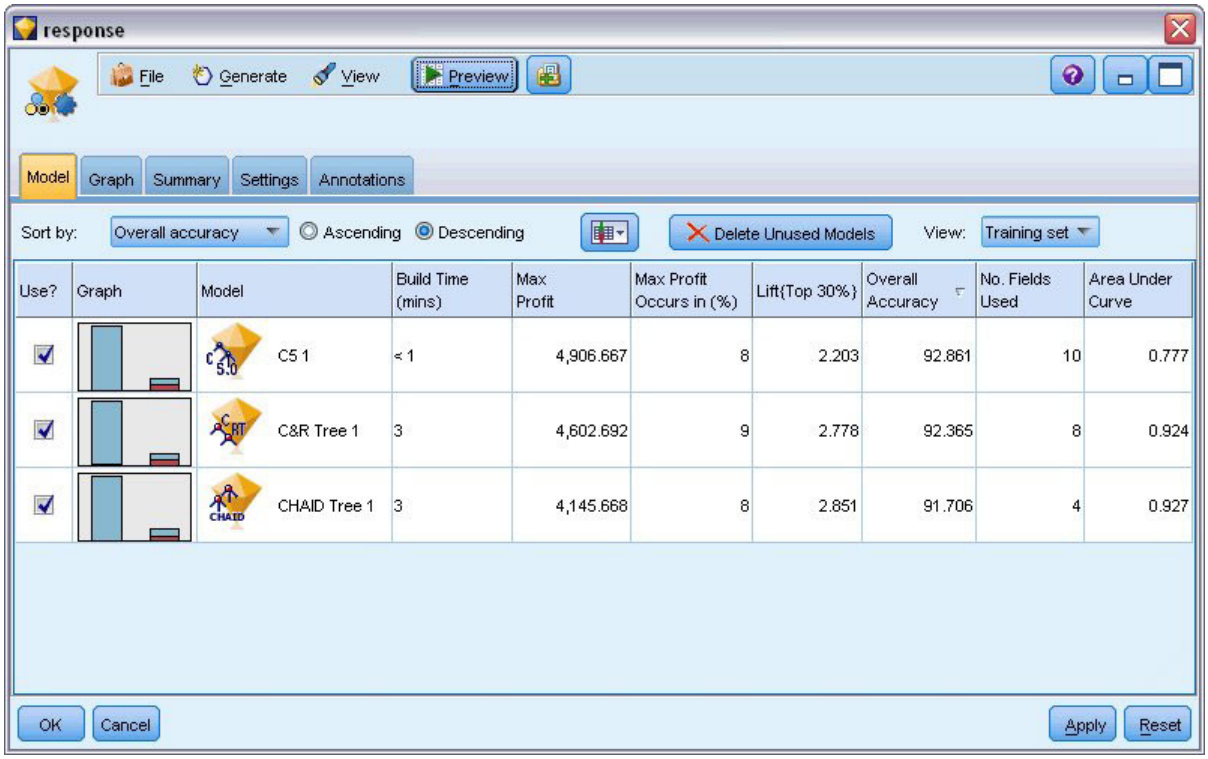


図 40. 自動分類の結果

デフォルトでは、自動分類ノードの「モデル」タブで選択していた指標である、全体の精度に基づいてソートされます。C51 モデルはこの指標では最も優れていますが、C&R Tree、CHAID モデルも同じくらい正確です。

列の見出しをクリックして異なる列でソートすることができます。あるいは、ツールバーの「**ソート基準**」ドロップダウン・リストから目的の指標を選択することもできます。

これらの結果に基づいて、3 つの最も正確なモデルをすべて使用するように指定します。複数モデルの予測を組み合わせることにより、個々のモデルの制限を回避でき、全体の精度がより高くなります。

「**使用?**」列で C51、C&R Tree、および CHAID モデルを選択します。

モデル・ナゲットの後に、精度分析ノード (「出力」パレット) を接続します。精度分析ノードを右クリックし、「**実行**」をクリックしてストリームを実行します。

アンサンブル・モデルで生成された集計済みスコアは、*$XF-response* という名前のフィールドに表示されます。学習データに対して測定する場合、予測値は実際の応答 (元の *response* フィールドに記録されているもの) に 92.82% の全体の精度で一致します。

このケースの 3 つの個別モデルの中で最善のモデル (C51 の 92.86%) ほど正確でありませんが、差は非常に小さいため、無視できます。一般的に、学習データ以外のデータ・セットに適用した場合、通常、アンサンブル・モデルは優れたパフォーマンスを示すことが多くなっています。

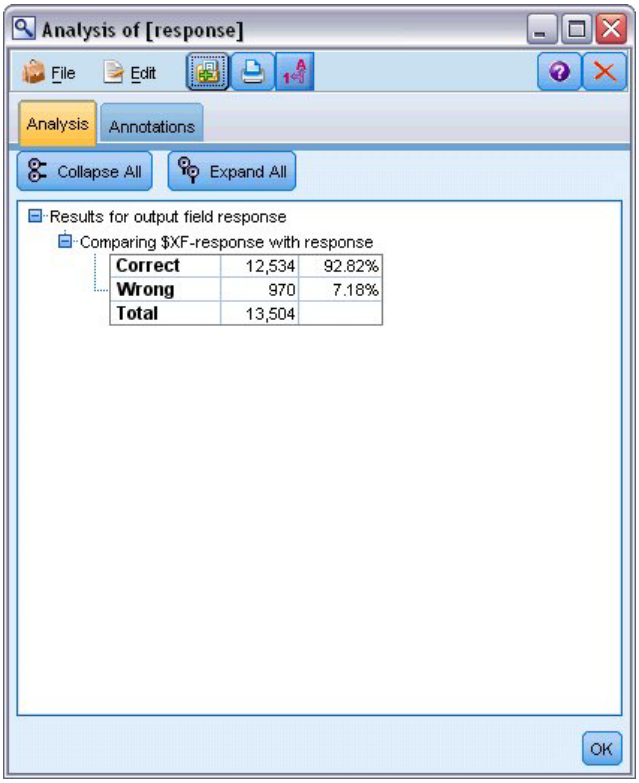


図 41. 3 つのアンサンブル・モデルの分析

要約

要約すると、自動分類ノードを使用してさまざまなモデルを比較し、3 つの最も正確なモデルを使用して、それをアンサンブル化された自動分類モデル・ナゲット内のストリームに追加しました。

- 全体の精度に基づいて、C51、C&R Tree、および CHAID モデルが学習データにおいて最もいい結果を出しました。
- このアンサンブル・モデルの結果は、個々のモデルのうちの最高のモデルとほぼ同程度であり、他のデータ・セットに適用した場合にこれを上回る結果を出す可能性があります。目標がプロセスを最大限自動化することである場合は、この方法によって、ほとんどの環境化で強固なモデルを獲得することが可能になり、任意の 1 つのモデルの詳細を深く掘り下げる必要はなくなります。

第 5 章 連続型対象の自動化モデル作成

プロパティー値 (自動数値)

自動数値ノードを使用して、資産の課税対象価格の予測など、連続型 (数値範囲) の結果のさまざまなモデルを自動的に作成して比較することができます。単一のノードで、候補モデルのセットを推定および比較し、より詳細な分析のためにモデルのサブセットを生成することができます。このノードは自動分類ノードと同じように動作しますが、フラグ型または名義型の対象ではなく、連続型に対して動作します。

ノードは、最良の候補モデルを単一の集計済み (アンサンブル) モデル・ナゲットに結合します。この方法は自動化の容易さと複数モデルの結合の利点を組み合わせるため、多くの場合、単一のモデルから取得するよりも精度が高い予測が得られます。

この例は、固定資産税を調整して評価する架空の自治体に焦点を当てています。より精度を上げるために、建物のタイプ、近隣、サイズ、その他の既知の要因に基づき、資産の価値を予測するモデルを構築します。

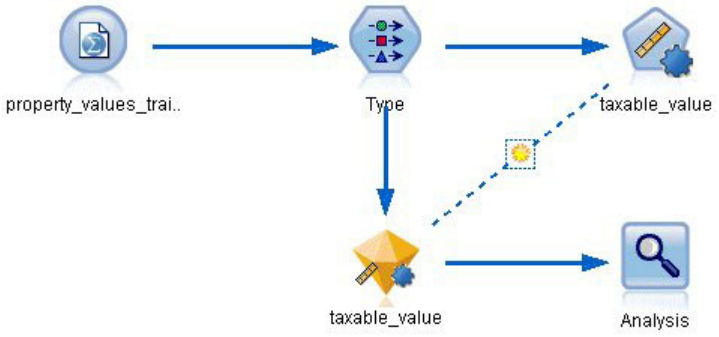


図 42. 自動数値のサンプル・ストリーム

この例では、Demos フォルダーの *streams* にインストールされているストリーム *property_values_numericpredictor.str* を使用します。使用するデータ・ファイルは *property_values_train.sav* です。詳細については、5 ページの『「Demos」フォルダー』を参照してください。

データの学習

データ・ファイルには *taxable_value* というフィールドが含まれており、これは、予測する**対象フィールド**または値です。その他のフィールドには、近隣、建物のタイプ、インテリアの数などの情報が含まれ、予測として使用される場合があります。

フィールド名	ラベル
property_id	プロパティー ID
neighborhood	市内のエリア
building_type	建築のタイプ
year_built	建築年数
volume_interior	インテリアの数
volume_other	ガレージや追加の建物の大きさ

フィールド名	ラベル
lot_size	ロット・サイズ
taxable_value	課税対象価格

property_values_score.sav というスコアリング・データ・ファイルも Demos フォルダーに含まれています。これには同じフィールドが含まれていますが、*taxable_value* フィールドはありません。課税対象価格が分かっているデータ・セットを使用してモデルを学習した後、この値がまだ分からないレコードをスコアリングすることができます。

ストリームの構築

1. IBM SPSS Modeler インストール環境の *Demos* フォルダーにある *property_values_train.sav* を指し示す Statistics ファイル入力ノードを追加します。(このフォルダーを参照するショートカットとして、ファイル・パスに `$CLEO_DEMOS/` を指定できます。表示されているとおり、パスにはバックスラッシュではなく普通のスラッシュを使用する必要があります)。

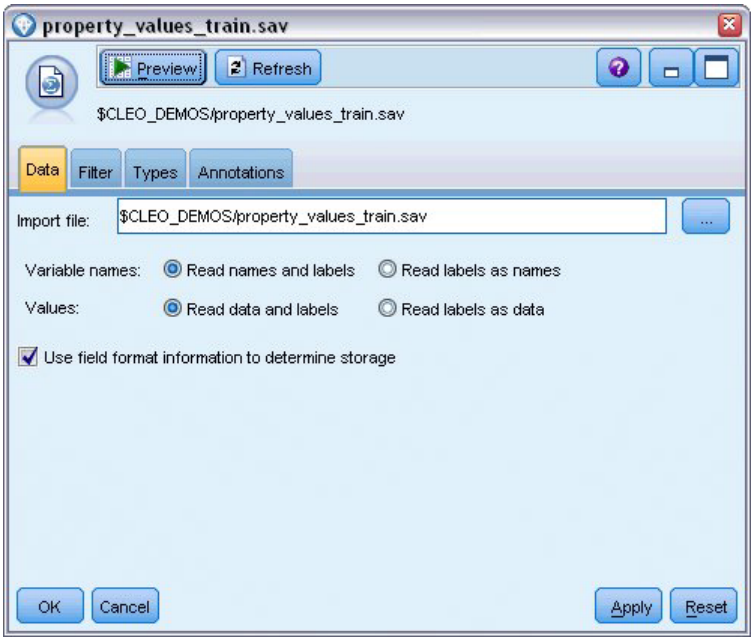


図 43. データの読み取り

2. データ型ノードを追加し、*taxable_value* を対象フィールドとして選択します (役割は「**対象**」)。他のすべてのフィールドの役割は、「**入力**」 (予測として使用されることを示す) に設定します。

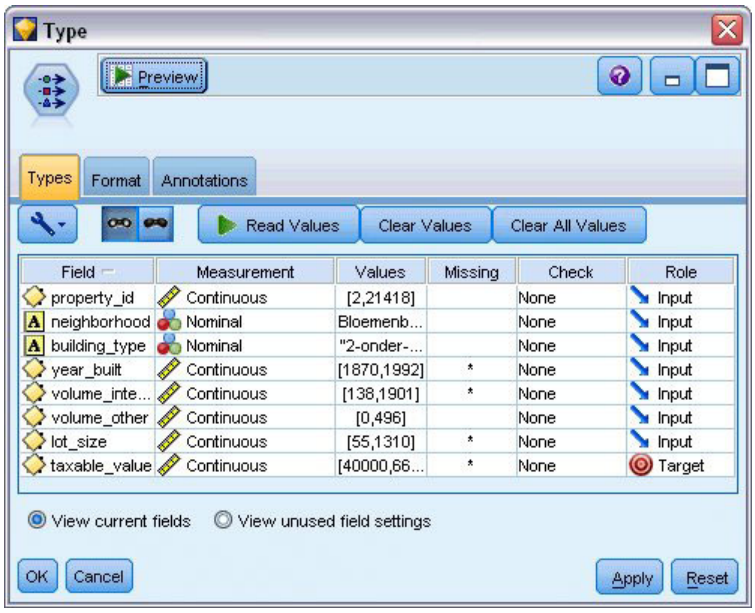


図 44. 対象フィールドの設定

3. 自動数値ノードを接続して、「**相関**」をモデルのランク付けに使用する測定基準として選択します。
4. 「**使用するモデルの数**」を 3 に設定します。 つまり、ノードを実行すると 3 つの最適モデルが作成されます。

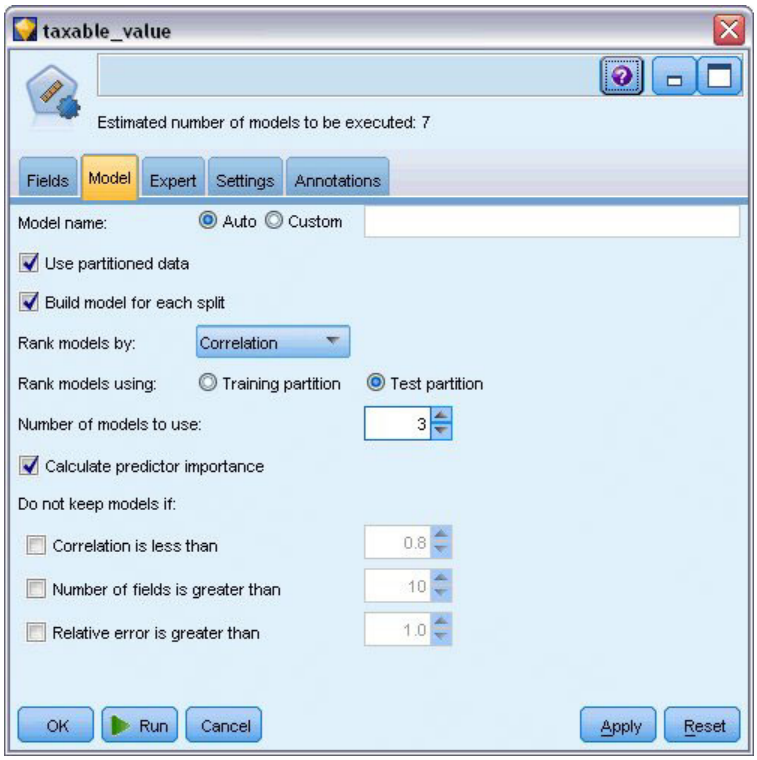


図 45. 自動数値ノードの「モデル」タブ

5. 「エキスパート」タブで、デフォルト設定をそのままにします。ノードは、7 つのモデルの合計について、各アルゴリズムの 1 つのモデルを推定します。(また、これらの設定を変更して、各モデル・タイプの複数のバリアントを比較することもできます)。

   「モデル」タブで **「使用するモデル数」**を 3 に設定しているため、ノードは 7 つのアルゴリズムの精度を計算し、3 つの最も正確なアルゴリズムを含む単一のモデル・ナゲットを構築します。

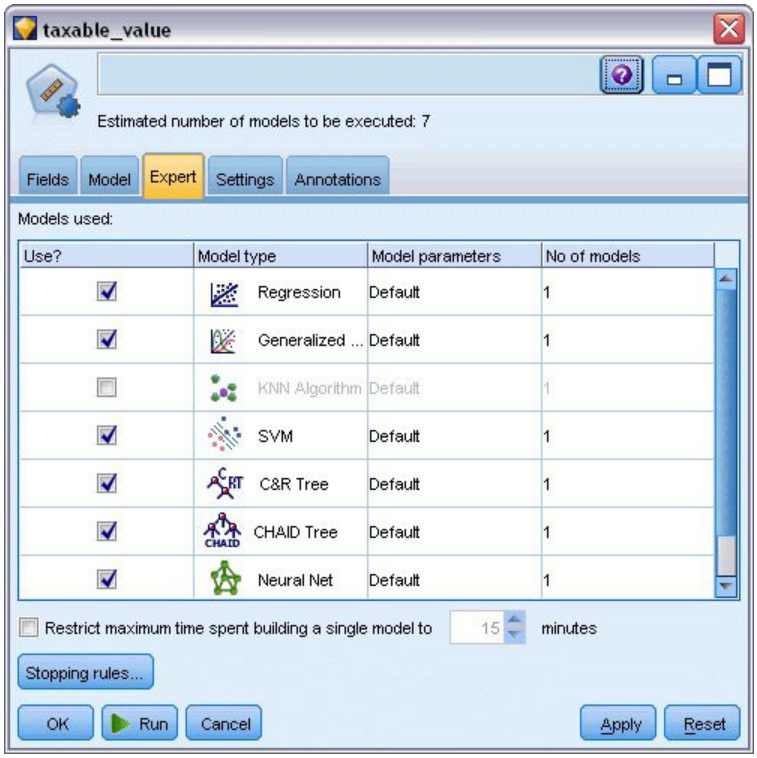


図 46. 自動数値ノードの「エキスパート」タブ

6. 「設定」タブで、デフォルト設定をそのままにします。これは連続型対象であるため、アンサンブル・スコアが各モデルのスコアを平均化することによって生成されます。

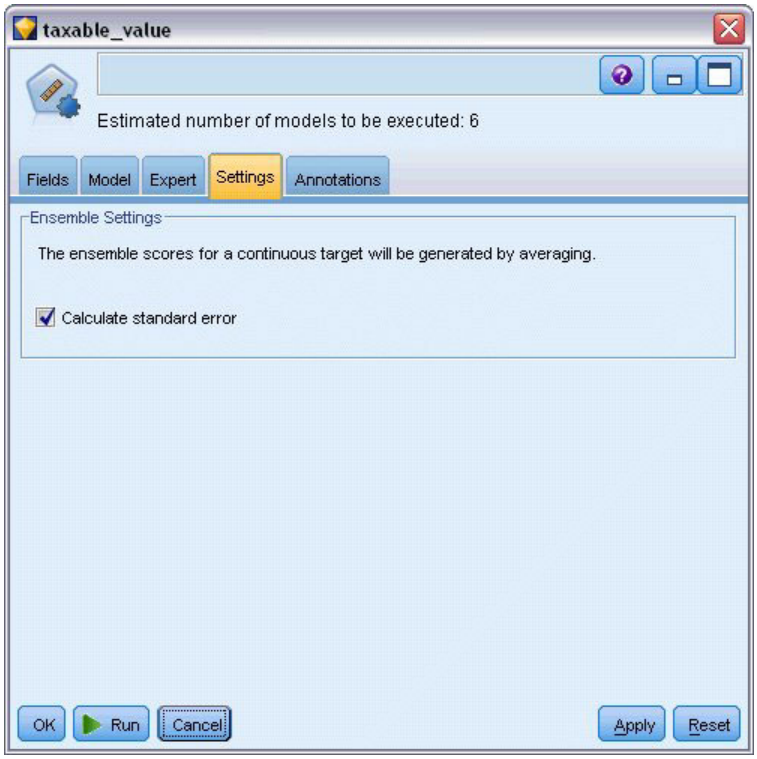


図 47. 自動数値ノードの「設定」タブ

モデルの比較

1. 「実行」ボタンをクリックします。

モデル・ナゲットが作成され、領域内、およびウィンドウの右上隅の「モデル」パレットに配置されます。ナゲットをさまざまな方法で参照、保存または展開することができます。

モデル・ナゲットを開きます。実行時に作成された各モデルの詳細がリストされます。(大規模なデータ・セットで何百ものモデルが推定される現実の状況では、これには何時間もかかることがあります)。 51 ページの図 42を参照してください。

個々のモデルをより詳細に探る場合、「**モデル**」列のモデル・ナゲットのアイコンをダブルクリックし、個別のモデルの結果をドリルダウンして参照することができます。そこから、モデル作成ノード、モデル・ナゲット、または評価グラフを生成することができます。

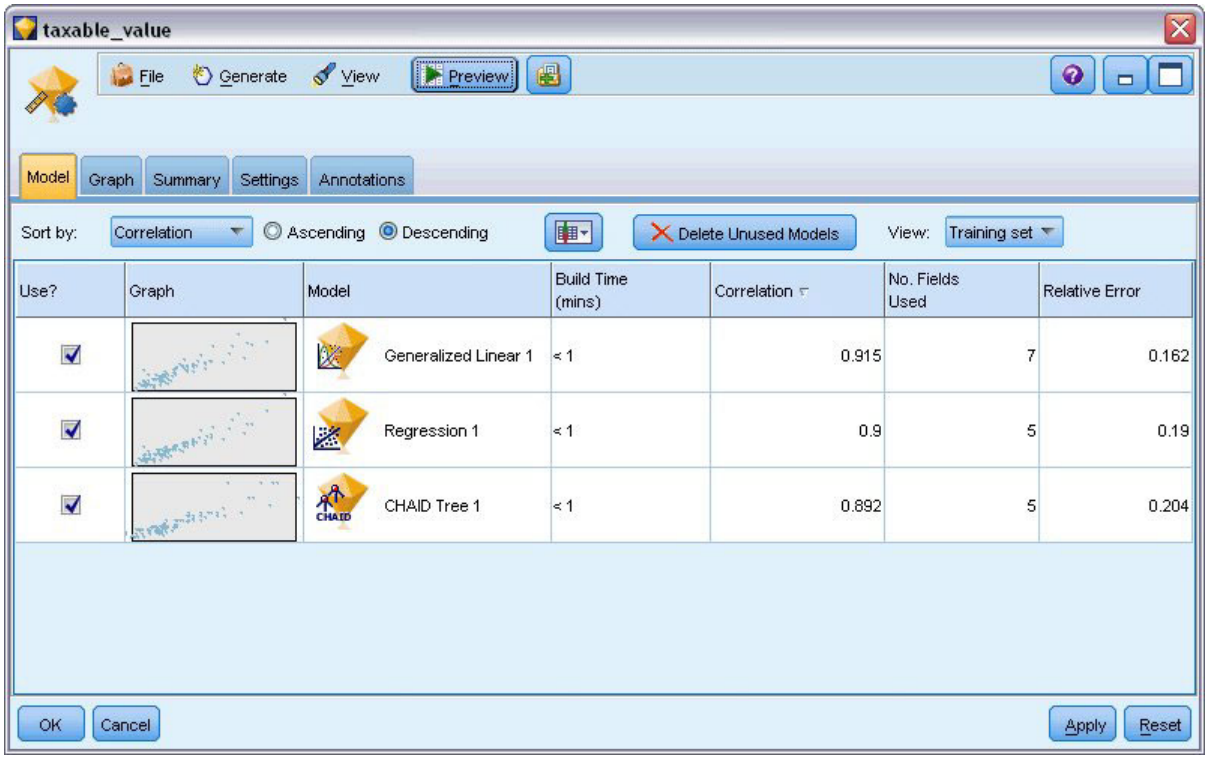


図 48. 自動数値の結果

デフォルトでは、モデルは自動数値ノードで選択した指標である、相関に基づいてソートされます。ランク付けのために、相関の絶対値が使用されます。1 に近い値ほど強力な関係を示します。一般化線型モデルはこの指標では最も優れていますが、他のいくつかのモデルも、ほぼ同じくらい正確です。一般化線型モデルにも、最低相対誤差があります。

列の見出しをクリックして異なる列でソートすることができます。あるいは、ツールバーの「**ソート基準**」リストから目的の指標を選択することもできます。

各グラフでは、モデルの予測値に対する観察値のプロットが表示され、それらの間の相関が迅速に視覚的に示されます。良好なモデルの場合、ポイントは対角線に沿って集まります。これは、この例のすべてのモデルに当てはまります。

「**グラフ**」列で、サムネー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。

これらの結果に基づいて、3 つの最も正確なモデルをすべて使用するように指定します。複数モデルの予測を組み合わせることにより、個々のモデルの制限を回避でき、全体の精度がより高くなります。

「**使用?**」列で、3 つすべてのモデルが選択されていることを確認します。

モデル・ナゲットの後に、精度分析ノード (「出力」パレット) を接続します。精度分析ノードを右クリックし、「**実行**」をクリックしてストリームを実行します。

アンサンブル・モデルが生成した平均化されたスコアは、*$XR-taxable_value* というフィールドに追加されます。相関は 0.922 で、3 つの個別モデルの相関よりも高くなっています。また、アンサンブル・スコアは低い絶対平均誤差を示し、他のデータ・セットに適用された場合、個々のモデルのどれよりもパフォーマンスが良くなる可能性があります。

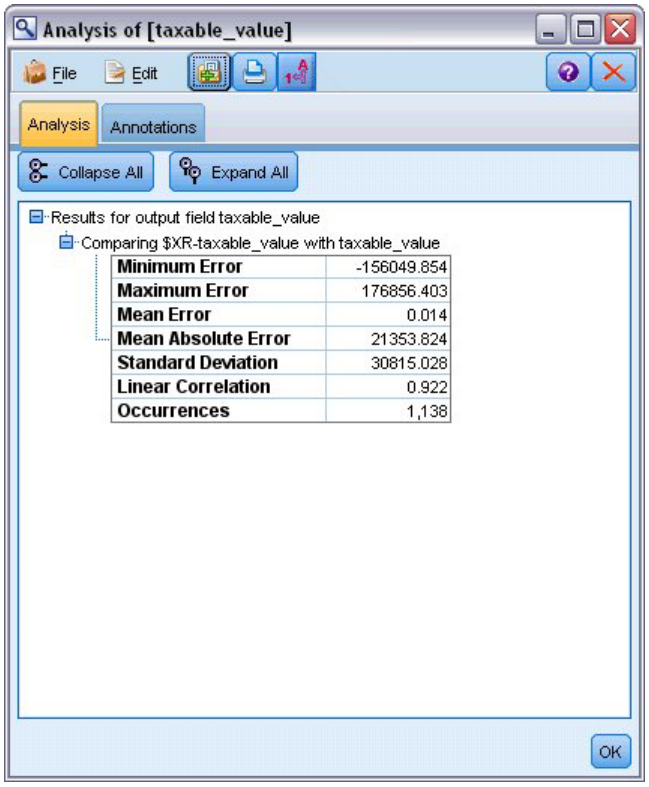


図 49. 自動数値のサンプル・ストリーム

要約

要約すると、自動数値ノードを使用してさまざまなモデルを比較し、3 つの最も正確なモデルを選択して、それをアンサンブル化された自動数値モデル・ナゲット内のストリームに追加しました。

- 全体の精度に基づいて、一般化線型、回帰、CHAID モデルは学習データにおいて最も良い結果を出しました。
- アンサンブル・モデルのパフォーマンスは各モデルの他の 2 つの個別モデルより良好であり、他のデータ・セットに適用した場合にさらに良いパフォーマンスを示す可能性があります。目標がプロセスを最大限自動化することである場合は、この方法によって、ほとんどの環境化で強固なモデルを獲得することが可能になり、任意の 1 つのモデルの詳細を深く掘り下げる必要はなくなります。

第 6 章 自動データ準備 (ADP)

分析に向けてデータを準備することは、データ・マイニング・プロジェクトにおいて最も重要な手順の 1 つですが、従来は最も時間のかかる手順の 1 つでもありました。自動データ準備 (ADP) ノードでは、データ分析、固定値の識別、問題のあるまたは役に立たない可能性の高いフィールドのスクリーニング、必要に応じた新しい属性の派生、高度なスクリーニング手法を使用したパフォーマンスの向上などのタスクを自動的に処理します。完全に自動化された方法でノードを使用し、ノードで固定値を選択および適用できます。また、必要に応じて、変更の作成前に変更をプレビューし、その変更を承認または拒否できます。

ADP ノードを使用すると、関係する統計コンセプトを事前に理解しなくとも、迅速かつ容易にデータ・マイニング用にデータを準備できます。デフォルトの設定でノードを実行した場合、モデルの作成およびスコアリングを素早く行うことができます。

この例では、*ADP_basic_demo.str* というストリームを使用します。これは *telco.sav* というデータ・ファイルを参照して、モデル作成時にデフォルトの ADP ノード設定を使用して実現できる精度の向上を実演します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*ADP_basic_demo.str* ファイルは、*streams* ディレクトリー内にあります。

ストリームの構築

1. ストリームを構築するには、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにある *telco.sav* を指し示す Statistics ファイル入力ノードを追加します。

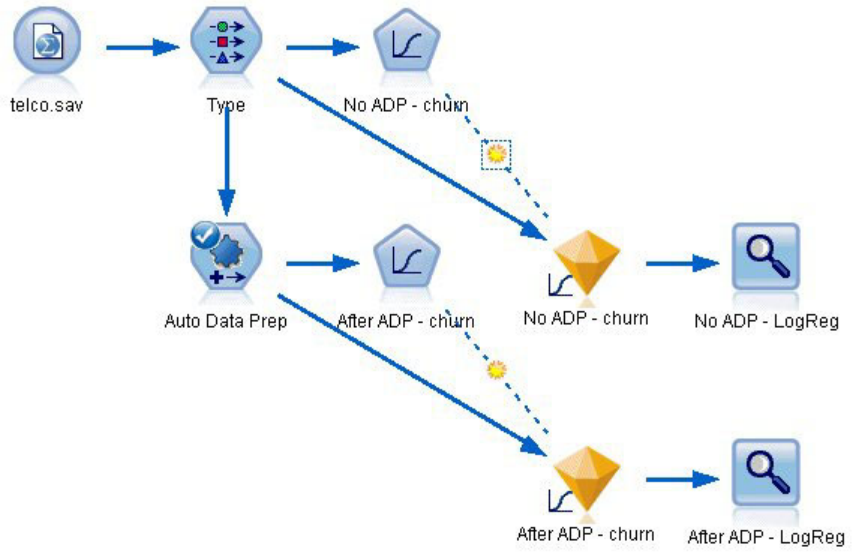


図 50. ストリームの構築

2. データ型ノードを入力ノードに接続し、*churn* フィールドの測定の尺度を「**フラグ**」に設定し、役割を「**対象**」に設定します。その他のフィールドの役割は、すべて「**入力**」に設定します。

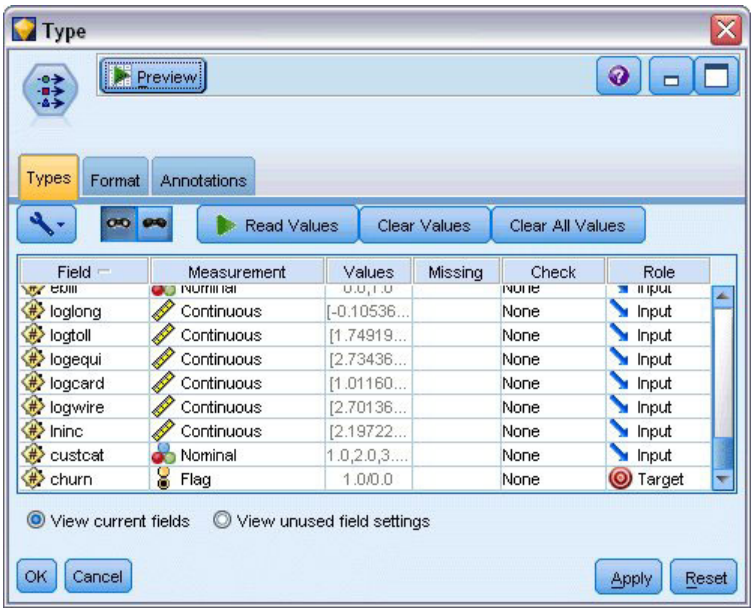


図 51. 対象の選択

3. ロジスティック・ノードをデータ型ノードに接続します。
4. ロジスティック・ノードで「モデル」タブをクリックし、「**2 項検定**」手続きを選択します。「モデル名」フィールドで、「**カスタム**」を選択し、「ADP なし - 解約」と入力します。

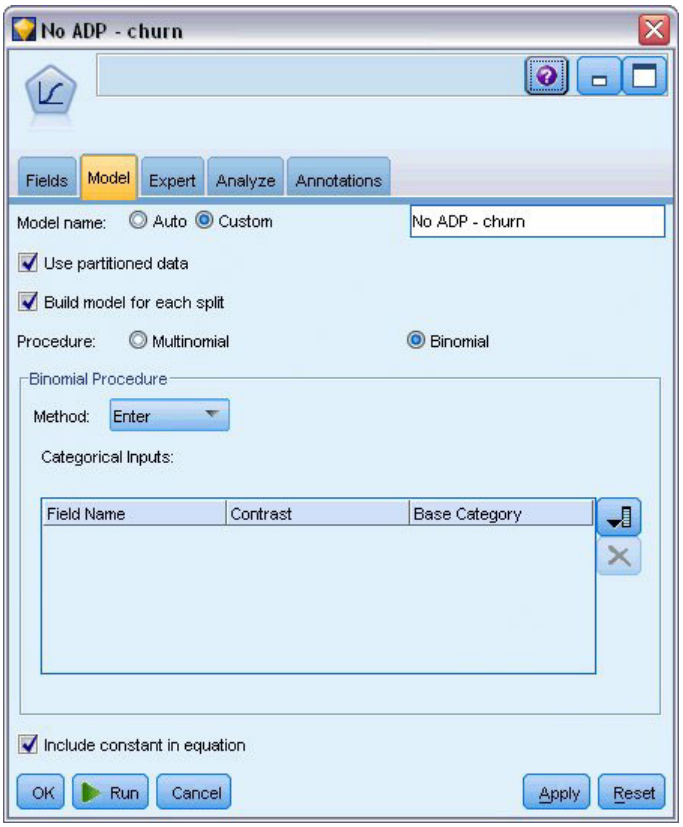


図 52. モデル・オプションの選択

5. ADP ノードをデータ型ノードに接続します。「目的」タブはデフォルトの設定のままにし、速度と精度の両方のバランスを調整して、データを分析および準備します。
6. 「目的」タブの上部で、「**データの分析**」をクリックして、データを分析および処理します。

ADP ノードの他のオプションを使用すると、精度により重きを置くか、処理の速度により重きを置くかを指定したり、多くのデータ準備プロセスの手順を微調整したりすることができます。

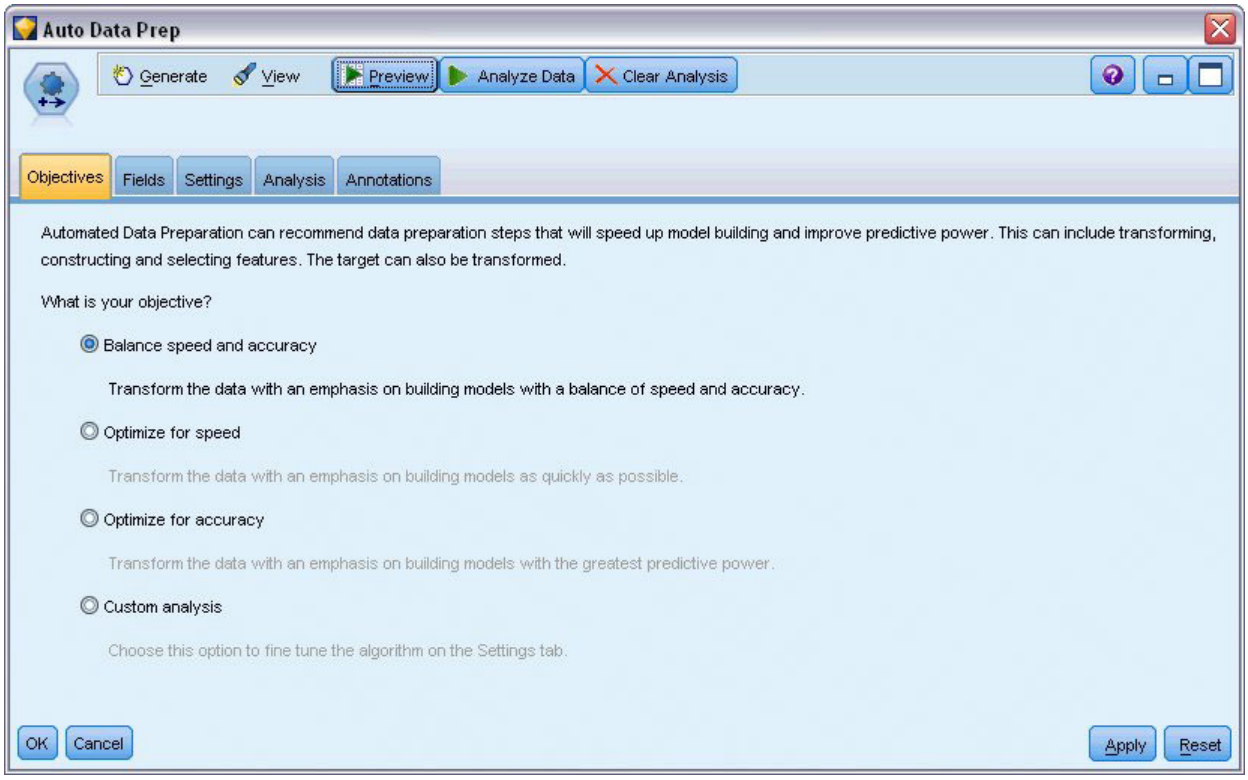


図 53. ADP のデフォルトの目的

データ処理の結果が「分析」タブに表示されます。「**フィールド処理の要約**」には、ADP ノードに取り込まれた 41 件のデータ・フィーチャーのうち、19 件が変換されて処理に使用され、3 件が破棄されて使用されていないことが示されています。

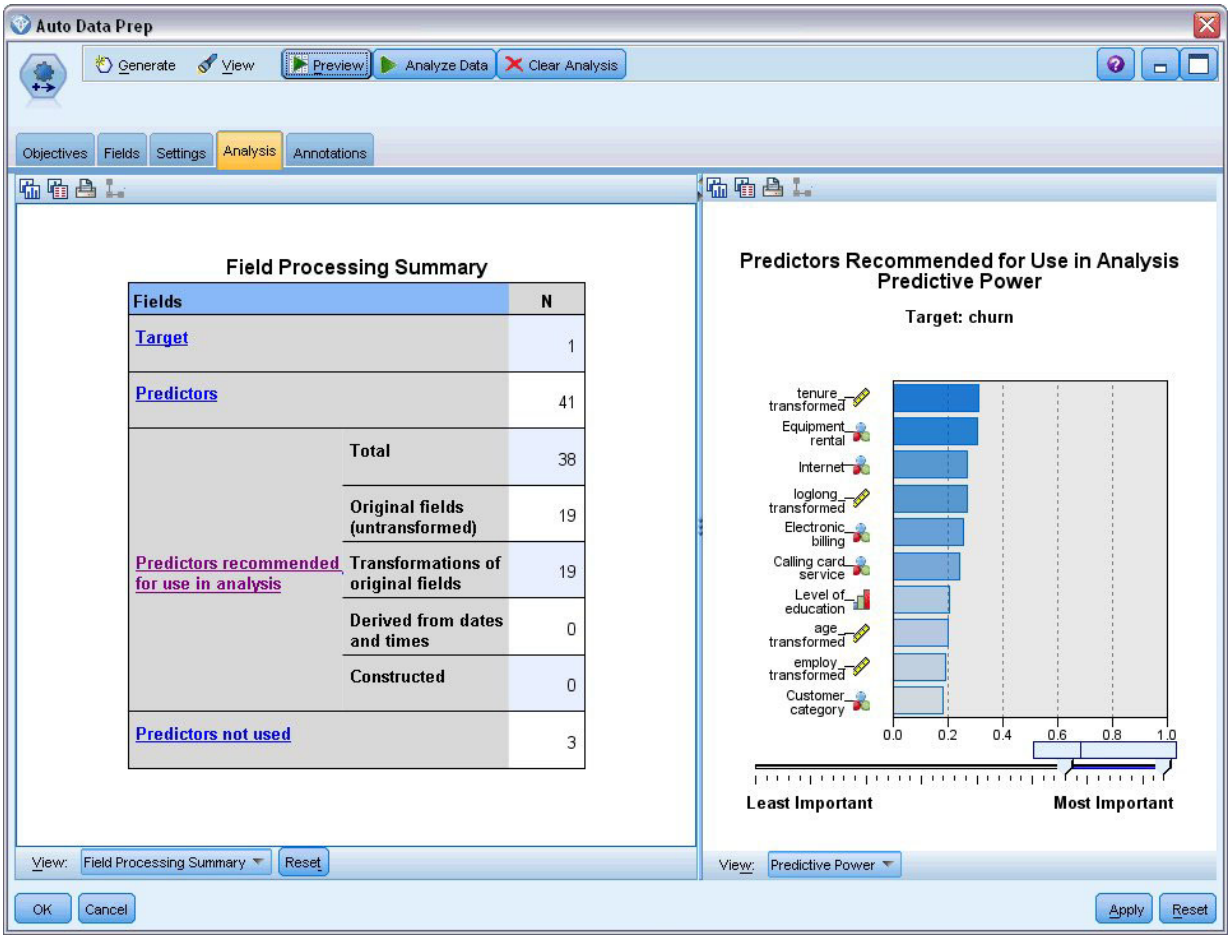


図 54. データ処理の要約

7. ロジスティック・ノードを ADP ノードに接続します。
8. ロジスティック・ノードで「モデル」タブをクリックし、「**2 項検定**」手続きを選択します。「モデル名」フィールドで、「**カスタム**」を選択し、「ADP 後 - 解約」と入力します。

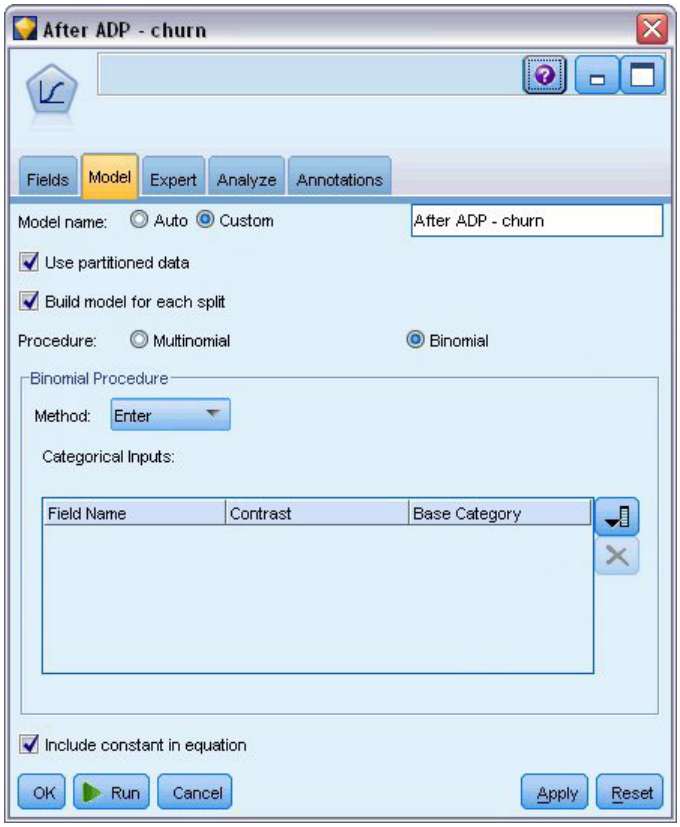


図 55. モデル・オプションの選択

モデルの精度の比較

1. 両方のロジスティック・ノードを実行して、モデル・ナゲットを作成します。モデル・ナゲットは、ストリームおよび右上隅のモデル・パレットに追加されます。

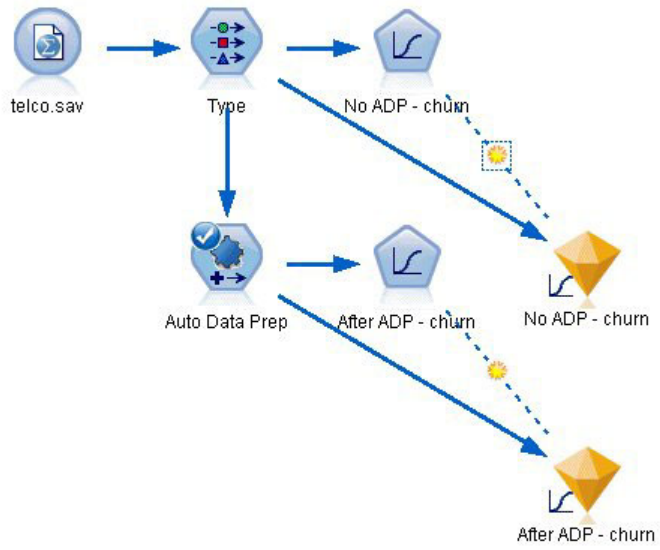


図 56. モデル・ナゲットの接続

2. 精度分析ノードをモデル・ナゲットに接続し、デフォルトの設定を使用して精度分析ノードを実行します。

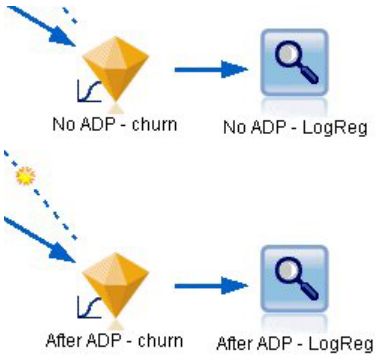


図 57. 精度分析ノードの接続

非 ADP 派生モデルの分析では、ロジスティック回帰ノードを使用してデフォルト設定でデータを実行すると、モデルの精度が低くなる (わずか 10.6%) ことが分かります。

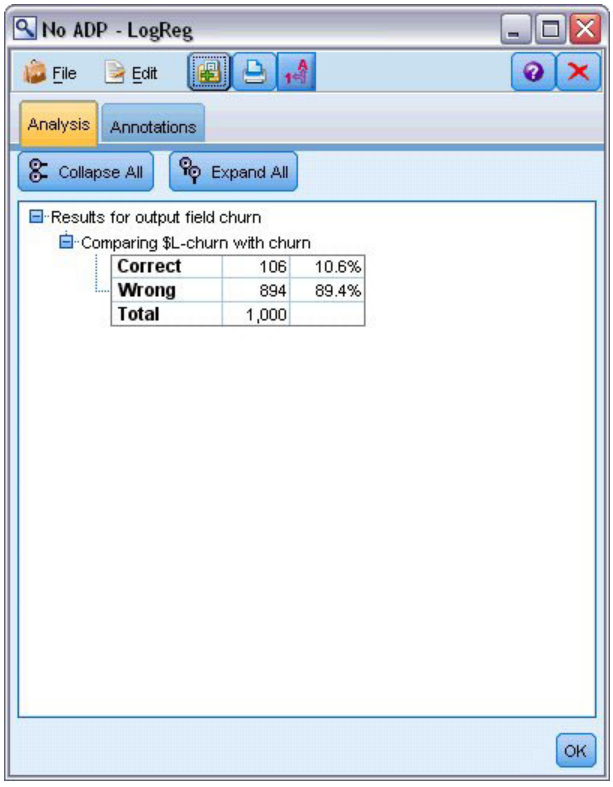


図 58. 非 ADP 派生モデルの結果

ADP 派生モデルの分析では、デフォルトの ADP 設定でデータを実行すると、はるかに高い精度の (78.8% が正しい) モデルが作成されることが分かります。

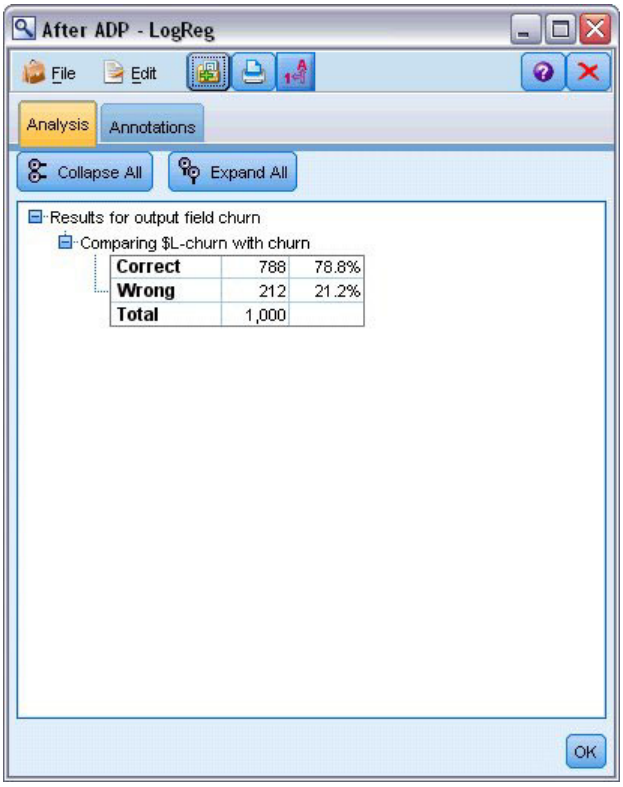


図 59. ADP 派生モデルの結果

要約すると、ADP ノードを実行してデータの処理を微調整するだけで、ほとんど直接データを操作することなく、より正確なモデルを作成することができました。

明らかに特定の理論の証明または反証に関心がある場合、または特定のモデルを作成する場合は、モデル設定を直接操作した方がよい場合もあります。ただし、時間がない場合、または準備するデータ量が多い場合は、ADP ノードを利用すると便利なことがあります。

IBM SPSS Modeler で使用されるモデル作成方法の数学的な基礎の説明については、インストール・ディスクの *¥Documentation* ディレクトリーにある「*IBM SPSS Modeler アルゴリズム・ガイド*」を参照してください。

この例の結果は学習データのみに基づくことに注意してください。モデルが実際の世界の他のデータにどれだけうまく一般化されるかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。

第 7 章 分析用のデータの準備 (データ監査)

データ監査ノードは、IBM SPSS Modeler に取り込むデータをまず広範囲に検査するための手段を提供します。データ監査レポートは、初期データ探索時に頻繁に使用され、各データ・フィールドのヒストグラムや分布図に加えて、要約統計量を表示し、また欠損値、外れ値、および極値の処理を指定できます。

この例では、*telco.sav* というデータ・ファイルを参照する *telco_dataaudit.str* というストリームを使用します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*telco_dataaudit.str* ファイルは、*streams* ディレクトリー内にあります。

ストリームの構築

1. ストリームを構築するには、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにある *telco.sav* を指し示す Statistics ファイル入力ノードを追加します。

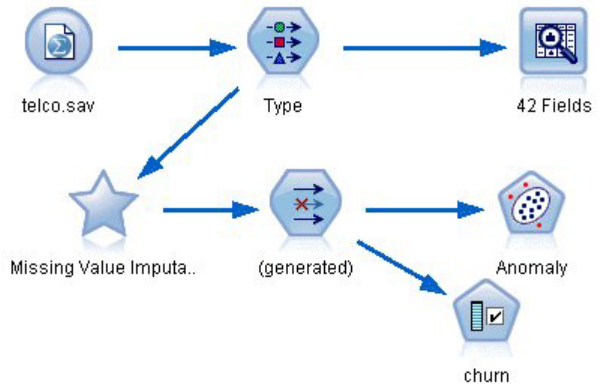


図 60. ストリームの構築

2. データ型ノードを追加してフィールドを定義し、*churn* を対象フィールドとして指定します (役割は「**対象**」)。これが唯一の対象フィールドになるように、他のすべてのフィールドの役割を「**入力**」に設定します。

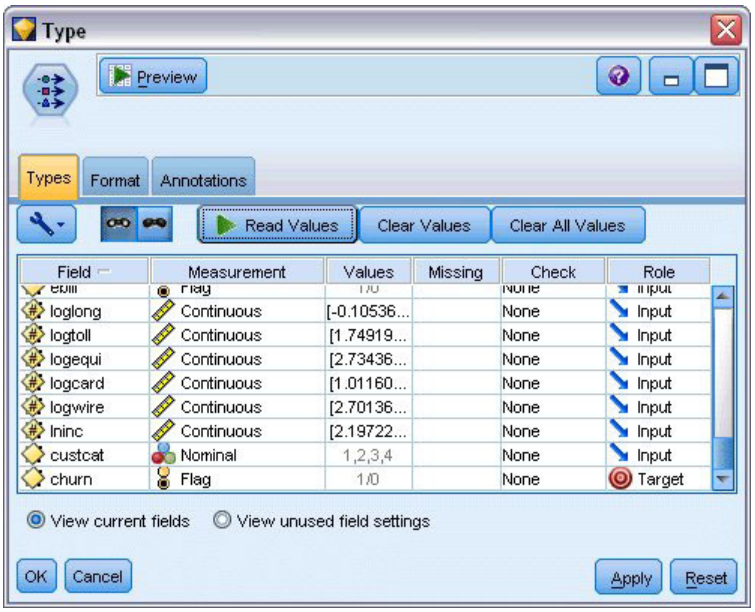


図 61. 対象の設定

3. フィールドの測定の尺度が適切に定義されていることを確認します。例えば、値 0 および 1 を持つほとんどのフィールドはフラグ型と見なすことができますが、性別などの特定のフィールドは、 2 つの値を持つ名義型フィールドの方がより正確に認識されます。

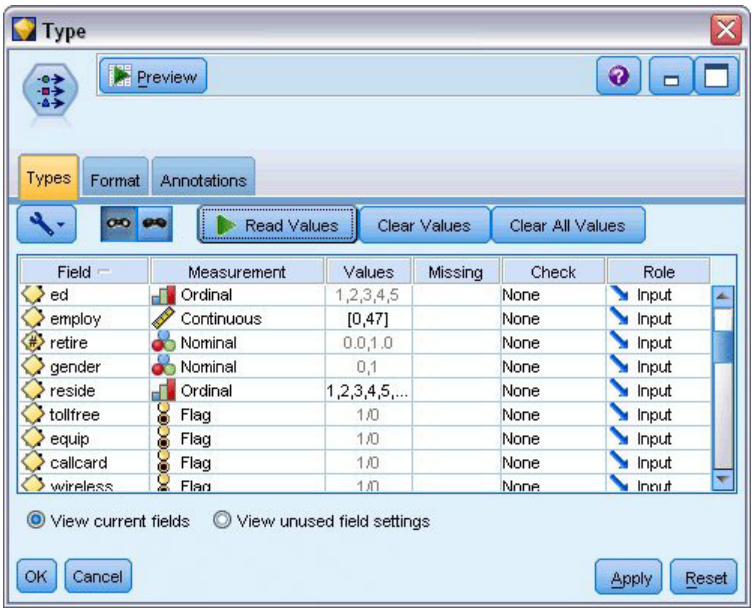


図 62. 測定の尺度の設定

ヒント: 類似した値 (0/1 など) を持つ複数のフィールドに対してプロパティーを変更するには、「*値*」列のヘッダーをクリックしてその列でフィールドをソートし、Shift キーを使用して、変更するフィールドをすべて選択します。その後、選択範囲を右クリックして、選択したすべてのフィールドの測定の尺度または他の属性を変更することができます。

4. データ監査ノードをストリームに接続します。「設定」タブでは、デフォルト設定はそのままにして、すべてのフィールドをレポートに含めます。*churn* はデータ型ノードで定義されている唯一の対象フィールドなので、オーバーレイとして自動的に使用されます。

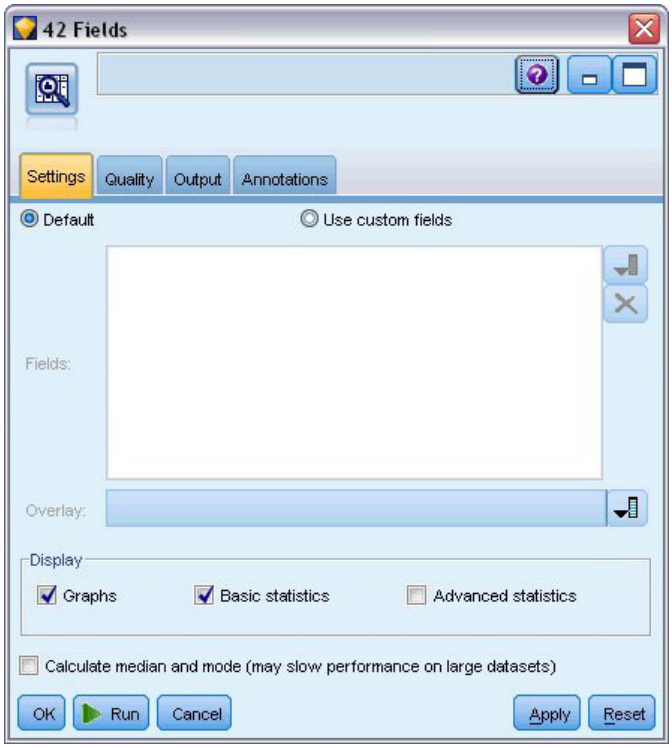


図 63. データ監査ノードの「設定」タブ

「欠損値検査」タブで、欠損値、外れ値、および極値の検出のデフォルト設定はそのままにしておき、「**実行**」をクリックします。

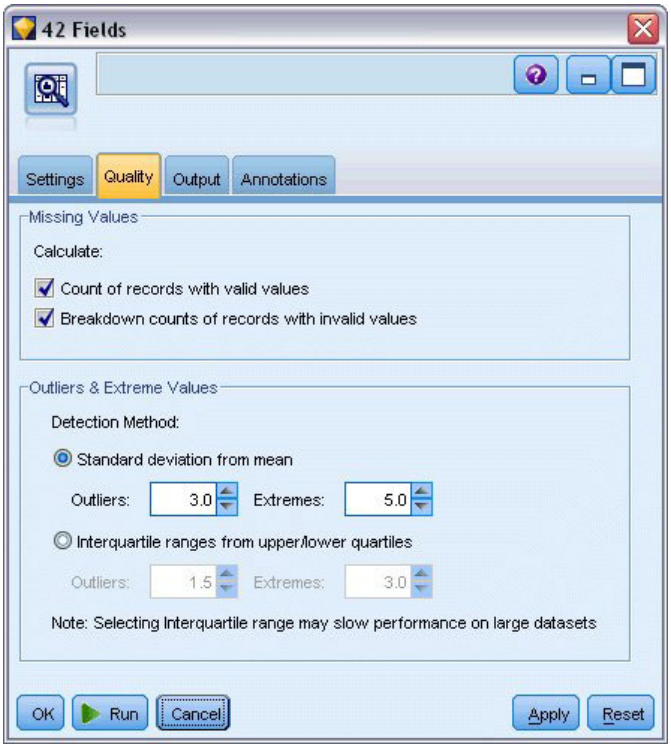


図 64. データ監査ノードの「欠損値検査」タブ

統計とグラフの参照

「データ監査」ブラウザーが表示され、フィールドごとにサムネール・グラフや記述統計が表示されます。

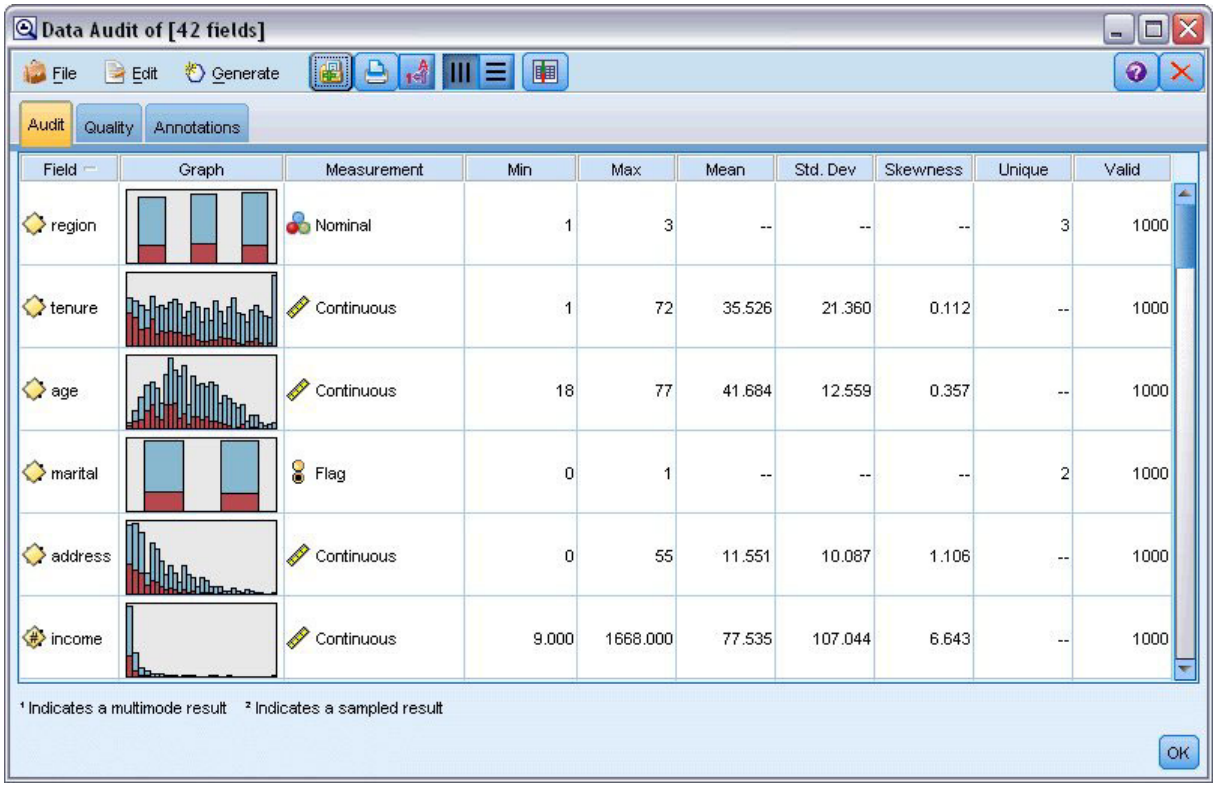


図 65. データ監査ブラウザー

ツールバーを使用してフィールドと値のラベルを表示し、グラフの位置合わせを水平から垂直に切り替えます (カテゴリー別フィールドの場合のみ)。

1. ツールバーまたは「編集」メニューを使用して、表示する統計を選択することもできます。

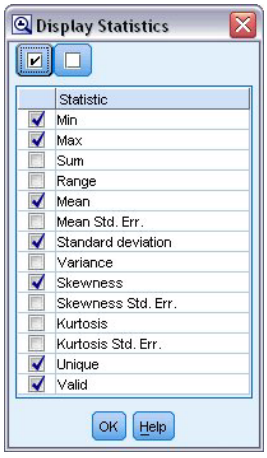


図 66. 統計の表示

監査レポートのサムネール・グラフをダブルクリックして、フルサイズ・バージョンのグラフを表示します。*churn* はストリーム内の唯一の対象フィールドであるため、オーバーレイとして自動的に使用されます。グラフをさらにカスタマイズするには、グラフ・ウィンドウのツールバーを使用してフィールドや値のラベルの表示を切り替えるか、あるいは「編集モード」ボタンをクリックします。

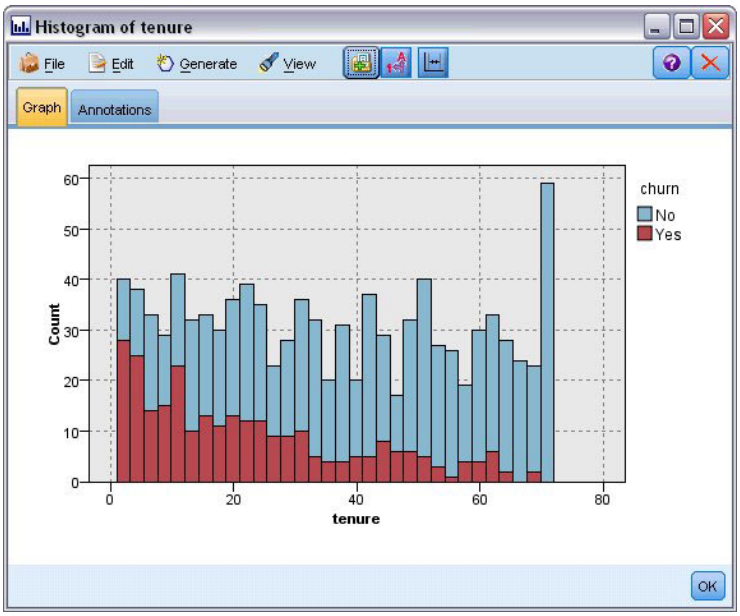


図 67. 保有期間のヒストグラム

あるいは、1 つ以上のサムネールを選択してそれぞれのグラフ・ノードを生成することもできます。生成されたノードはストリーム領域に配置され、ストリームに追加して特定のグラフを再作成できます。

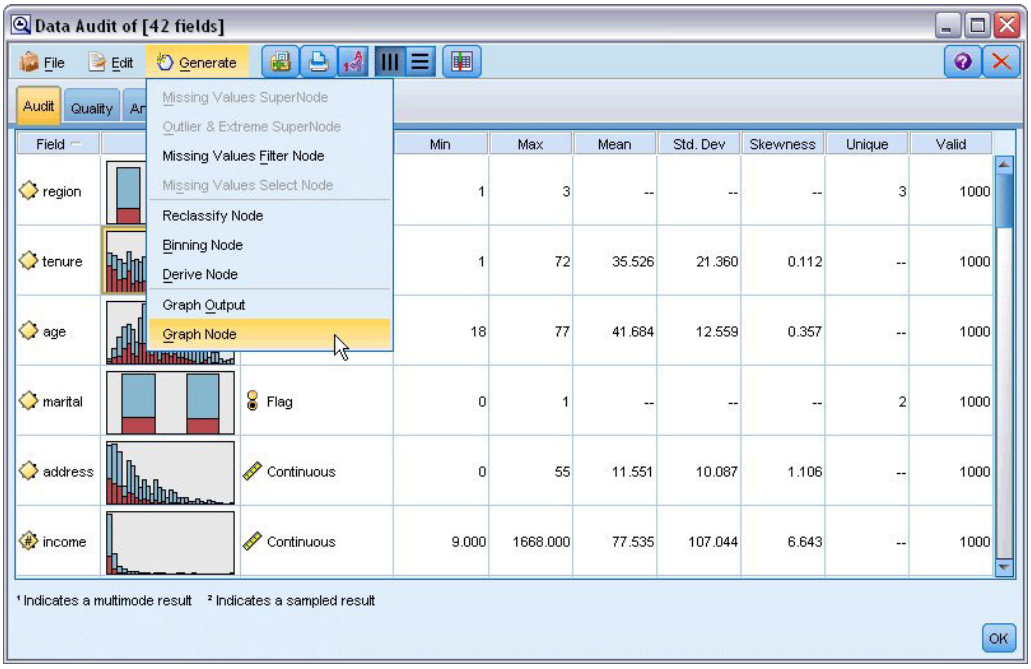


図 68. グラフ・ノードの生成

外れ値および欠損値の処理

監査レポートの「欠損値検査」タブは、外れ値、極値、および欠損値に関する情報を示します。

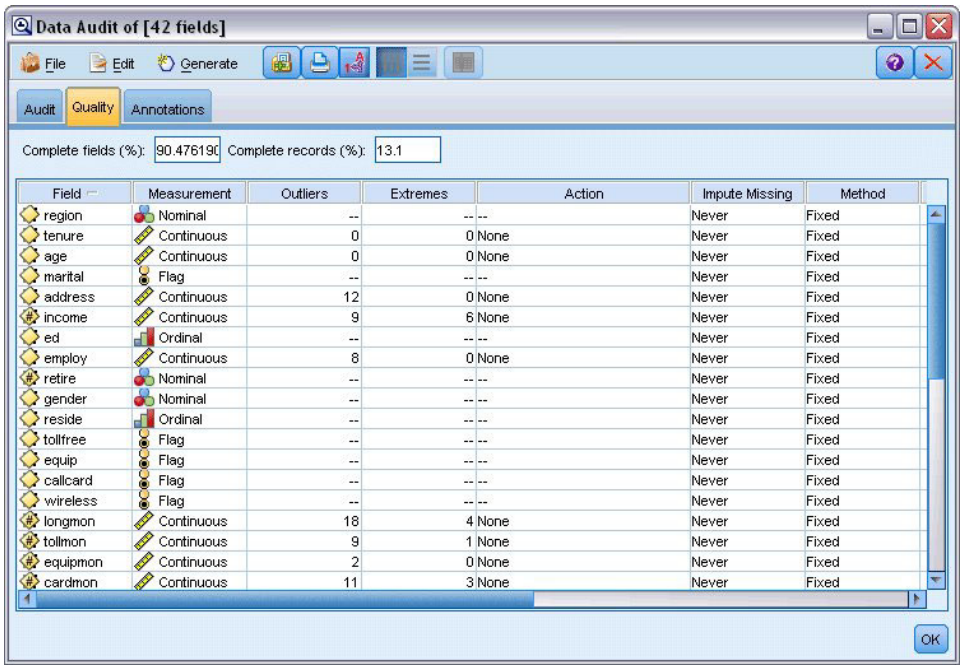


図 69. データ監査ブラウザーの「欠損値検査」タブ

これらの値の処理方法を指定し、スーパーノードを生成して、変換を自動的に適用することもできます。例えば、1 つ以上のフィールドを選択したり、C&RT アルゴリズムなどのさまざまな方法を使用してこれらのフィールドの欠損値に代入または置き換えたりすることができます。

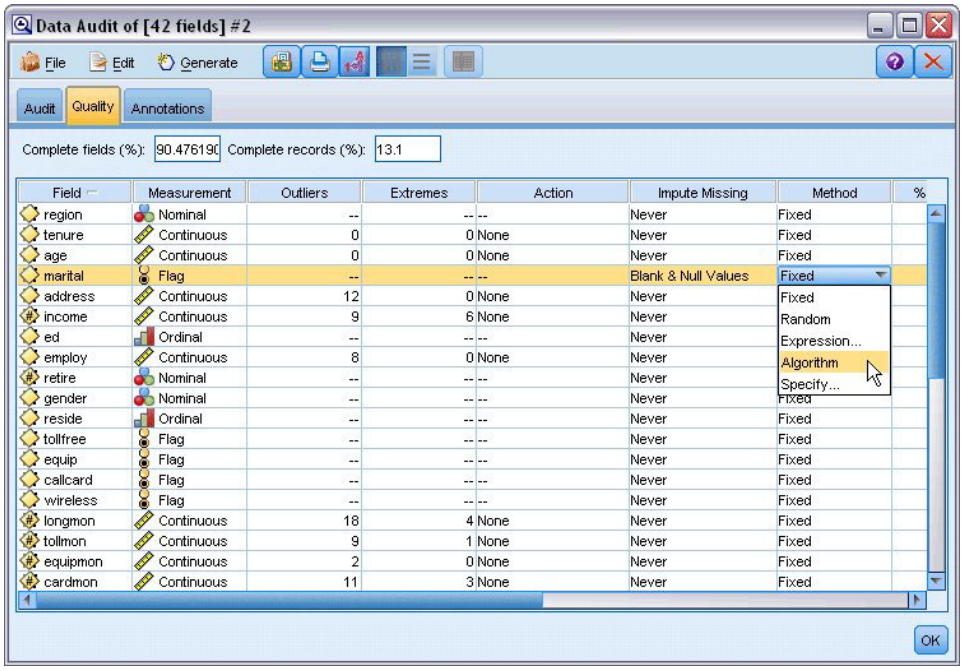


図 70. 代入法の選択

1 つ以上のフィールドに代入法を指定した後、欠損値スーパーノードを生成するには、メニューから以下のように選択します。

「生成」 > 「欠損値スーパーノード」

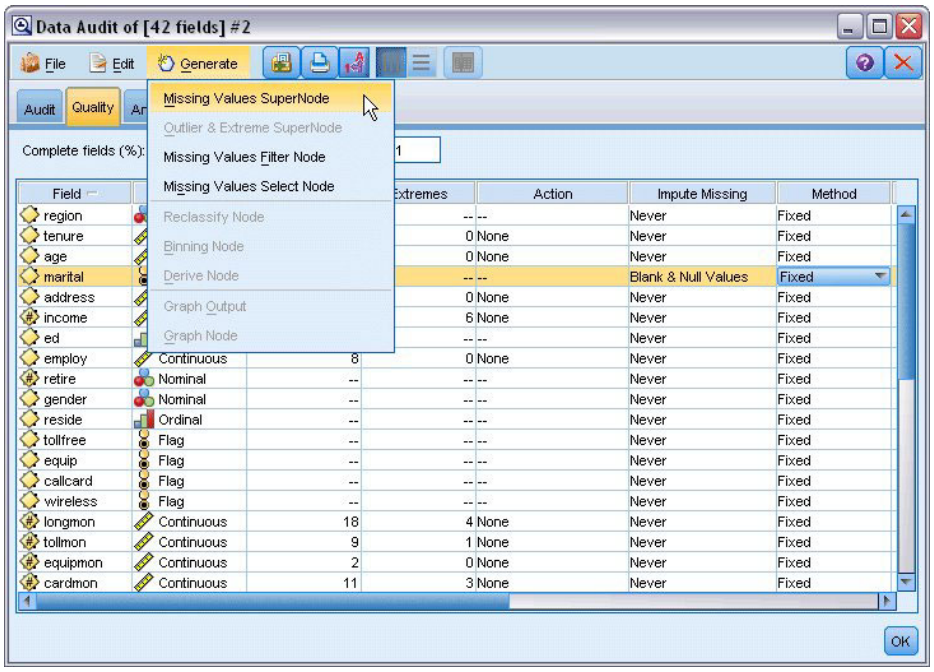


図 71. スーパーノードの生成

生成されたスーパーノードはストリーム領域に追加され、そこでそのスーパーノードをストリームに接続して変換を適用できます。

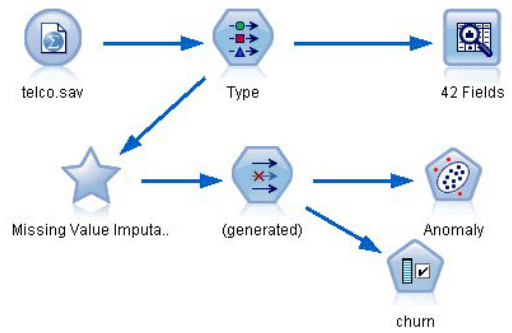


図 72. 欠損値スーパーノードによるストリーム

スーパーノードには、実際、要求された変換を実行する一連のノードが含まれています。これがどのように機能するかを理解するには、スーパーノードを編集し、「**ズームイン**」をクリックします。

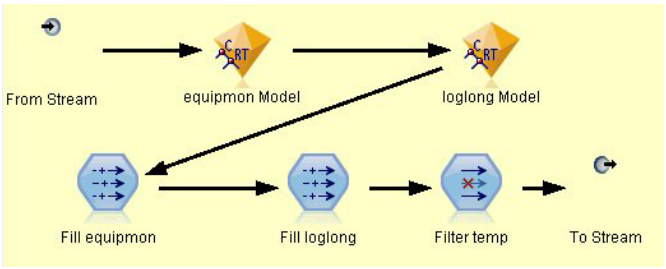


図 73. スーパーノードでのズームイン

例えば、このアルゴリズム法を使用して代入された各フィールドに対し、空白値や Null 値をモデルで予測された値と置き換える置換ノードとともに、個別の C&RT モデルがあります。スーパーノード内の特定のノードを追加、編集、あるいは削除して、動作をさらにカスタマイズできます。

あるいは、条件抽出ノードまたはフィルター・ノードを生成して、欠損値を含むフィールドまたはレコードを削除できます。例えば、指定したしきい値未満の品質パーセンテージのすべてのフィールドをフィルタリングできます。

図 74. フィルター・ノードの生成

外れ値と極値は同じ方法で処理できます。フィールドごとに実行する動作 (強制、破棄、無効化のいずれか) を指定し、スーパーノードを生成して変換を適用します。

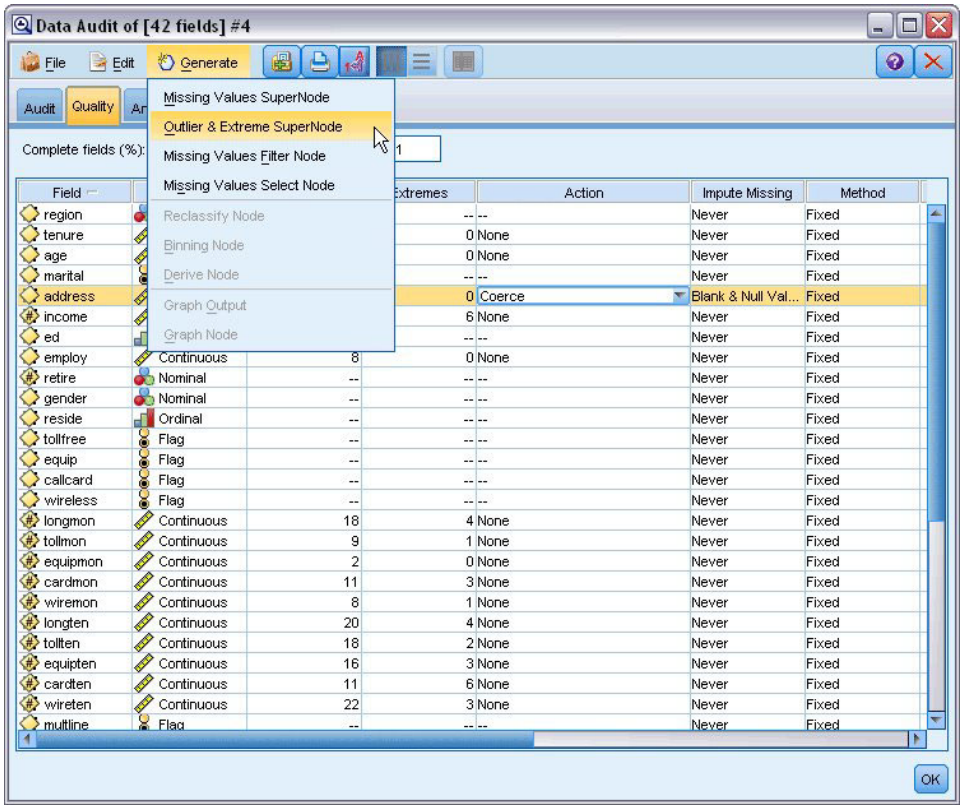


図 75. フィルター・ノードの生成

監査を完了し、生成されたノードをストリームに追加したら、分析を開始できます。必要に応じて、異常値検出、フィールド選択、またはその他のさまざまな方法を使用してデータをさらに選別できます。

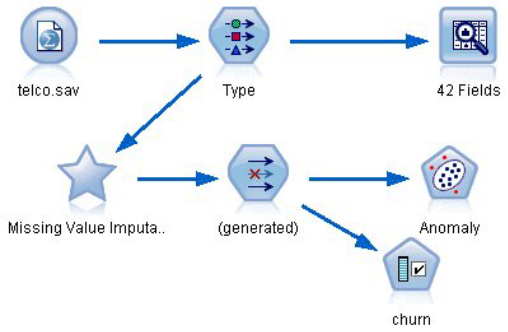


図 76. 欠損値スーパーノードによるストリーム

第 8 章 薬品による治療 (調査用グラフ/C5.0)

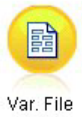
ここでは、医療研究者が研究用データを整理する場合を考えてみましょう。全員が同じ疾患を持つ患者のグループについてデータを収集しました。治療過程において、それぞれの患者は 5 種類の薬品のうちのいずれかで効果がありました。そこで、今後同じ疾患を持つ患者にどの薬品が効果的かを、データ・マイニングを使用して特定していきます。

この例では、*druglearn.str* という名前のストリームを使用します。このストリームでは *DRUG1n* という名前のデータ・ファイルを参照します。これらのファイルは、 IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*druglearn.str* ファイルは*streams* ディレクトリーにあります。

デモで使用するデータ・フィールドは、次のとおりです。

データ・フィールド	説明
年齢	(数値)
性別	*M* または *F*
血圧	血圧: *HIGH*、*NORMAL*、または *LOW*
コレステロール値	血中コレステロール: *NORMAL* または *HIGH*
ナトリウム値	血液中のナトリウム濃度
カリウム値	血液中のカリウム濃度
薬品	患者に効果があった処方薬

テキスト・データの読み取り



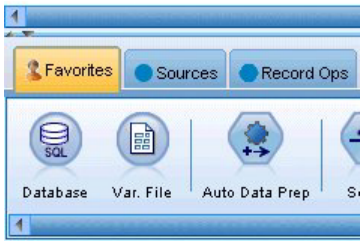


図 77. 可変長ファイル・ノードの追加

可変長ファイル・ノードを使用して区切られたテキスト・データを読み取ることができます。可変長ファイル・ノードはパレットから追加できます。「**ソース**」タブをクリックしてこのノードを探すか、このノードがデフォルトで含まれている「**お気に入り**」タブを使用します。次に、新しく配置したノードをダブルクリックして、そのダイアログ・ボックスを開きます。

「ファイル」ボックスの右側にある、省略符号 (...) の付いたボタンをクリックして、IBM SPSS Modeler がインストールされているシステム上のディレクトリーを参照します。「*Demos*」ディレクトリーを開き、*DRUG1n* というファイルを選択します。

「**ファイルからフィールド名を取得**」が選択されていることを確認し、ダイアログ・ボックスに読み込まれたフィールドおよび値を確認します。

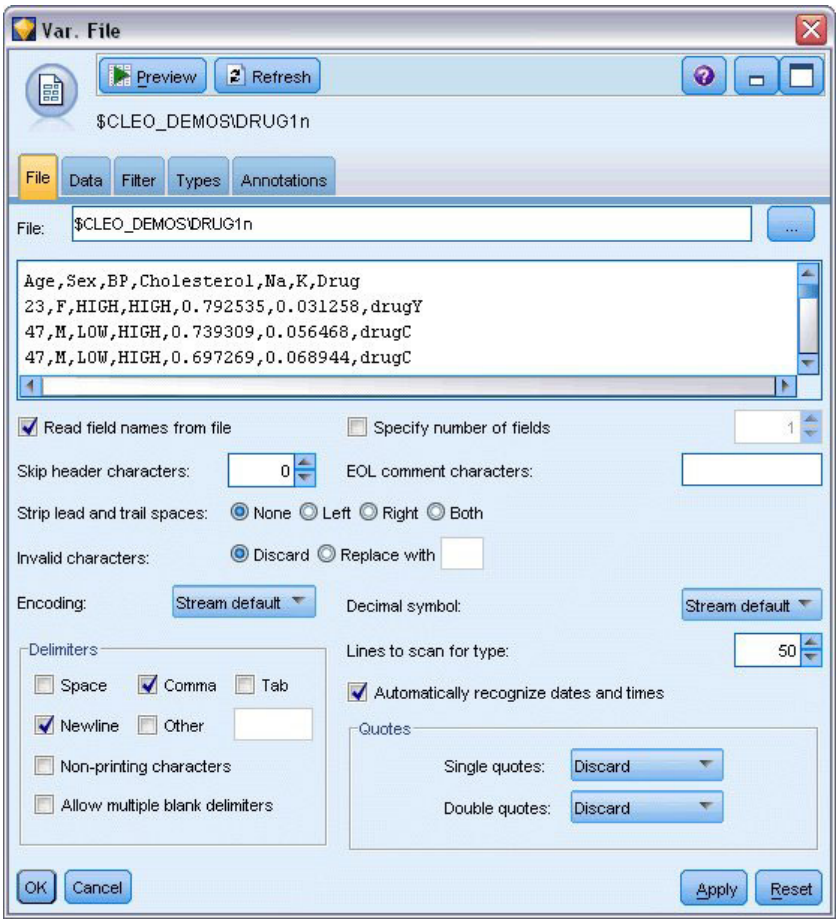


図 78. 「可変長ファイル」ダイアログ・ボックス

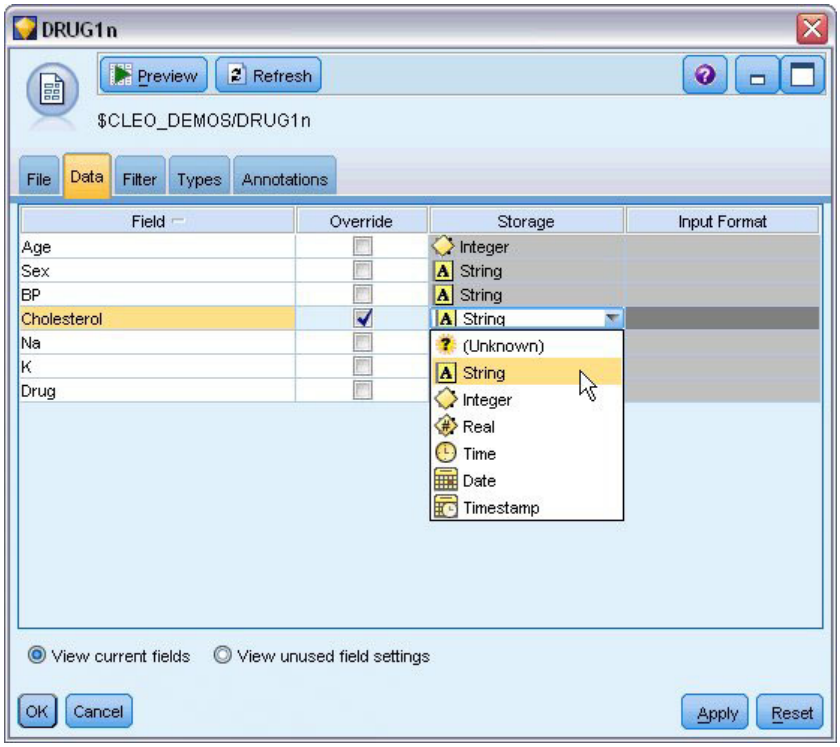


図 79. フィールドのストレージ・タイプの変更

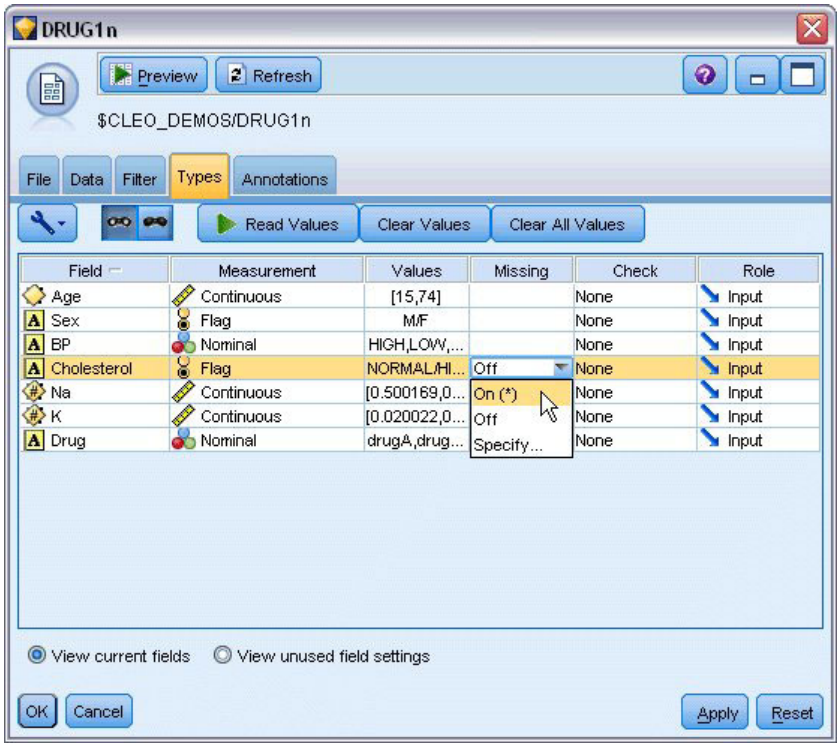


図 80. 「タイプ」タブでの「値」オプションの選択

「**データ**」タブをクリックして、フィールドの「**ストレージ**」を上書きして変更します。ストレージは「**尺度**」、つまりデータ・フィールドの測定の尺度 (または使用タイプ) とは異なります。「**タイプ**」タブは、

データ内のフィールドのタイプを確認する場合に役立ちます。また、「**値の読み込み**」を選択して、「*値*」列での選択に基づく各フィールドの実際の値を表示することもできます。このプロセスは、**インスタンス化**と呼ばれます。

テーブルの追加

データ・ファイルの読み込みが完了したので、いくつかのレコードの値を見てみましょう。レコードの値を表示する方法の 1 つに、テーブル・ノードを含むストリームの構築があります。テーブル・ノードをストリームに配置するには、パレットでアイコンをダブルクリックするか、アイコンを領域にドラッグ・アンド・ドロップします。

図 81. データ・ソースに接続されたテーブル・ノード

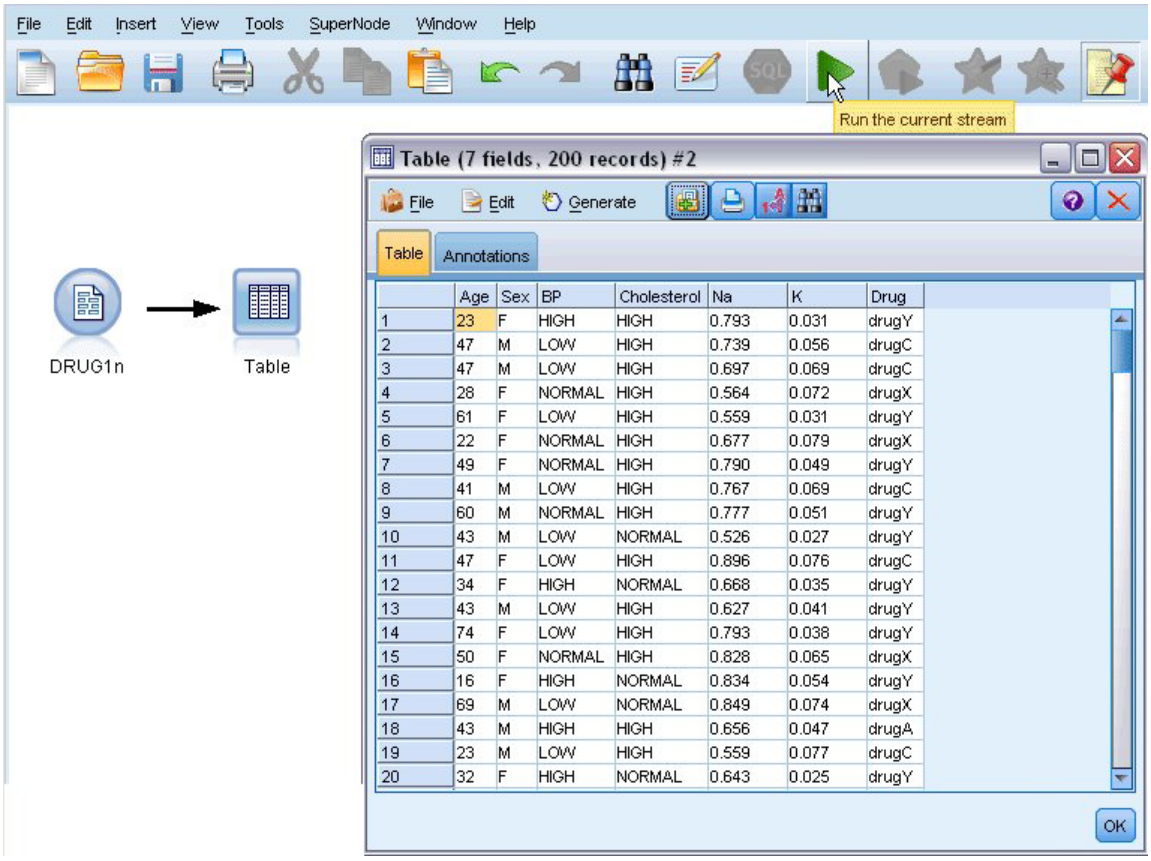


図 82. ツールバーからのストリームの実行

パレットでノードをダブルクリックすると、ストリーム領域で選択されているノードに自動的に接続されます。また、ノードがまだ接続されていない場合、マウスの中央ボタンを使用して入力ノードをテーブル・ノードに接続できます。マウスの中央ボタンをシミュレートするには、Alt キーを押しながらマウスを使用し

ます。テーブルを表示するには、ツールバーの緑色の矢印ボタンをクリックしてストリームを実行するか、テーブル・ノードを右クリックして「**実行**」を選択します。

分布図の作成

データ・マイニングの際は、多くの場合、視覚的な要約を作成してデータを検討すると便利です。IBM SPSS Modeler では、要約するデータの種類に応じて、さまざまな種類のグラフを選択できます。例えば、薬品ごとに、効果の出た患者の比率を調べるには、分布ノードを使用します。

ストリームに分布ノードを追加して入力ノードに接続し、そのノードをダブルクリックして表示のオプションを編集します。

分布を表示する対象フィールドとして、「*薬品*」を選択します。その後、ダイアログ・ボックスで「**実行**」をクリックします。

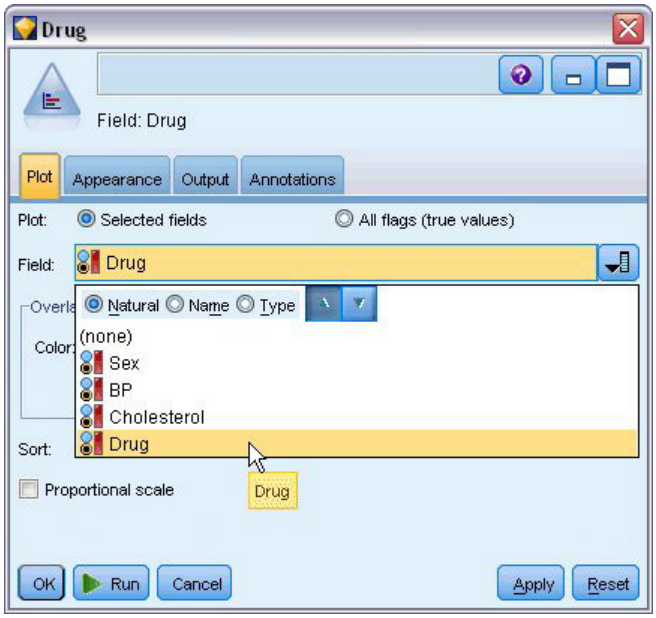


図 83. 対象フィールドとして薬品を選択

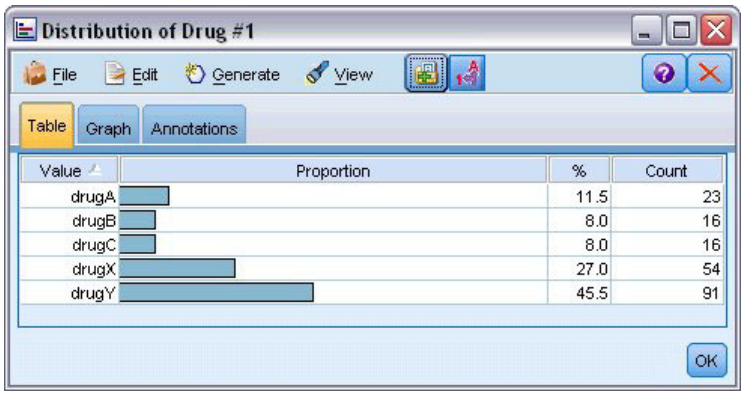


図 84. 薬品の種類に対する効果の分布

作成されたグラフは、データの「形状」を確認するのに役に立ちます。薬品 *Y* で効果が出た患者が最も多く、薬品 *B* および *C* が最も少ないことが分かります。

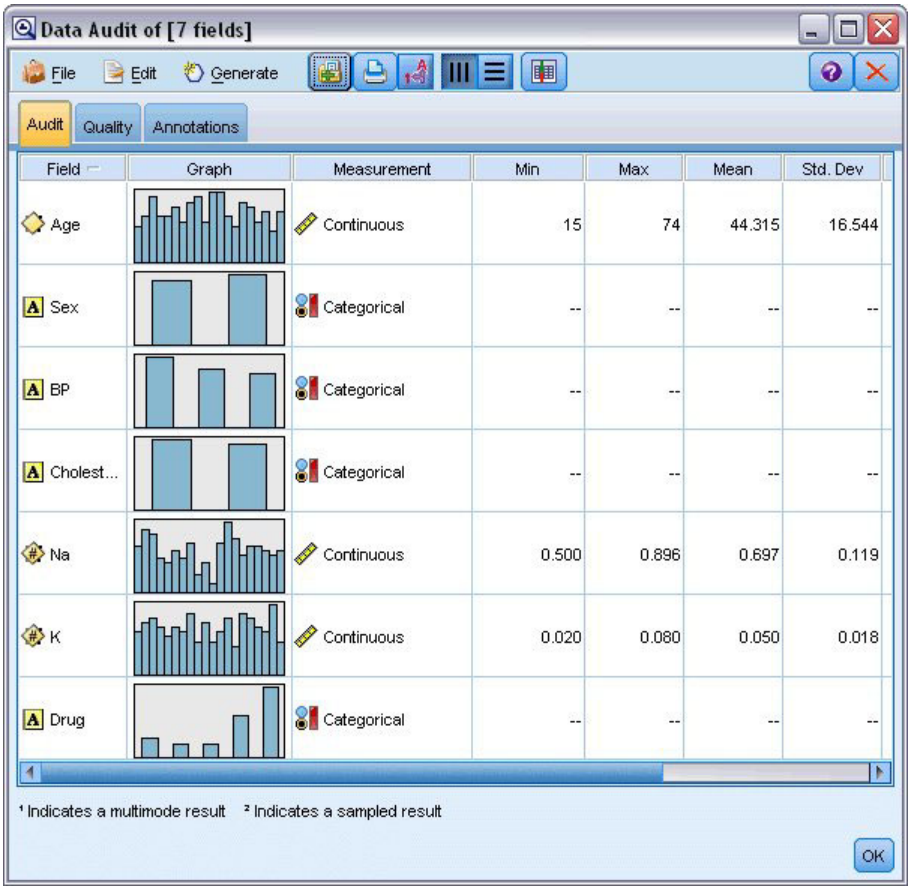


図 85. データ監査の結果

また、データ監査ノードを関連付けて実行すると、一度にすべてのフィールドの分布およびヒストグラムを一目で確認できます。データ監査ノードは、「出力」タブで使用可能です。

散布図の作成

どのような因子が *薬品* (対象変数) に影響を与えるかを見てみましょう。研究者としては、ナトリウムおよびカリウムの血中濃度が重要な因子であることが分かっています。これらは両方とも数値であるため、薬品のカテゴリーを色のオーバーレイとして使用して、ナトリウムとカリウムの散布図を作成することができます。

ワークスペースに散布図ノードを配置して入力ノードに接続し、ダブルクリックしてノードを編集します。

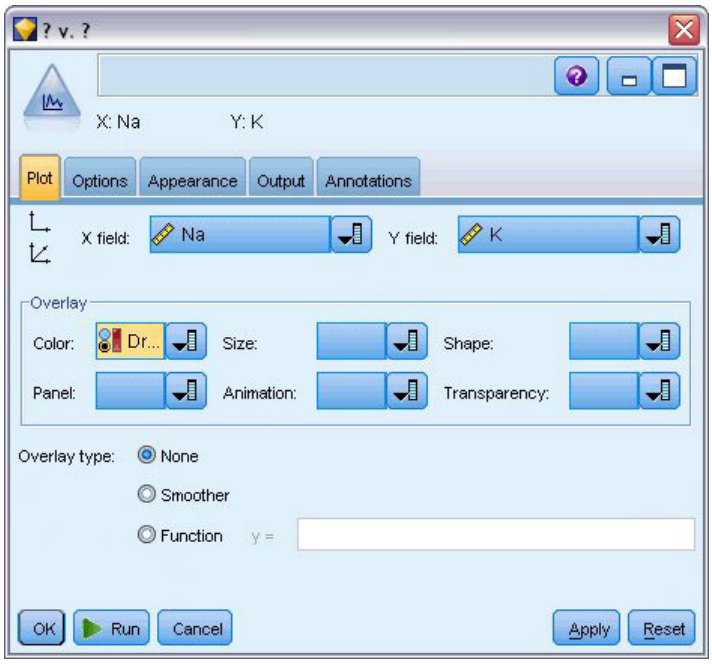


図 86. 散布図の作成

「プロット」タブで、「X フィールド」にナトリウム、「Y フィールド」にカリウム、オーバーレイ・フィールドに*薬品* を選択します。次に、「**実行**」をクリックします。

このプロットは明らかなしきい値を示しており、このしきい値を上回る場合に正しい薬品は常に薬品 *Y* であり、このしきい値を下回る場合に正しい薬品は決して薬品 *Y* ではありません。このしきい値は比率であり、ナトリウム (*Na*) とカリウム (*K*) の比率を示しています。

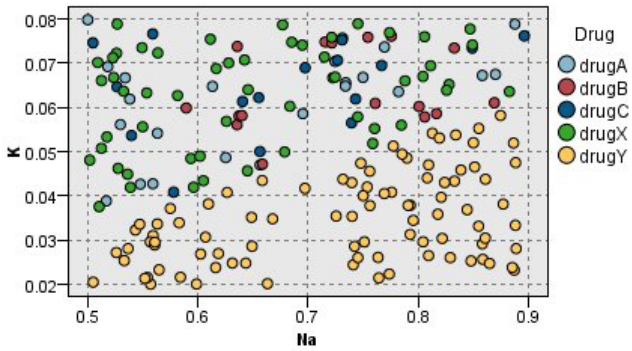


図 87. 薬品分布の散布図

Web グラフの作成

データ・フィールドの多くはカテゴリーのため、Web グラフのプロットを試すこともできます。これによってさまざまなカテゴリー間の関連性をマップすることができます。Web グラフ・ノードをワークスペースの入力ノードに接続して開始します。「Web グラフ・ノード」ダイアログ・ボックスで、血圧を表す「*BP*」と、「*薬品*」を選択します。次に、「**実行**」をクリックします。

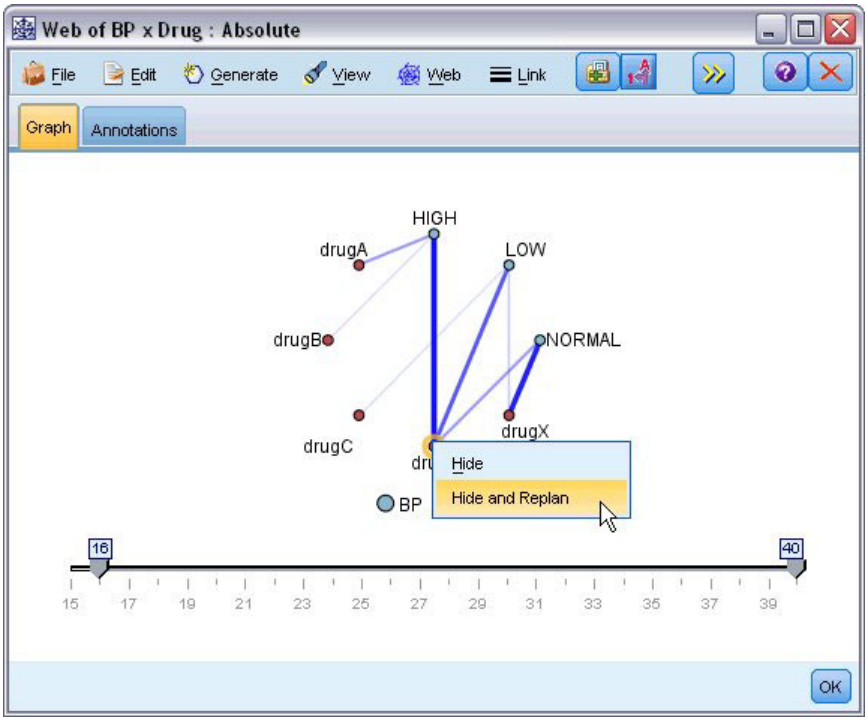


図 88. 薬品と血圧の Web グラフ

プロットから、薬品 *Y* がすべての 3 つのレベルの血圧と関連することが分かります。薬品 *Y* が最適な状況を既に特定したとしても驚きではありません。その他の薬品にも注目するために、薬品 *Y* を非表示にできます。「**表示**」メニューで、 「**編集モード**」を選択し、薬品 *Y* の点を右クリックして、「**非表示にして再計算**」を選択します。

簡略化されたプロットでは、薬品 *Y* およびそのすべてのリンクは非表示になります。ここで、薬品 *A* および *B* のみが高血圧と関係することが、はっきり分かるようになりました。薬品 *C* および *X* のみが低血圧と関係しています。血圧が通常の場合は、薬品 *X* のみに関係することも分かります。この時点で、薬品 *A* と *B* のどちらか、および薬品 *C* と *X* のどちらかを、対象の患者にたいしてどのように選択するかはまだ不明です。ここで、モデル作成が役立ちます。

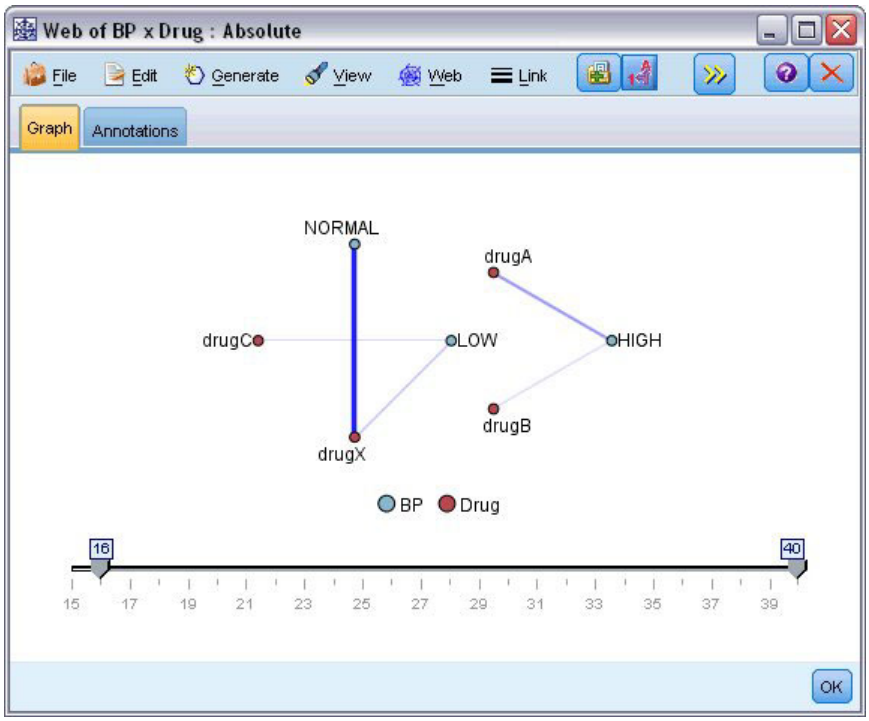


図 89. 薬品 Y とそのリンクが非表示になっている Web グラフ

新規フィールドの作成

カリウムに対するナトリウムの比率が、薬品 *Y* の使用タイミングを予測すると考えられるため、各レコードにこの比率の値を含むフィールドを作成します。このフィールドは、後で 5 つの薬品のそれぞれを使用するタイミングを予測するモデルを構築する際に役立つこともあります。ストリームのレイアウトを単純化するため、DRUG1n 入力ノード以外のすべてのノードを削除することから始めます。DRUG1n にフィールド作成ノード (「フィールド操作」タブ) を接続し、フィールド作成ノードをダブルクリックして編集します。

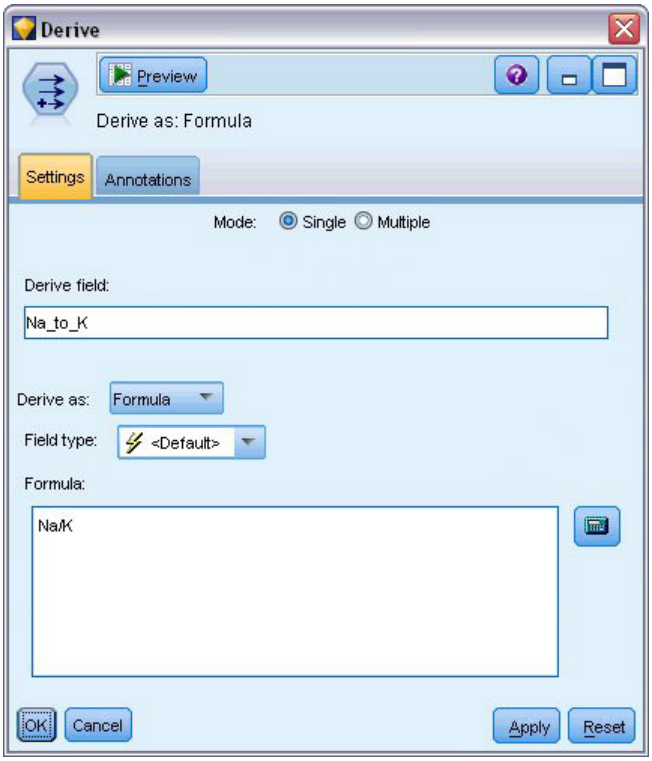


図 90. フィールド作成ノードの編集

新しいフィールドの名前を *Na_to_K* とします。ナトリウム値をカリウム値で除算することで新しいフィールドを取得するため、式に `Na/K` を入力します。フィールドの右側にあるアイコンをクリックして式を作成することもできます。その場合は式ビルダーが開くので、そこで組み込みの関数、オペランド、フィールドとその値のリストを使用して、対話的に式を作成できます。

新規フィールドの分布を確認するには、ヒストグラム・ノードをフィールド作成ノードに接続します。「ヒストグラム・ノード」ダイアログ・ボックスで、プロットするフィールドとして「*Na_to_K*」を指定し、オーバーレイ・フィールドとして「*薬品*」を指定します。

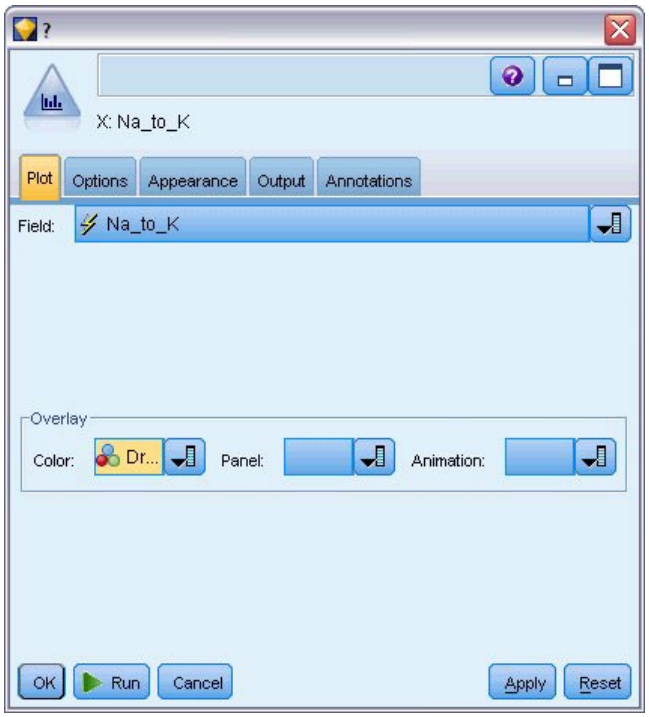


図 91. ヒストグラム・ノードの編集

このストリームを実行すると、次のようなグラフが表示されます。表示内よって、*Na_to_K* 値が約 15 以上の場合は薬品 *Y* が適していると結論づけることができます。

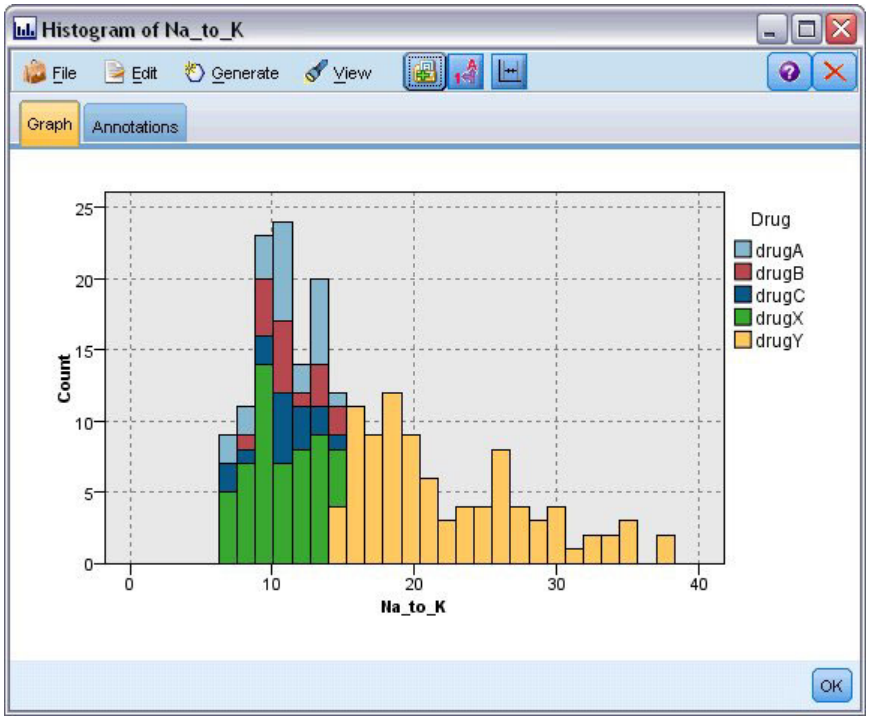


図 92. ヒストグラムの表示

モデルの構築

データの探索と操作により、いくつかの仮説を立てることができます。血中のカリウムに対するナトリウムの比率が、血圧と同様に薬品の選択に影響するようです。ただし、これですべての関係性を完全に説明することはできません。このような場合にモデルを作成すると、答えが得られることがあります。この例では、ルール構築モデル C5.0 を使用して、データへの適合を試みます。

派生フィールド (*Na_to_K*) を使用しているため、元のフィールド (*Na* および *K*) をフィルターで除外して、これらのフィールドがモデル作成アルゴリズムで 2 度使用されないようにできます。これはフィルター・ノードを使用して実現できます。

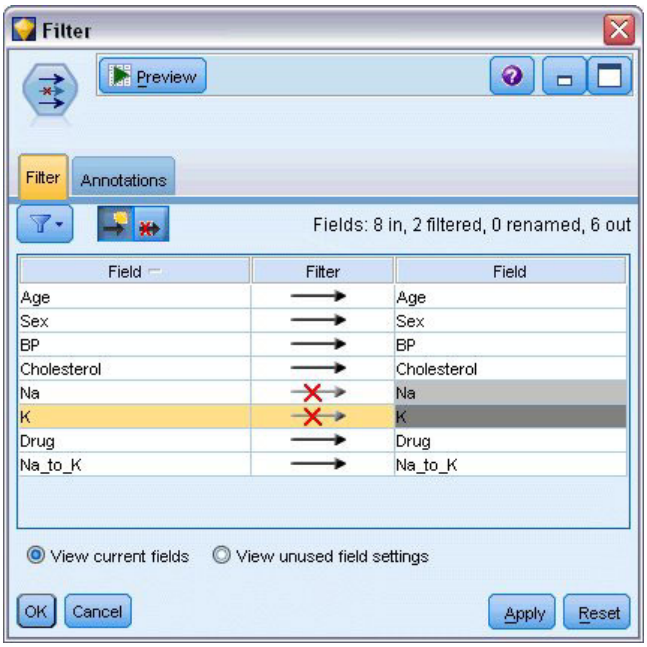


図 93. フィルター・ノードの編集

「フィルター」タブで、「*Na*」および「*K*」の横の矢印をクリックします。その矢印の上に赤い X が表示されて、そのフィールドがフィルターで除外されることが示されます。

次に、データ型ノードをフィルター・ノードに接続します。このデータ型ノードにより、使用するフィールドのデータ型、およびそれを結果の予測にどのように使用するかを示すことができます。

「データ型」タブで、「*薬品*」フィールドの役割を「**対象**」に設定して、「*薬品*」が予測対象フィールドであることを指定します。その他のフィールドの役割は「**入力**」のままにして、これらが予測値として使用されるようにします。

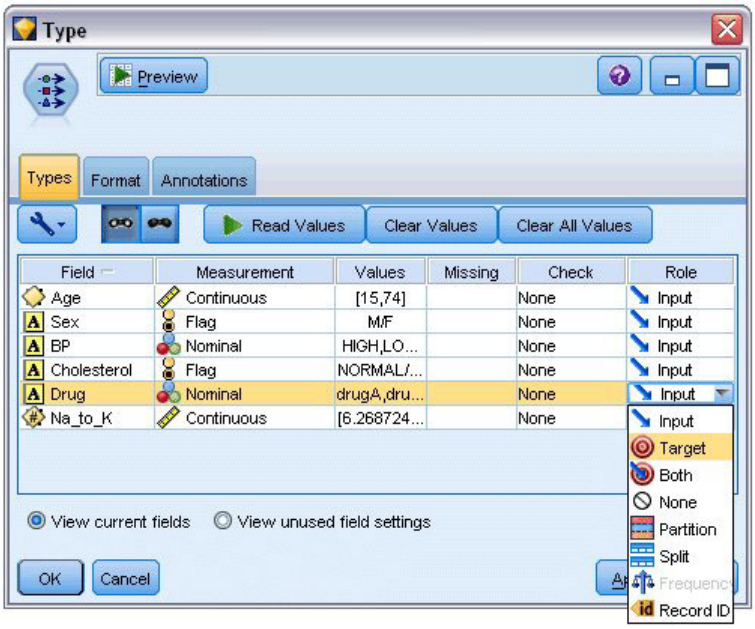


図 94. データ型ノードの編集

モデルを推定するには、次に示すように C5.0 ノードをワークスペースに配置し、ストリームの末尾に接続します。その後、ツールバーの緑色の「**実行**」ボタンをクリックしてストリームを実行します。

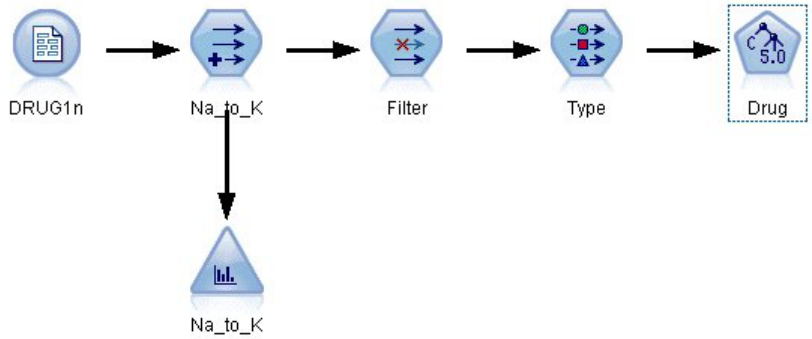


図 95. C5.0 ノードの追加

モデルの参照

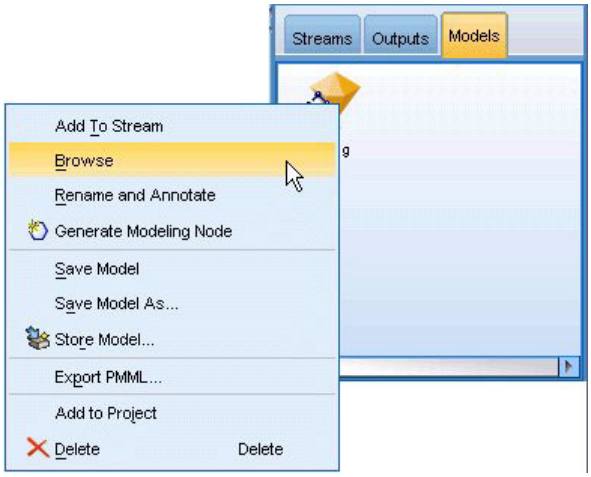


図 96. モデルの参照

C5.0 ノードを実行すると、モデル・ナゲットがストリーム、およびウィンドウ右上の「モデル」パレットにも追加されます。そのモデルを参照するには、いずれかのアイコンを右クリックして、コンテキスト・メニューから「**編集**」または「**参照**」を選択します。

「ルール」ブラウザーに、C5.0 ノードによって生成されたルール・セットがデシジョン・ツリー形式で表示されます。最初の状態では、ツリーは省略されています。展開するには、「**すべて**」ボタンをクリックしてすべてのレベルを表示します。

図 97. 「ルール」ブラウザー

これで、パズルの欠けていたピースが表示されました。*Na* 対 *K* の比率が 14.64 未満で高血圧の人は、年齢によって薬品の選択が決定されます。低血圧の人の場合は、コレステロール・レベルが最適な予測値と考えられます。

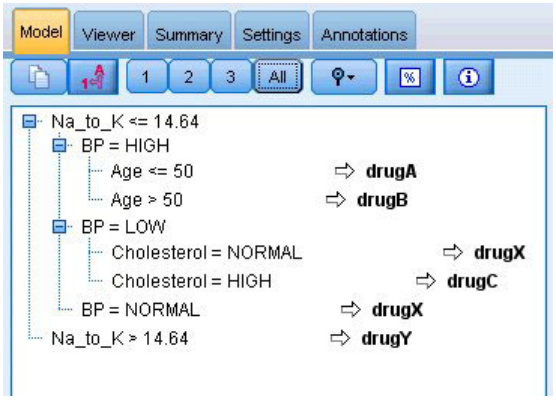


図 98. すべて展開された「ルール」ブラウザー

「**ビューアー**」タブをクリックすると、同じデシジョン・ツリーをより洗練されたグラフィカル形式で表示できます。ここでは、血圧カテゴリーごとのケース数およびケースのパーセントをより簡単に確認できます。

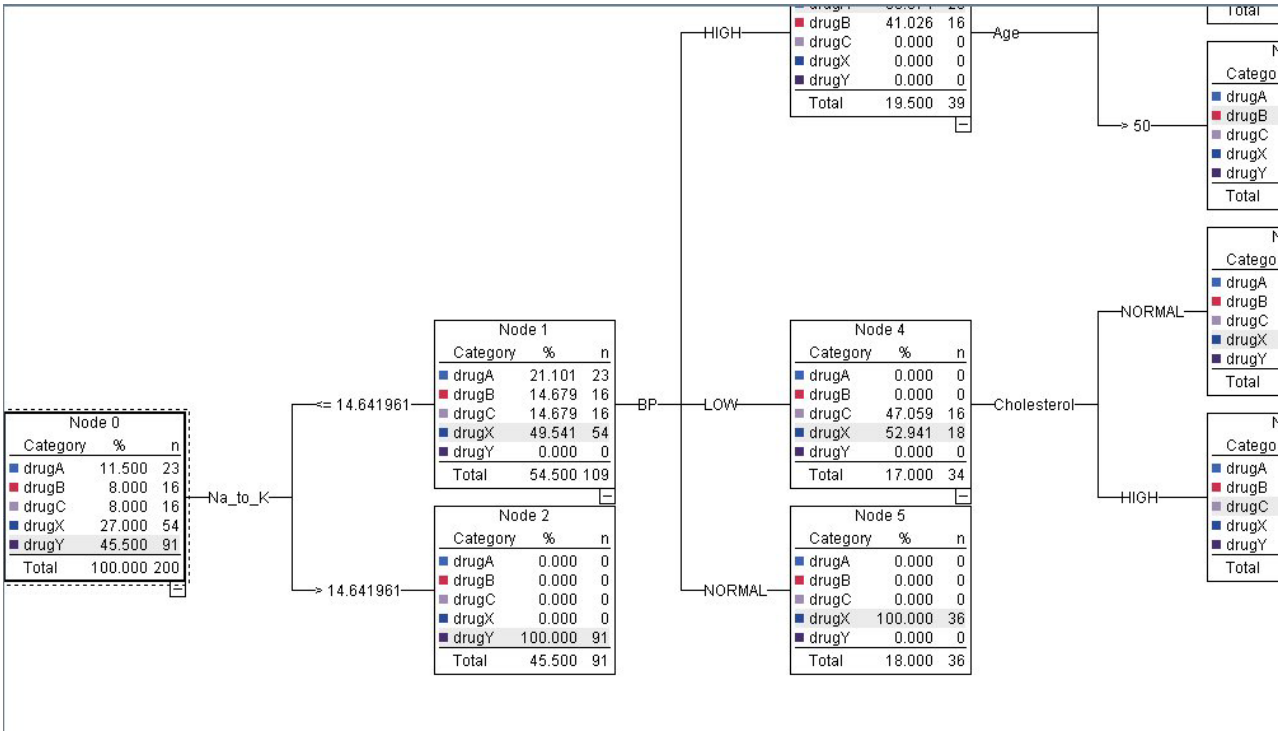


図 99. グラフィカル形式のデシジョン・ツリー

精度分析ノードの使用

精度分析ノードを使用して、モデルの精度を評価できます。精度分析ノードを出力ノード・パレットからモデル・ナゲットに接続し、精度分析ノードを開き、「**実行**」をクリックします。

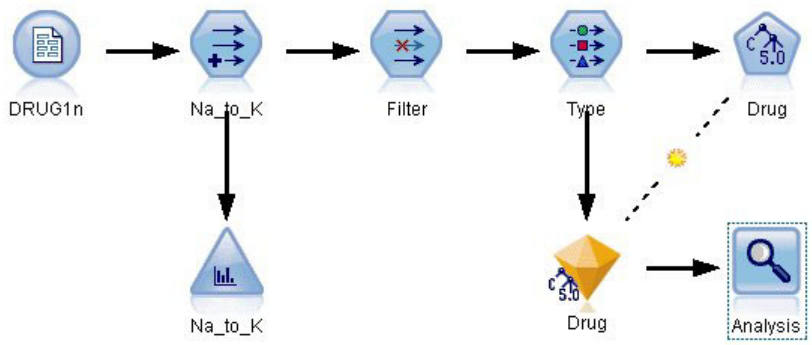


図 100. 精度分析ノードの追加

精度分析ノードの出力は、サンプル・データ・セットにおいて、モデルがこのデータ・セット内のすべてのレコードの薬品の選択を正しく予測したことを示しています。実際のデータ・セットでは、100% の精度はめったに実現しませんが、そのモデルが特定の適用に対して容認できる精度かどうかの判断を支援する上で、精度分析ノードを役立てることができます。

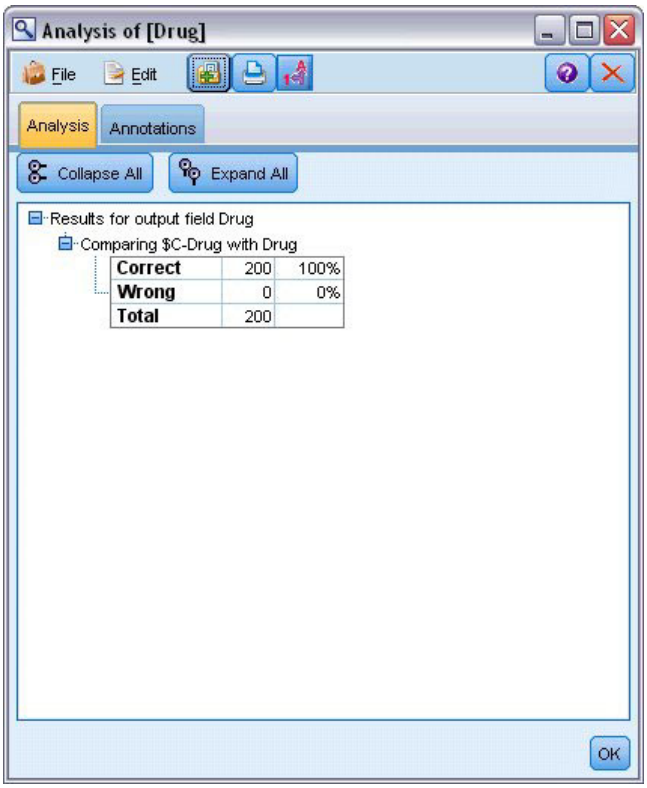


図 101. 精度分析ノードの出力

第 9 章 予測フィールドのスクリーニング (フィールド選択)

フィールド選択ノードは、ある結果を予測する上で最も重要なフィールドを識別するのに役立ちます。数百、数千の予測フィールド・セットから、フィールド選択ノードは最も重要と思われる予測フィールドをスクリーニング、ランク付け、および選択します。最終的には、より簡単でより効果的なモデル、すなわち少ない予測フィールドを使用して、すぐに実行できる分かりやすいモデルになります。

この例で使用するデータは、架空の電話会社のデータウェアハウスを想定しており、この会社の 5,000 人の顧客から得る特別プロモーションに対する応答に関する情報があります。このデータには、顧客の年齢、職業、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 3 つのオファーのそれぞれ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反応する可能性が最も高い顧客を予測したいと考えています。

この例では、*featureselection.str* という名前のストリームを使用します。これは、*customer_dbase.sav* という名前のデータ・ファイルを参照します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*featureselection.str* ファイルは、*streams* ディレクトリーにあります。

この例では、オファーの 1 つに注目して対象として取り上げます。CHAID ツリー構築ノードを使用して、販売促進活動に反応する可能性が最も高い顧客を示すモデルを作成します。ここで次の 2 つの方法を比較します。

- フィールド選択なし。データ・セットのすべての予測フィールドが CHAID ツリーへの入力として使用されます。
- フィールド選択あり。フィールド選択ノードを使用して、上位 10 の予測フィールドを選択します。それが CHAID ツリーに入力されます。

この 2 つの結果ツリー・モデルを比較することで、フィールド選択がいかに有効な成果を生むかがわかります。

ストリームの構築

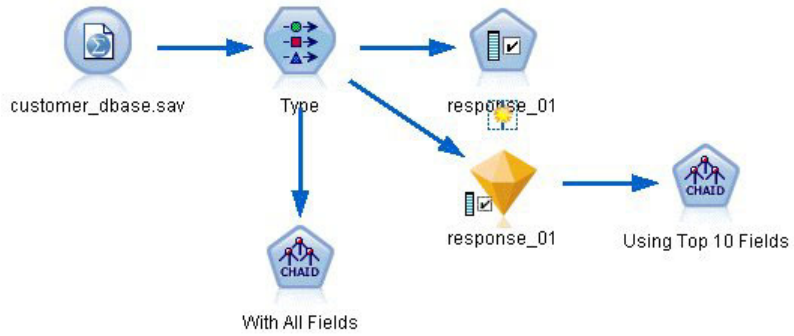


図 102. フィールド選択ストリームの例

1. Statistics ファイル入力ノードを空白のストリーム領域に移動します。このノードをサンプル・データ・ファイル *customer_dbase.sav* に接続します。このファイルは、IBM SPSS Modeler インストール・フォルダーの下の *Demos* ディレクトリーにあります。(あるいは、*streams* ディレクトリーのサンプル・ストリーム・ファイル *featureselection.str* を開きます。)
2. データ型ノードを追加します。「データ型」タブで、一番下までスクロールし、*response_01* の役割を「対象」に変更します。他の応答フィールド (*response_02* および *response_03*)、さらにリスト最上部の顧客 ID (*custid*) については、役割を「なし」に変更します。これら以外のすべてのフィールドの役割は、「入力」のままにして、「**値の読み取り**」ボタンをクリックしてから、「**OK**」をクリックします。

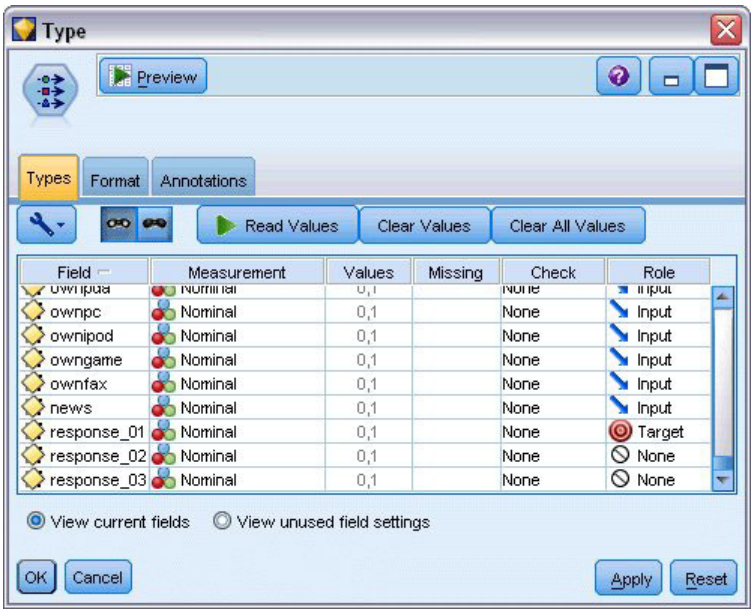


図 103. データ型ノードの追加

3. フィールド選択モデル作成ノードをストリームに追加します。このノードには、スクリーニングまたは不認定フィールドのルールおよび基準を指定できます。
4. ストリームを実行して、フィールド選択モデル・ナゲットを作成します。
5. ストリーム上またはモデル・パレット内のモデル・ナゲットを右クリックし、「**編集**」または「**参照**」を選択して結果を確認します。

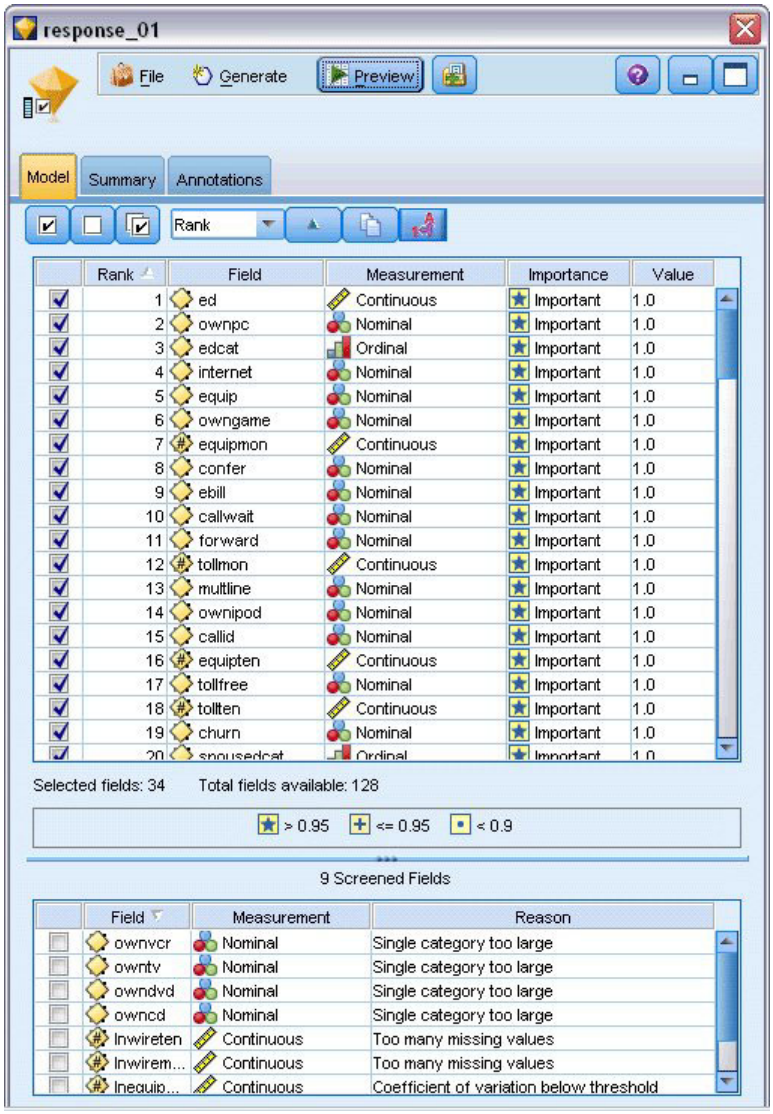


図 104. フィールド選択モデル・ナゲットの「モデル」タブ

上部のパネルには、予測に有用と判断されたフィールドが表示されます。これらのフィールドは重要度に基づいてランク付けされています。下部のパネルには、解析によってスクリーニングされたフィールドと、その理由が示されます。上部のパネルにあるフィールドを検証して、この後のモデリング・セッションに使用するフィールドを決定することができます。

6. これで、下流で使用するフィールドを選択できるようになりました。本来は 34 のフィールドが重要なフィールドとして識別されていましたが、予測値セットをさらに絞り込む必要があります。
7. 先頭の列にチェック・マークを付けて、上位 10 の予測値のみを選択し、不要な予測値を選択解除します。(11 行目のチェック・マークをクリックし、Shift キーを押しながら 34 行目のチェック・マークをクリックします。)モデル・ナゲットを閉じます。
8. フィールド選択なしで結果を比較するには、2 つの CHAID モデル作成ノードをストリームに追加する必要があります。一方ではフィールド選択を使用し、もう一方では使用しません。
9. 1 つの CHAID ノードをデータ型ノードに接続し、もう 1 つをフィールド選択モデル・ナゲットに接続します。

10. 各 CHAID ノードを開き、「作成オプション」タブを選択して、「目的」ペインの「**新規モデルを作成**」、「**単一ツリーを作成**」、「**インタラクティブ・セッションを起動**」の 3 つのオプションが選択されていることを確認します。

「基本」ペインで、「**ツリーの最大深度**」が 5 に設定されていることを確認します。

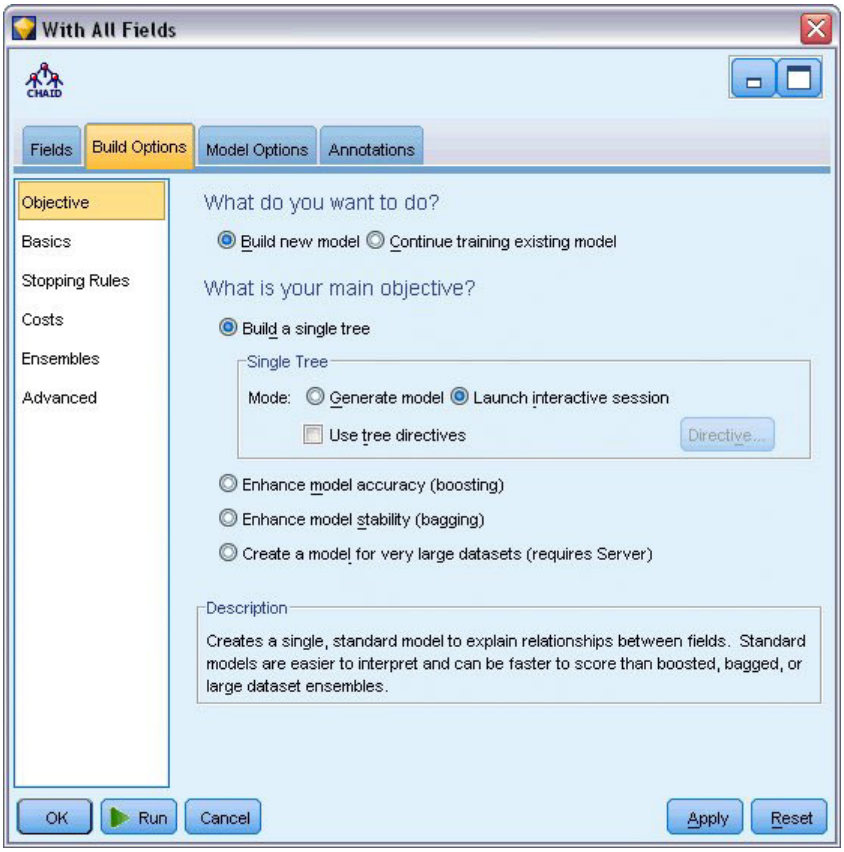


図 105. すべての予測フィールドに使用する CHAID モデル作成ノードの「目的」設定

モデルの構築

1. データ・セット内のすべての予測値を使用する CHAID ノード (データ型ノードに接続した方) を実行します。その際、実行に掛かる時間に注意します。結果ウィンドウにテーブルが表示されます。
2. メニューから、「**ツリー**」 > 「**ツリーを成長**」を選択し、展開したツリーを成長させて表示します。

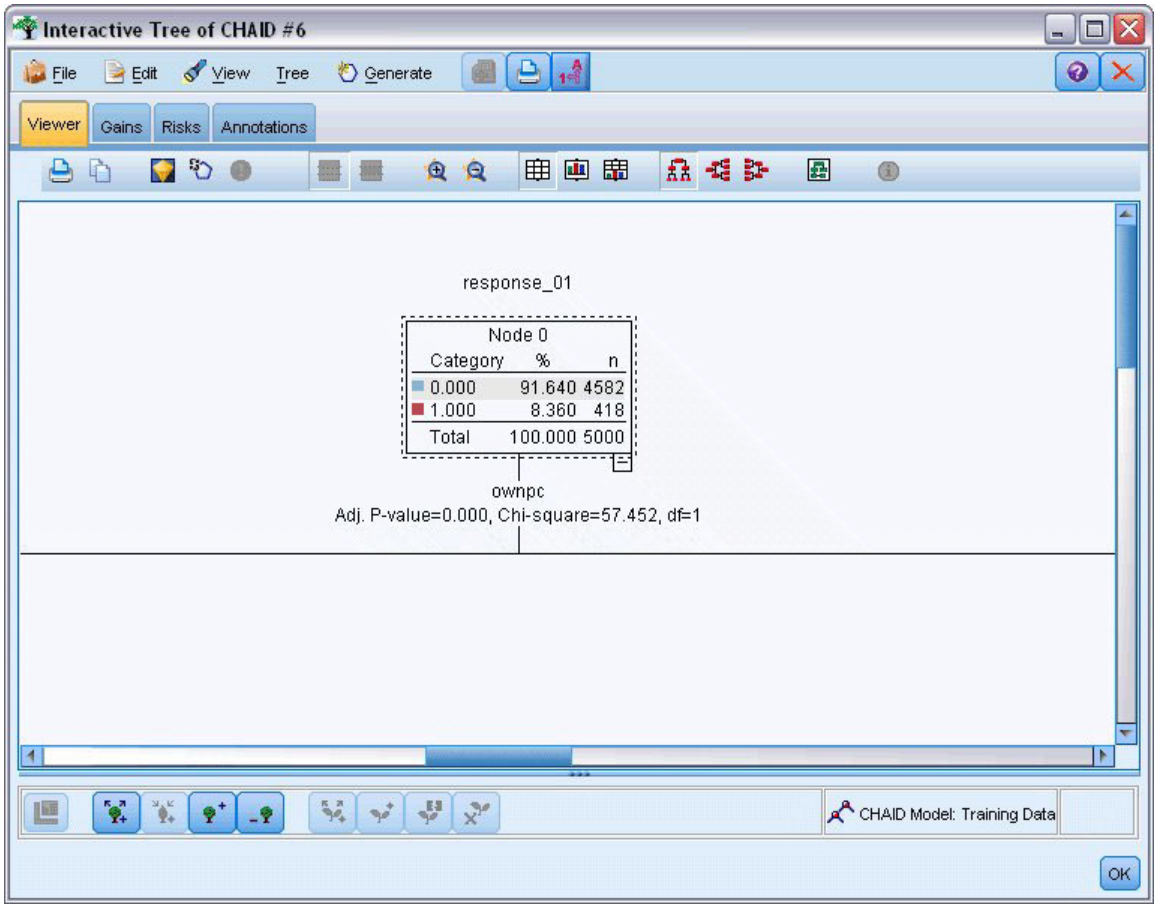


図 106. ツリー・ビルダーでのツリーの成長

3. 次に、10 個の予測値のみを使用するその他の CHAID ノードに対して同じ手順を実行します。ツリー・ビルダーが開いたら、同じようにツリーを成長させます。

2 番目のモデルは、1 番目のモデルよりも短時間で実行されたはずです。このデータ・セットはかなり小さいので、実行時間の差は数秒かもしれませんが、実際のデータ・セットはもっと大きいため、この差は数分から数時間と顕著になる可能性があります。フィールド選択を使用すると、処理時間を大幅に短縮することができます。

また、2 番目のツリーは 1 番目のツリーに比べて、ツリー・ノードも少数です。そのため理解しやすくなっています。ただし、2 番目の方法を採用する前に、それが効果的かどうか、またすべての予測値を使用するモデルとの比較方法を確認する必要があります。

結果の比較

2 つの結果を比較するには、効率の測定が必要です。これには、ツリー・ビルダーの「ゲイン」タブを使用します。ここで**リフト**を調べます。これは、データ・セットのすべてのレコードを比較した場合、ノード内のどのレコードが対象カテゴリーに分類される可能性がより高いかを示します。例えば、リフト値が 148% の場合、データ・セット内のすべてのレコードよりも、ノード内のレコードの方が対象カテゴリーに分類される可能性が 1.48 倍高いことを示します。リフトは「ゲイン」タブの「インデックス」列に示されます。

1. すべての予測値を示すツリー・ビルダーで、「ゲイン」タブをクリックします。「対象カテゴリー」を 1.0 に変更します。「分位」ツールバー・ボタンを最初にクリックして、表示を四分位に切り替えます。次に、このボタンの右側のドロップダウン・リストから、**「四分位」**を選択します。

2. 10 の予測値に対してツリー・ビルダーでこの手順を繰り返すと、次の図のように 2 つの同じようなゲイン・テーブルが表示され、比較できるようになります。

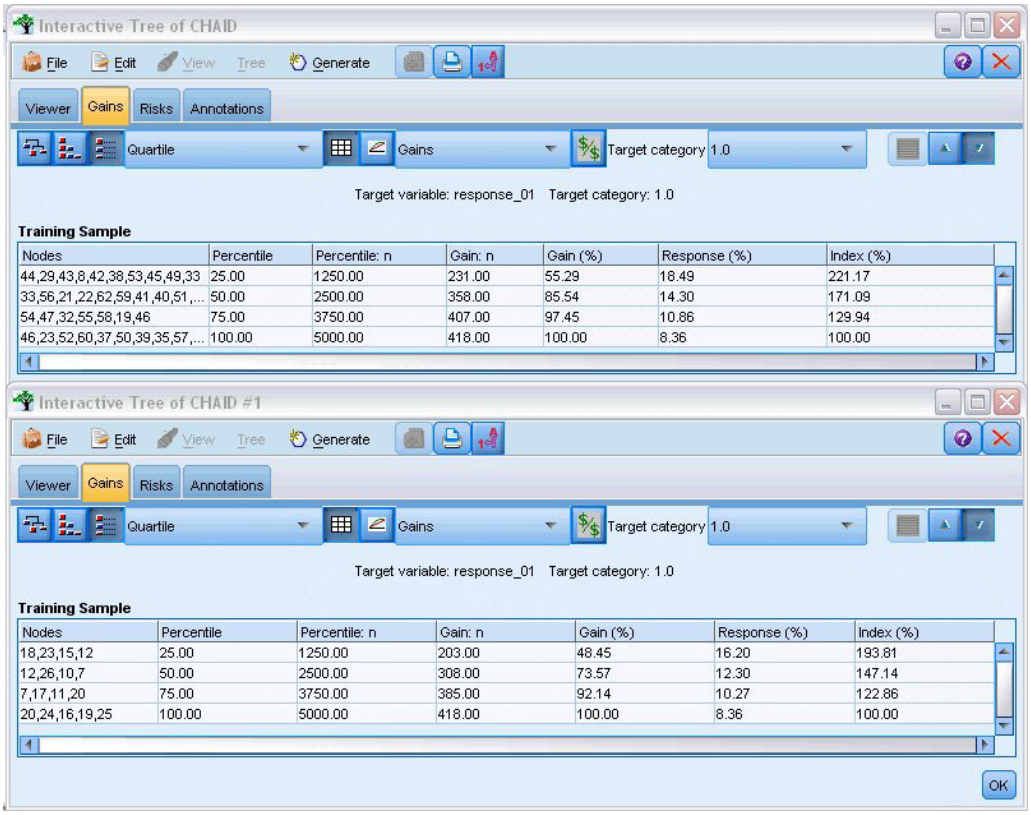


図 107. 2 つの CHAID モデルのゲイン・チャート

各ゲイン・テーブルは、そのツリーのターミナル・ノードを四分位にグループ化しています。2 つのモデルの有効性を比較するには、各テーブルの最上位四分位のリフト (インデックス 値) を調べます。

すべての予測値が含まれるときは、モデルは 221% のリフトを示しています。つまり、これらのノードの特徴を持つケースは、対象の販売促進活動に対して 2.2 倍反応する可能性が高いと思われます。これらの特徴の内容を確認するには、最上位の行をクリックして選択します。次に、「ビューアー」タブに切り替えると、そこに該当するノードが外枠を黒で縁取られて表示されます。強調表示されている各ターミナル・ノードまでツリーを下にたどり、予測値がどのように分割されているかを確認します。上位四分位だけで 10 ノードが含まれています。実際のスコアリング・モデルに変換すると、10 件の顧客プロファイルの管理は困難になります。

フィールド選択によって識別された上位 10 の予測値だけが含まれると、リフトは約 194% になります。このモデルは、すべての予測値を使用するモデルほど優れているとは言えませんが、有用であることは間違いありません。ここでは、上位四分位に含まれるノードは 4 つだけなので、よりシンプルになります。したがって、フィールド選択モデルの方が、すべての予測値を含むモデルよりも望ましいと判断できます。

要約

フィールド選択の利点を確認してみましょう。使用する予測値が少ないほど低コストになります。つまり、収集、処理、モデルに送信するデータが少なくなります。計算時間が短縮されます。この例では、フィールド選択の手順が増えたにも関わらず、モデル構築は予測数が少ない方が大幅に速くなりました。実際のデータ・セットはさらに大きくなるため、節約できる時間が大幅に増えるのは確かです。

使用する予測値が少ないほど、スコアリングはシンプルになります。例が示すように、販売促進活動に反応しそうな顧客のプロファイルを 4 つだけ識別します。予想値の数が多くなると、モデルでオーバーフィッティングが発生する恐れがあります。シンプルなモデルの方が他のデータ・セットに対してよりうまく一般化する可能性があります (ただし念のため、テストする必要があります)。

ツリー構築アルゴリズムを使用してフィールド選択作業を行うと、最も重要な予測値をツリーで識別できるようになります。実際に、CHAID アルゴリズムはこの目的のために使用されることが多く、1 レベルずつツリーを成長させてツリーの深度と複雑性をコントロールすることも可能です。ただし、フィールド選択ノードの方が処理が速く、使い方も簡単です。すべての予測値を 1 ステップで素早くランク付けし、最も重要なフィールドを迅速に識別できるようにします。また、含める予測値の数を変更することもできます。上位 10 個の代わりに、上位 15 個または 20 個の予測値を使用してこの例を再度実行し、結果を比較して、最適モデルを決定することも簡単にできます。

第 10 章 入力データ文字列の長さの短縮 (データ分類ノード)

入力データ文字列の長さの短縮 (分類)

2 項ロジスティック回帰モデル、および 2 項ロジスティック回帰モデルを含む自動分類モデルの場合、文字列フィールドは最大 8 文字に制限されます。文字列が 8 文字を超える場合、データ分類ノードを使用して再コード化することができます。

この例では、*drug_long_name* というデータ・ファイルを参照する *reclassify_strings.str* というストリームを使用します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*reclassify_strings.str* ファイルは、*streams* ディレクトリー内にあります。

この例では、ストリームの小さな一部に焦点を当て、長すぎる文字列で生成されることがある種類のエラーを示し、データ分類ノードを使用して文字列の詳細を受け入れられる長さに変更する方法を説明します。この例では 2 項ロジスティック回帰ノードを使用しますが、自動分類ノードを使用して 2 項ロジスティック回帰モデルを生成した場合にも適用されます。

データの再分類

1. 変数ファイル入力ノードを使用して、*Demos* フォルダーのデータ・セット *drug_long_name* に接続します。

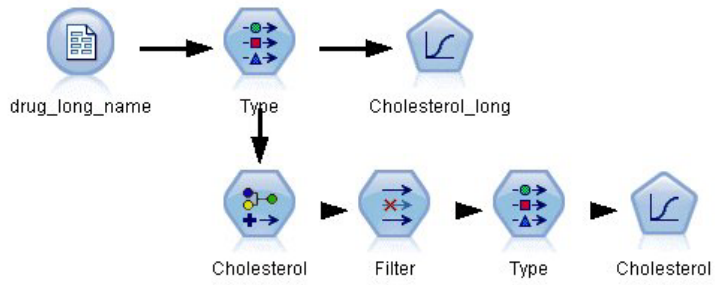


図 108. 2 項ロジスティック回帰の文字列分類を示すサンプル・ストリーム

2. データ型ノードを入力ノードに追加して、対象として **Cholesterol_long** を選択します。
3. ロジスティック回帰ノードをデータ型ノードに追加します。
4. ロジスティック回帰ノードで「モデル」タブをクリックし、「**2 項検定**」手続きを選択します。

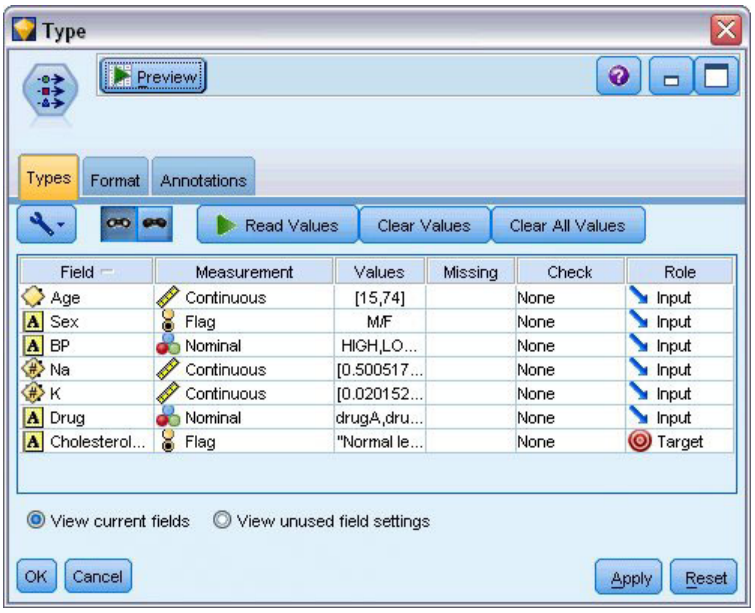


図 109. 「Cholesterol_long」フィールドの長い文字列の詳細

5. *reclassify_strings.str* でロジスティック回帰ノードを実行すると、**Cholesterol_long** 文字列値が長すぎることを警告するエラー・メッセージが表示されます。

   このタイプのエラー・メッセージが表示された場合、この例で後述している手順に従って、データを変更します。

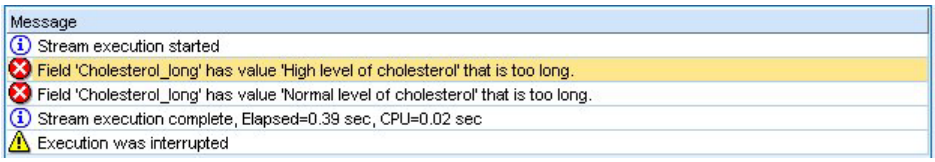


図 110. 2 項ロジスティック回帰ノード実行時に表示されたエラー・メッセージ

6. データ分類ノードをデータ型ノードに追加します。
7. データ分類フィールドで、「**Cholesterol_long**」 を選択します。
8. 新規フィールド名として 「**Cholesterol**」 と入力します。
9. 「**取得**」ボタンをクリックして、「**Cholesterol_long**」 値を元の値列に追加します。
10. 新しい値の列で、「**高レベルのコレステロール**」の元の値の隣に「**高**」と入力し、「**正常レベルのコレステロール**」の元の値の隣に「**正常**」と入力します。

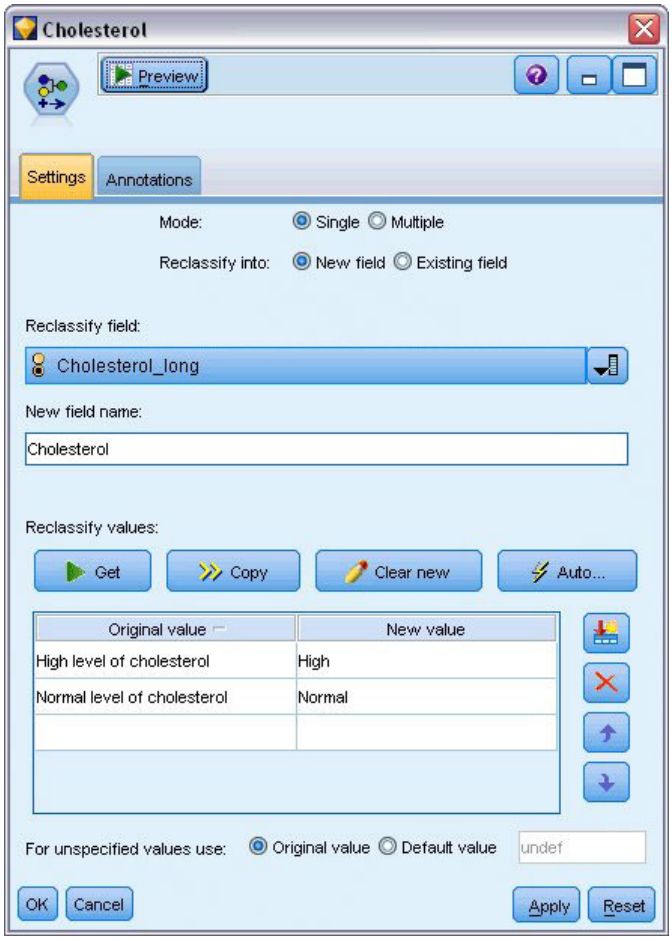


図 111. 長い文字列のデータ分類

11. フィルター・ノードをデータ分類ノードに追加します。
12. 「フィルター」列で、「**Cholesterol_long**」 をクリックして削除します。

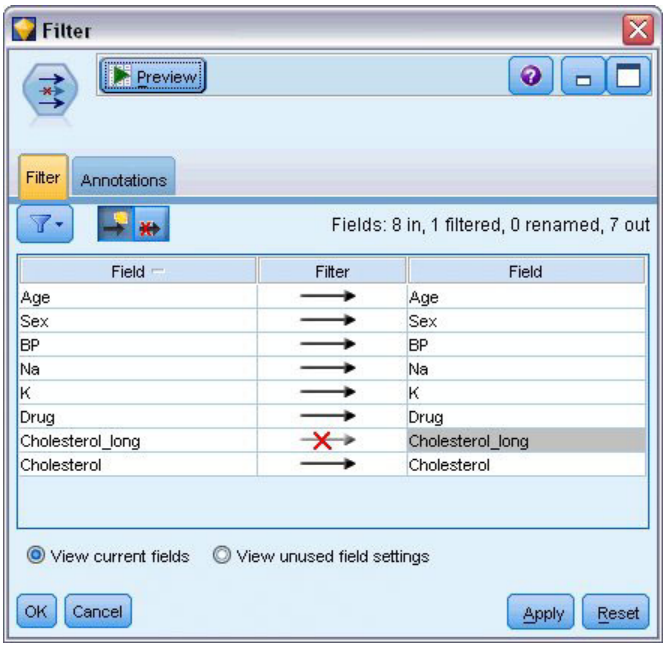


図 112. データからの「Cholesterol_long」フィールドのフィルタリング

13. データ型ノードをフィルター・ノードに追加して、対象として 「**Cholesterol**」 を選択します。

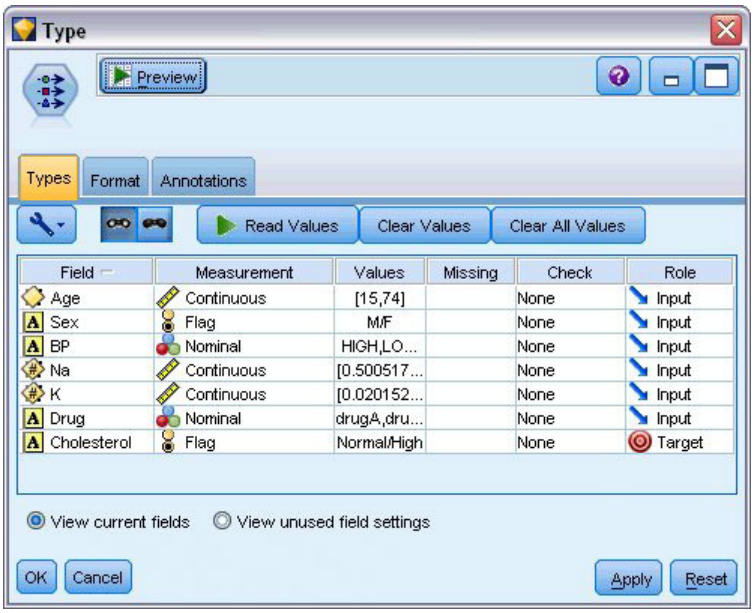


図 113. 「Cholesterol」フィールドの短い文字列の詳細

14. ロジスティック・ノードをデータ型ノードに追加します。
15. ロジスティック・ノードで「モデル」タブをクリックし、「**2 項検定**」手続きを選択します。
16. ここでは、2 項ロジスティック・ノードを実行し、エラー・メッセージが表示されることなくモデルを生成することができます。

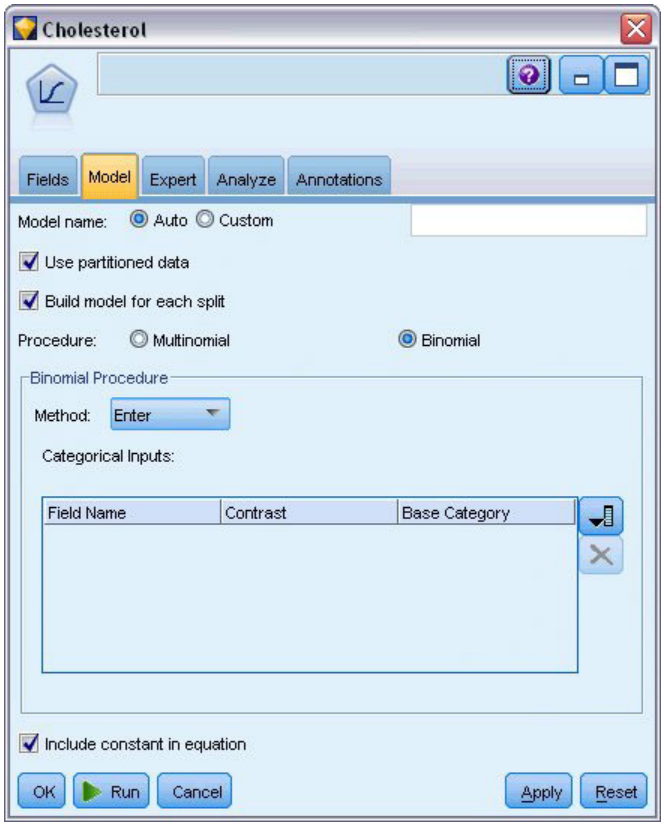


図 114. 手順としての 2 項検定の選択

この例はストリームの一部のみを示しています。長い文字列のデータ分類が必要になることがあるストリームのタイプの詳細については、次の例を参照してください。

- 自動分類ノード。詳細については、 39 ページの『顧客のレスポンスのモデル作成 (自動分類)』を参照してください。
- 2 項ロジスティック回帰ノード。詳細については、 141 ページの『第 13 章 電気通信会社の解約 (2 項検定ロジスティック回帰)』を参照してください。

IBM SPSS Modeler の使用方法の詳細 (ユーザーズ・ガイド、ノード・リファレンス、アルゴリズム・ガイドなど) は、インストール・ディスクの *¥Documentation* ディレクトリーからご利用いただけます。

第 11 章 顧客応答のモデル作成 (ディシジョン・リスト)

ディシジョン・リスト・アルゴリズムは、与えられた 2 値 (「はい」または「いいえ」) の結果のより高い、 またはより低い尤度を示す規則を生成します。ディシジョン・リストのモデルは、コール・センターやマーケティング・アプリケーションなどのカスタマー・リレーションシップ・マネジメントで幅広く使用されています。

この例は、それぞれの顧客に合った適切な提案を行うことで、今後のマーケティング・キャンペーンでさらに収益の高い結果を実現することを望んでいる架空の会社に基づいています。具体的にこの例では、ディシジョン・リスト・モデルを使用して、以前の販売促進活動に基づき、好意的な反応を示す可能性が高い顧客の特徴を識別し、その結果に基づいてメーリング・リストを生成します。

ディシジョン・リスト・モデルは特にインタラクティブ・モデル作成に適しており、モデル内のパラメーターを調整してすぐに結果を確認することができます。自動的にさまざまなモデルを作成して結果をランク付けすることができる別の手法として、自動分類ノードを代わりに使用することができます。

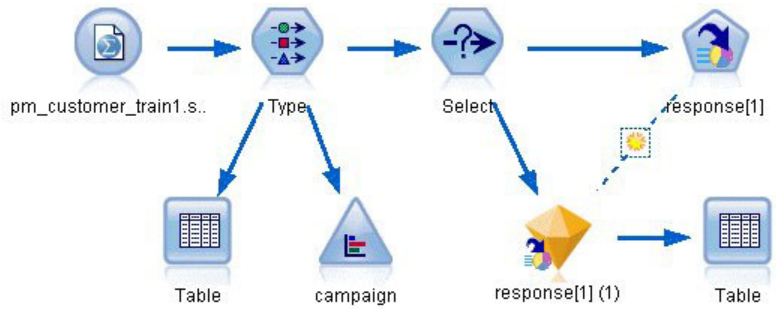
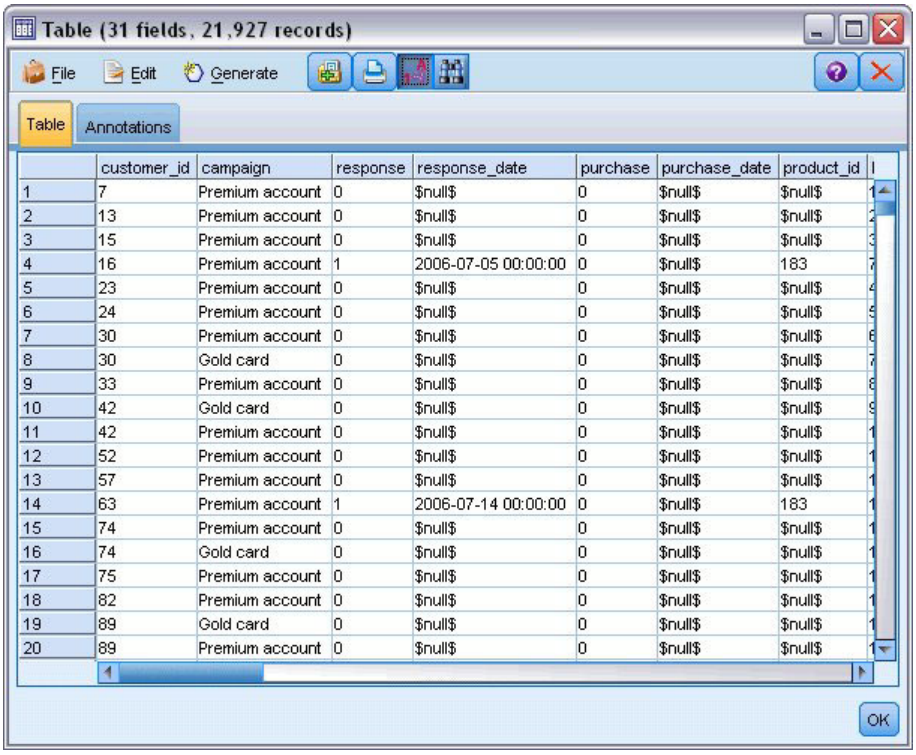


図 115. ディシジョン・リストのサンプル・ストリーム

この例では、データ・ファイル *pm_customer_train1.sav* を参照するストリーム *pm_decisionlist.str* を使用します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*pm_decisionlist.str* ファイルは、*streams* ディレクトリー内にあります。

履歴データ

ファイル *pm_customer_train1.sav* には、過去のキャンペーンでの特定のお客様に対する提案を追跡する履歴データが含まれており、「*campaign*」フィールドの値で示されています。最大多数のレコードが *Premium account* キャンペーンに分類されています。



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	l
1	7	Premium account	0	$null$	0	$null$	$null$	1
2	13	Premium account	0	$null$	0	$null$	$null$	2
3	15	Premium account	0	$null$	0	$null$	$null$	3
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183	7
5	23	Premium account	0	$null$	0	$null$	$null$	4
6	24	Premium account	0	$null$	0	$null$	$null$	5
7	30	Premium account	0	$null$	0	$null$	$null$	6
8	30	Gold card	0	$null$	0	$null$	$null$	7
9	33	Premium account	0	$null$	0	$null$	$null$	8
10	42	Gold card	0	$null$	0	$null$	$null$	9
11	42	Premium account	0	$null$	0	$null$	$null$	1
12	52	Premium account	0	$null$	0	$null$	$null$	1
13	57	Premium account	0	$null$	0	$null$	$null$	1
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183	1
15	74	Premium account	0	$null$	0	$null$	$null$	1
16	74	Gold card	0	$null$	0	$null$	$null$	1
17	75	Premium account	0	$null$	0	$null$	$null$	1
18	82	Premium account	0	$null$	0	$null$	$null$	1
19	89	Gold card	0	$null$	0	$null$	$null$	1
20	89	Premium account	0	$null$	0	$null$	$null$	1

図 116. 以前の販売促進活動に関するデータ

「*campaign*」フィールドの値は、実際には、データ内では整数としてコード化され、データ型ノードで定義されたラベルが付けられています (例えば、*2 = Premium account*)。ツールバーを使用して、テーブル内の値ラベルの表示を切り替えることができます。

このファイルには、各顧客のデモグラフィックおよび財務情報を含むフィールドが含まれており、これを使用して、特定の特徴に基づいてさまざまなグループの回答率を予測するモデルを作成または「学習」することができます。

ストリームの構築

1. IBM SPSS Modeler インストール環境の「*Demos*」フォルダーにある *pm_customer_train1.sav* を示す Statistics ファイル・ノードを追加します。(このフォルダーを参照するショートカットとして、ファイル・パスに `$CLEO_DEMOS/` を指定できます。)

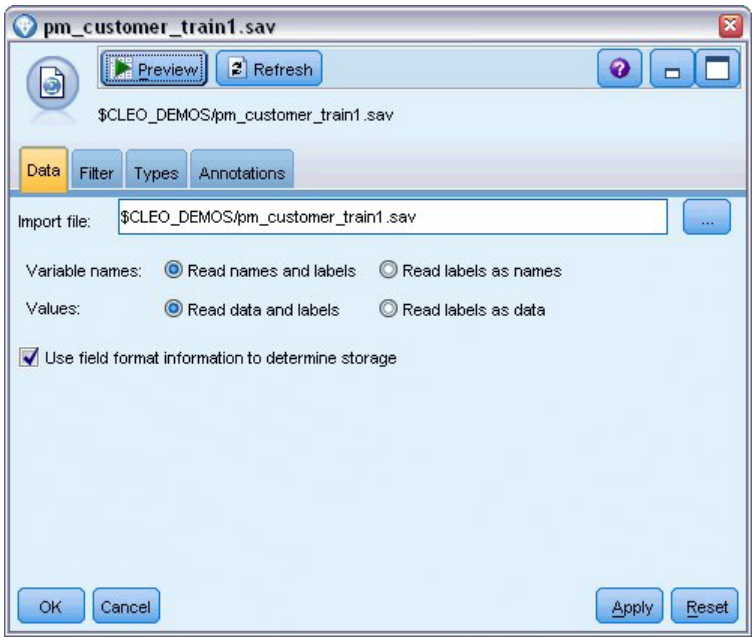


図 117. データの読み取り

2. データ型ノードを追加し、*response* を対象フィールドとして選択します (役割は「**対象**」)。このフィールドの測定の尺度を「**フラグ**」に設定します。

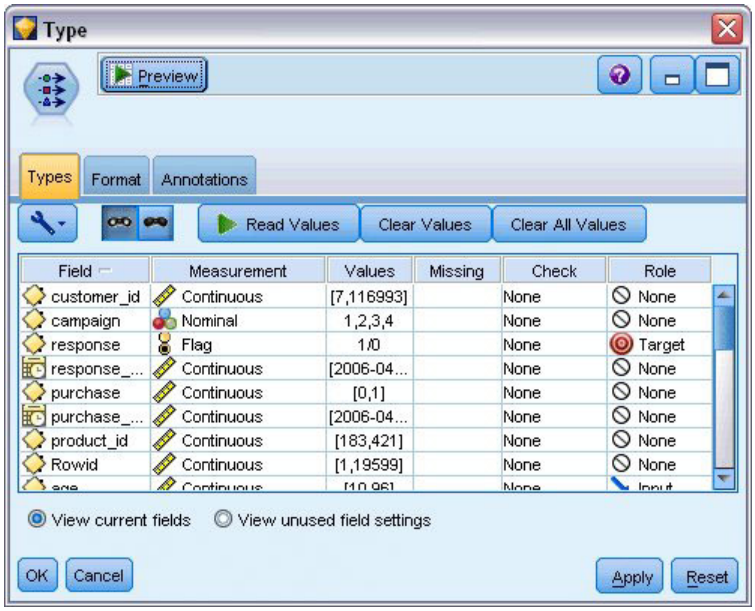


図 118. 測定の尺度および役割の設定

3. フィールド *customer_id*、*campaign*、*response_date*、*purchase*、*purchase_date*、*product_id*、*Rowid*、および *X_random* の役割を「**なし**」に設定します。これらのフィールドはすべてデータ内で使用されますが、実際のモデル作成には使用されません。
4. データ型ノードで「**値の読み込み**」ボタンをクリックして、値がインスタンス化されていることを確認します。

データには 4 つの異なるキャンペーンに関する情報が含まれますが、分析は一度に 1 つのキャンペーンに集中して行います。最も多くのレコードは Premium キャンペーンに分類される (データでは *campaign* = *2* にコード化されている) ため、条件抽出ノードを使用してこれらのレコードのみがストリームに含まれるようにできます。

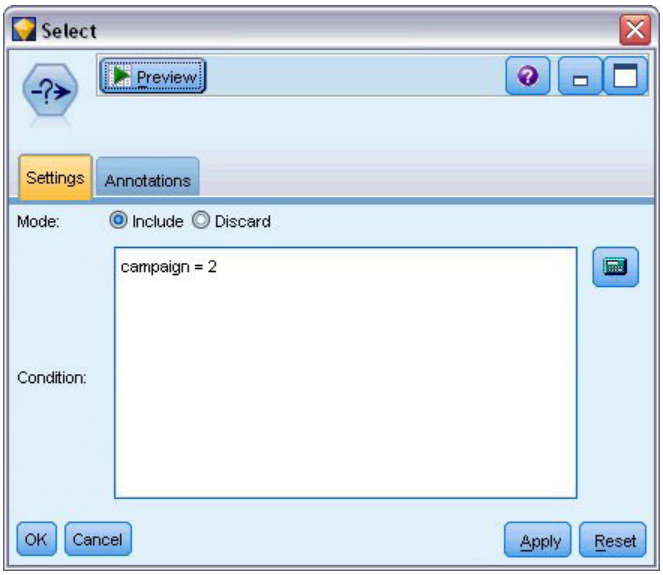


図 119. 単一キャンペーンのレコードの選択

モデルの作成

1. ディシジョン・リスト・ノードをストリームに接続します。「モデル」タブで、「**対象の値**」を 1 に設定して、検索する結果を指定します。この場合は、以前のオファーで「はい」と答えた顧客を検索します。

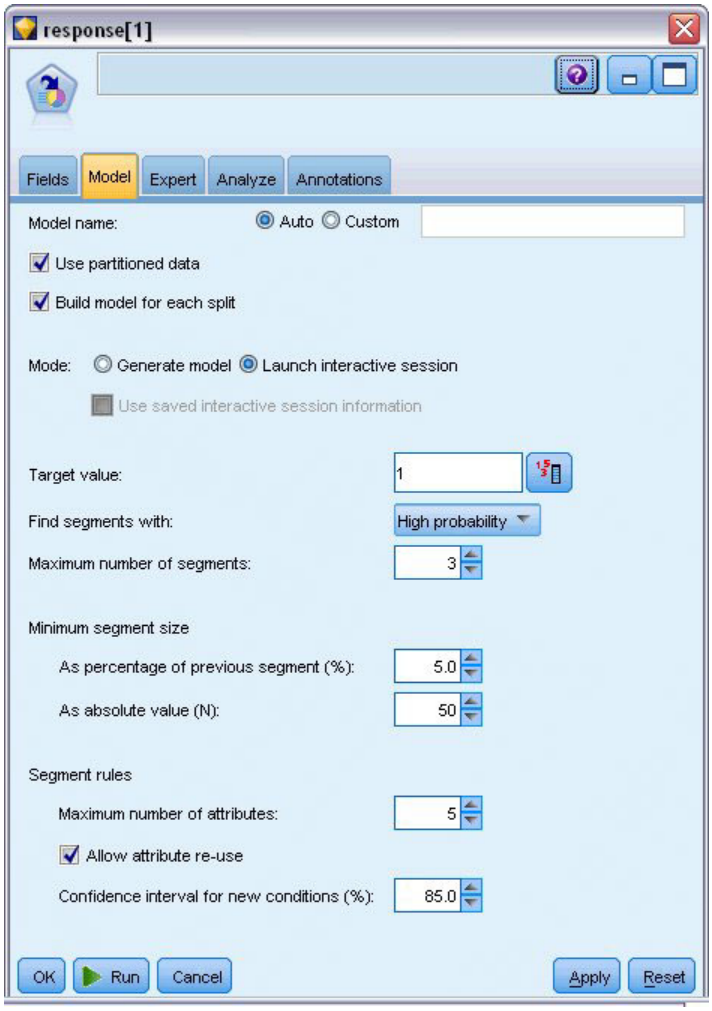


図 120. ディシジョン・リスト・ノードの「モデル」タブ

2. 「**インタラクティブ・セッションの起動**」を選択します。
3. この例のモデルをシンプルに保つために、セグメントの最大数を 3 に設定します。
4. 新しい条件の信頼区間を 85% に変更します。
5. 「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。

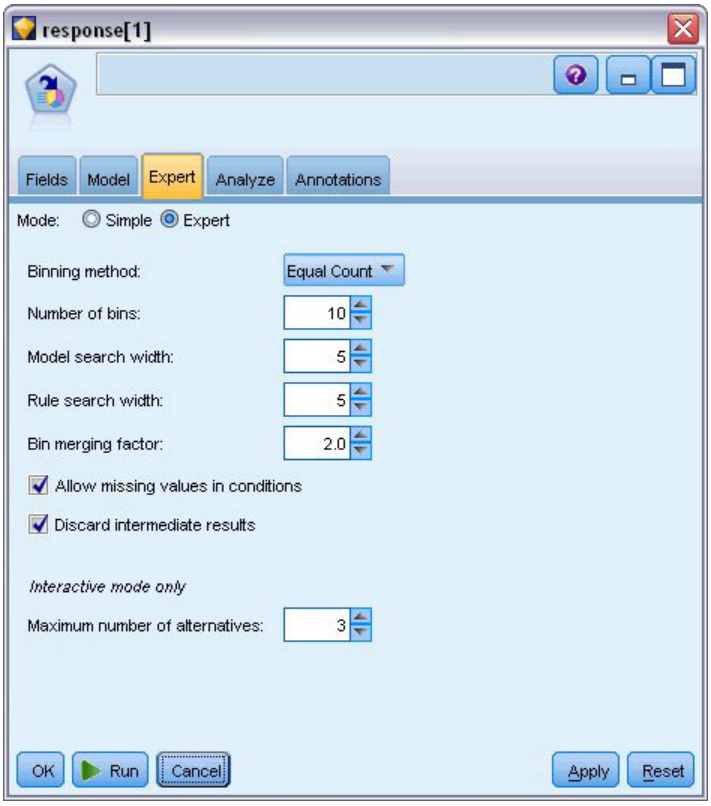


図 121. ディシジョン・リスト・ノードの「エキスパート」タブ

6. 「**代替の最大数**」を 3 に増やします。このオプションは、「モデル」タブで選択した「**インタラクティブ・セッションの起動**」設定と連携して機能します。
7. 「**実行**」をクリックして、Interactive List Viewer を表示します。

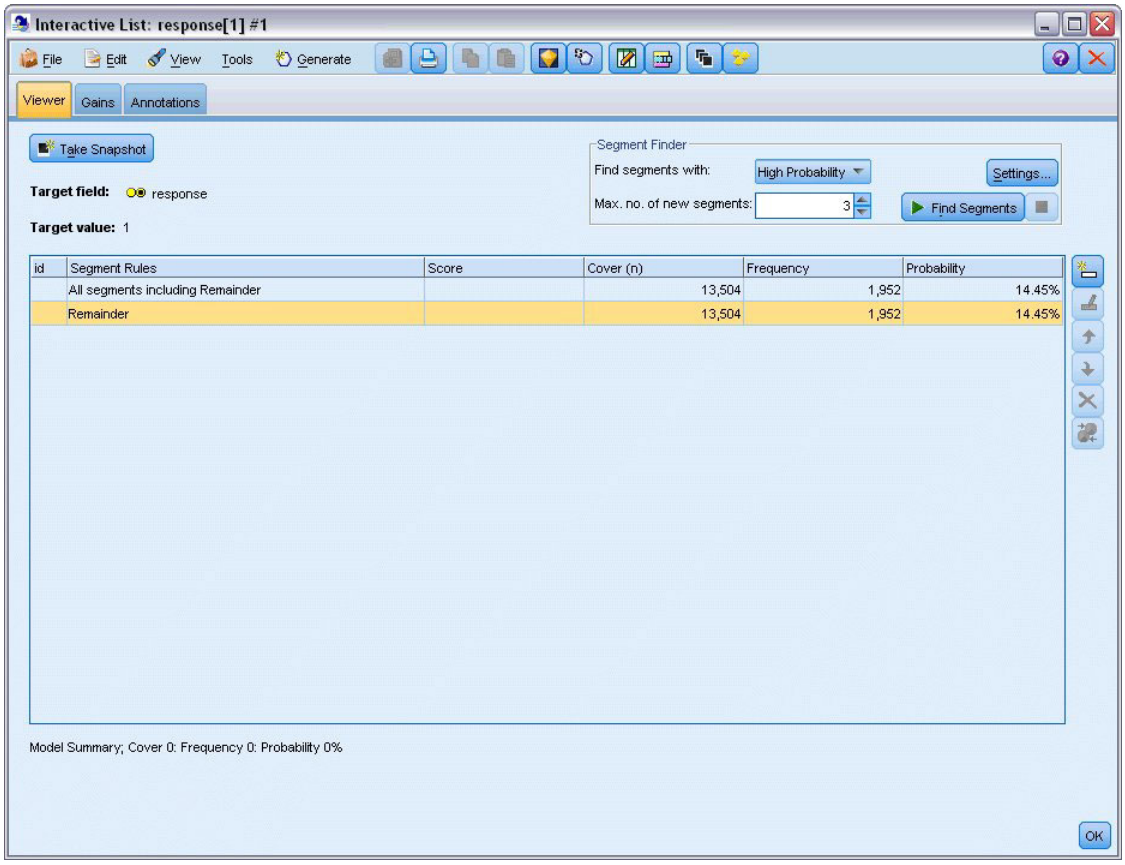


図 122. Interactive List Viewer

セグメントがまだ定義されていないため、すべてのレコードが残余に分類されます。サンプルの中の 13,504 件のレコードの中から、全体のヒット率 14.45% の 1,952 件が「はい」になっています。この割合を向上させるには、好意的なレスポンスをしそうな (またはしそうでない) 顧客のセグメントを特定します。

8. Interactive List Viewer で、メニューから次の項目を選択します。

「ツール」 > 「セグメントの検索」

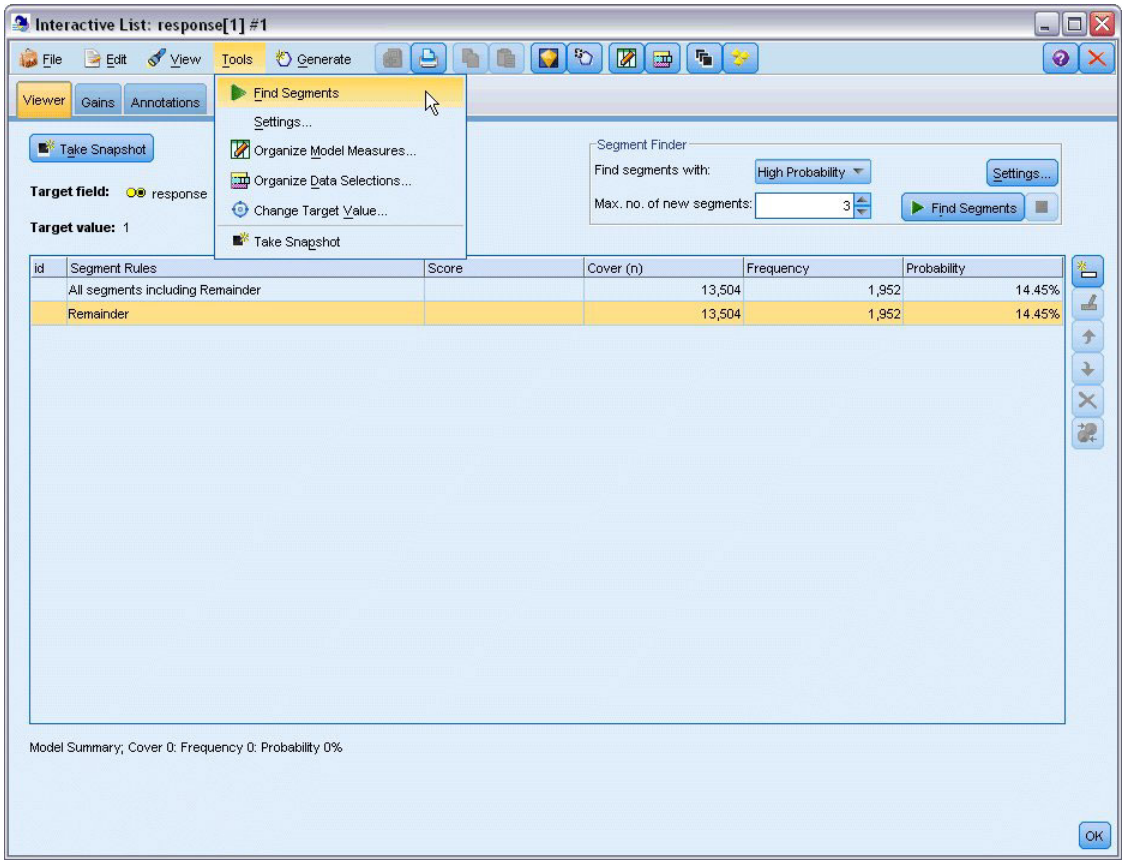


図 123. Interactive List Viewer

これにより、ディシジョン・リスト・ノードで指定した設定に基づいて、デフォルトのマイニング・タスクが実行されます。完了したタスクは 3 つの代替モデルを返し、それらのモデルは、「モデル・アルバム」ダイアログ・ボックスの「代替」タブにリストされます。

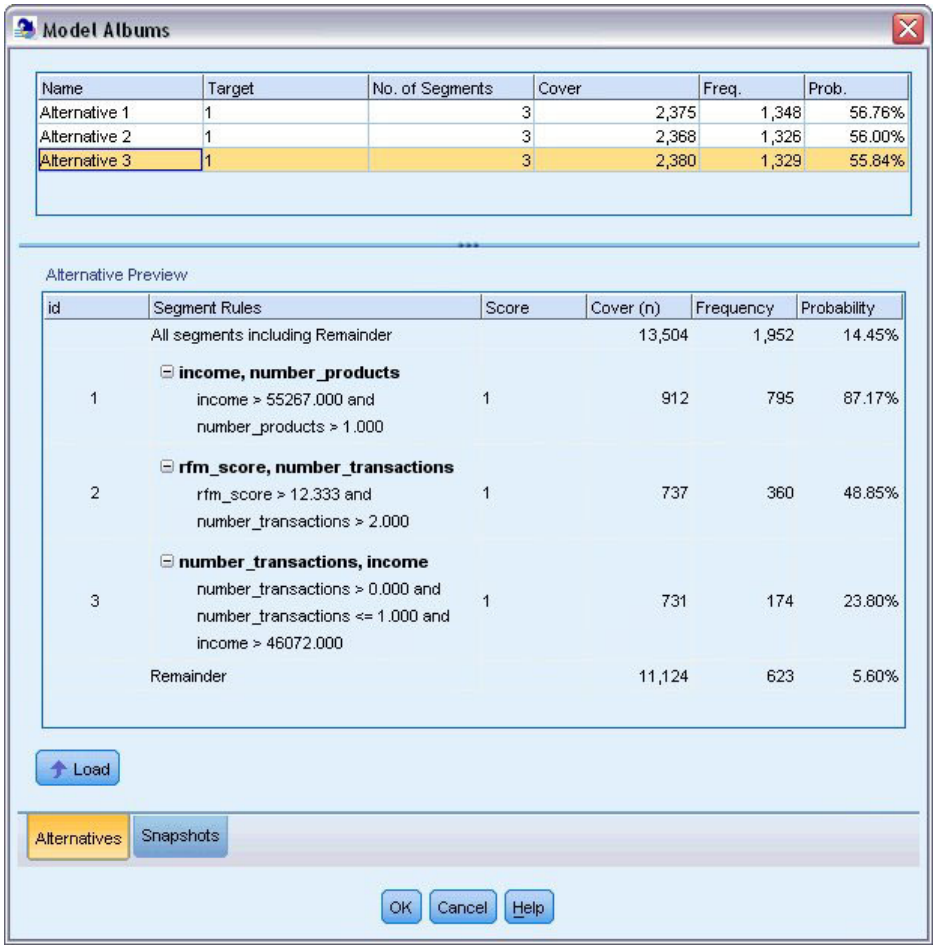


図 124. 使用可能な代替モデル

9. リストから最初の代替モデルを選択します。その詳細は、「代替プレビュー」パネルに表示されます。

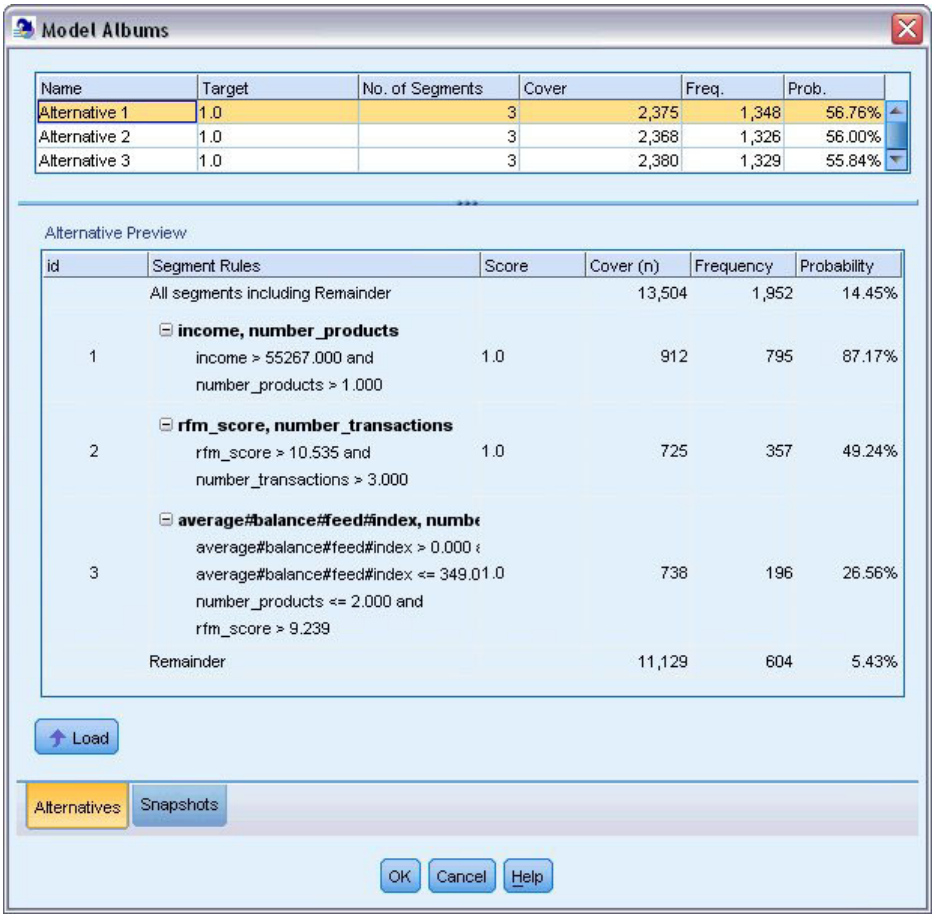


図 125. 選択された代替モデル

「代替プレビュー」パネルを使用して、作業モデルを変更することなく、任意の数の代替モデルを迅速に表示でき、さまざまなアプローチで容易に実験することができます。

注: モデルを見やすくするために、ここで示しているように、ダイアログ内の「代替プレビュー」パネルを最大化できます。これを行うには、パネルの境界線をドラッグします。

収入、月ごとの取引数、RFM スコアなどの予測値に基づいたルールを使用し、モデルはサンプル全体よりもはるかに高い回答率のセグメントを識別します。セグメントが結合されると、このモデルはヒット率が 56.76% まで向上したと示唆します。しかし、このモデルは、サンプル全体の小さな部分をカバーするに過ぎないため、数百件のヒットがある 11,000 件を上回るレコードが残余に分類されたままになります。成績の悪いセグメントを排除しながらこれらのヒットのより多くを捕捉するモデルが必要です。

10. 異なるモデル作成の方法を試すには、メニューから次の項目を選択します。

「ツール」 > **「設定」**

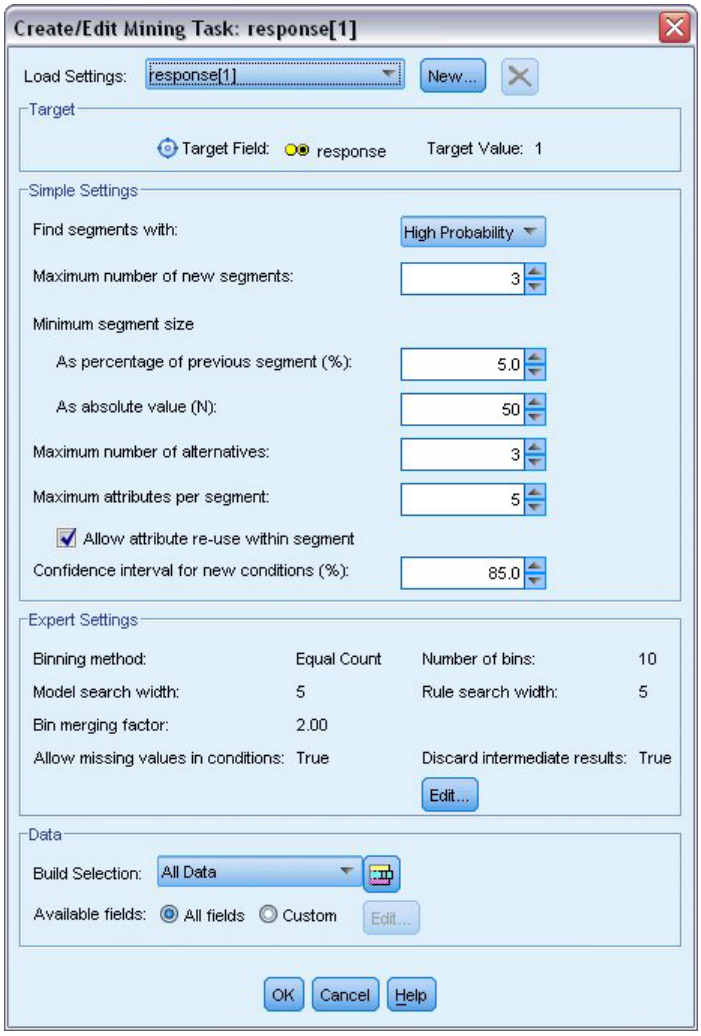


図 126. 「マイニング・タスクの作成/編集」ダイアログ・ボックス

11. 「**新規**」ボタン (右上隅にあります) をクリックして 2 番目のマイニング・タスクを作成し、「新規設定」ダイアログ・ボックスでタスク名として「*下方検索*」を指定します。

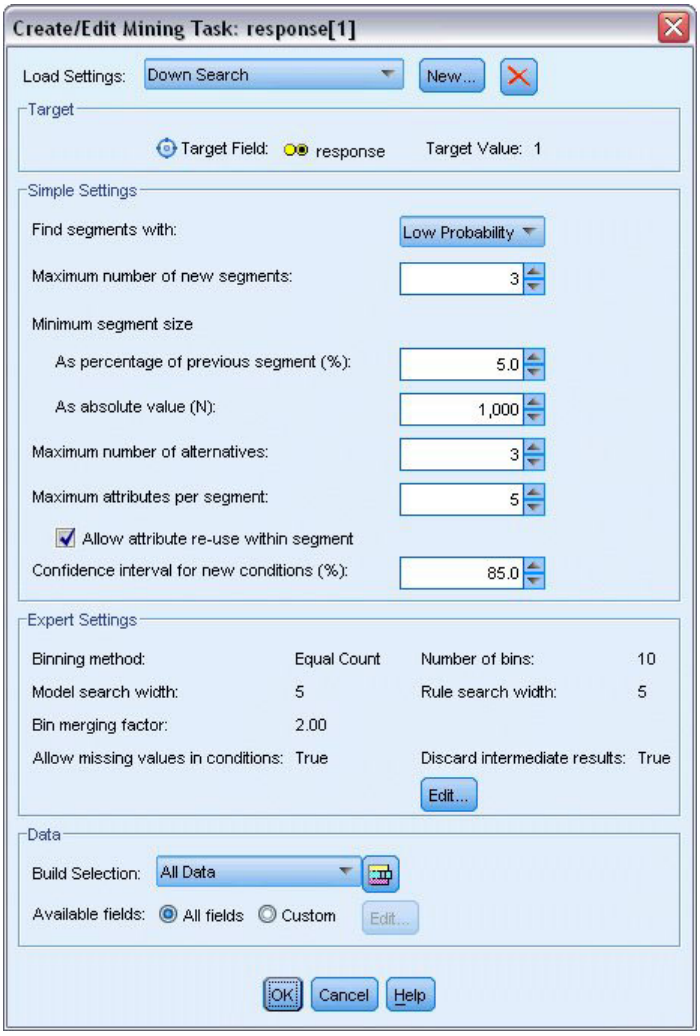


図 127. 「マイニング・タスクの作成/編集」ダイアログ・ボックス

12. タスクの検索方向を「**低い確率**」に変更します。こうすると、アルゴリズムは、最大ではなく最小 レスポンス率のセグメントを検索します。
13. 最小セグメント・サイズを 1,000 に増やします。「**OK**」をクリックして、Interactive List Viewer に戻ります。
14. Interactive List Viewer で、「セグメント・ファインダー」パネルに新しいタスクの詳細が表示されていることを確認し、「**セグメントの検索**」をクリックします。

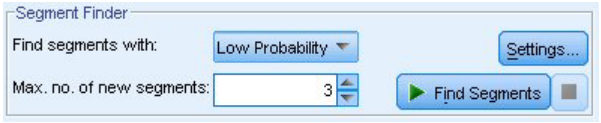


図 128. 新しいマイニング・タスクのセグメントの検索

タスクによって新しい代替モデルのセットが返され、それらは「モデル・アルバム」ダイアログ・ボックスの「代替」タブに表示され、以前の結果と同じ方法でプレビューできます。

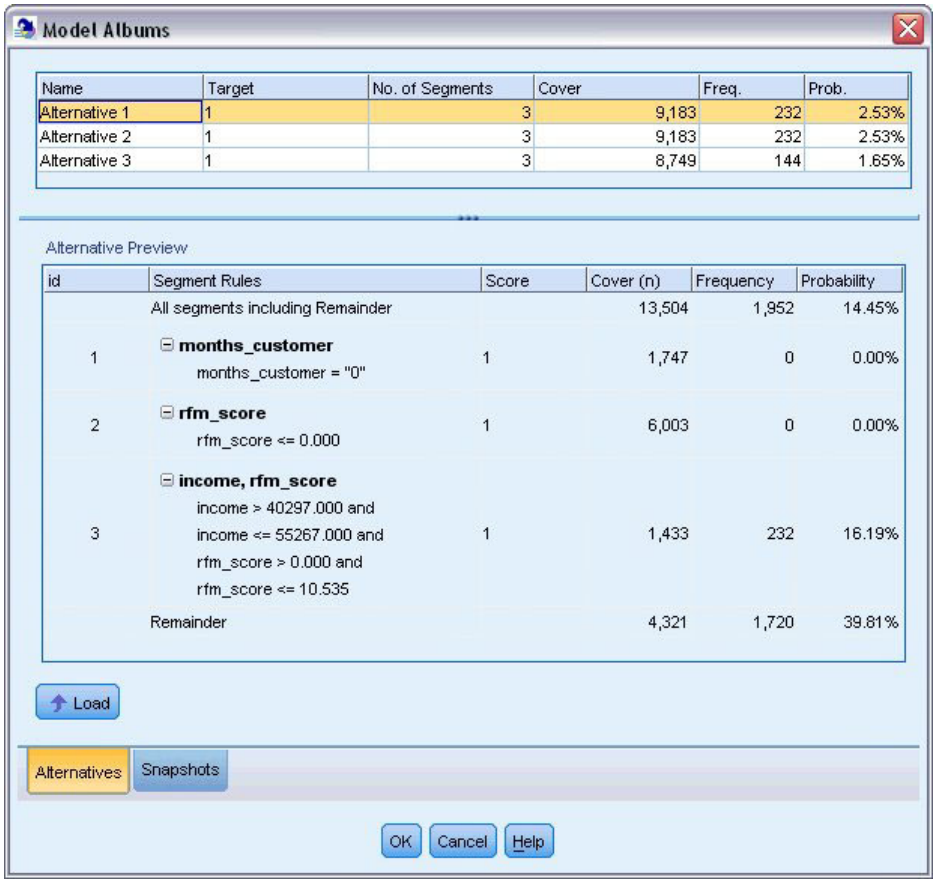


図 129. 「下方向検索」モデルの結果

今回は、各モデルで、高いレスポンス確率ではなく低いレスポンス確率のセグメントが特定されました。最初の代替モデルを調べると、これらのセグメントを単に除外するだけで、残りの部分のヒット率が 39.81% に増加します。これは、以前に調べたモデルより低くなっていますが、カバー率はより高くなっています (つまり合計ヒットが多い)。

2 つのアプローチを結合する (低い確率検索を使用して重要度の低いレコードを除外し、その後に高い確率の検索を使用する) ことで、この結果を改善させることができる可能性があります。

15. 「**ロード**」をクリックして、これ (最初の下方向検索代替モデル) を作業モデルにし、「**OK**」をクリックして、「モデル・アルバム」ダイアログ・ボックスを閉じます。

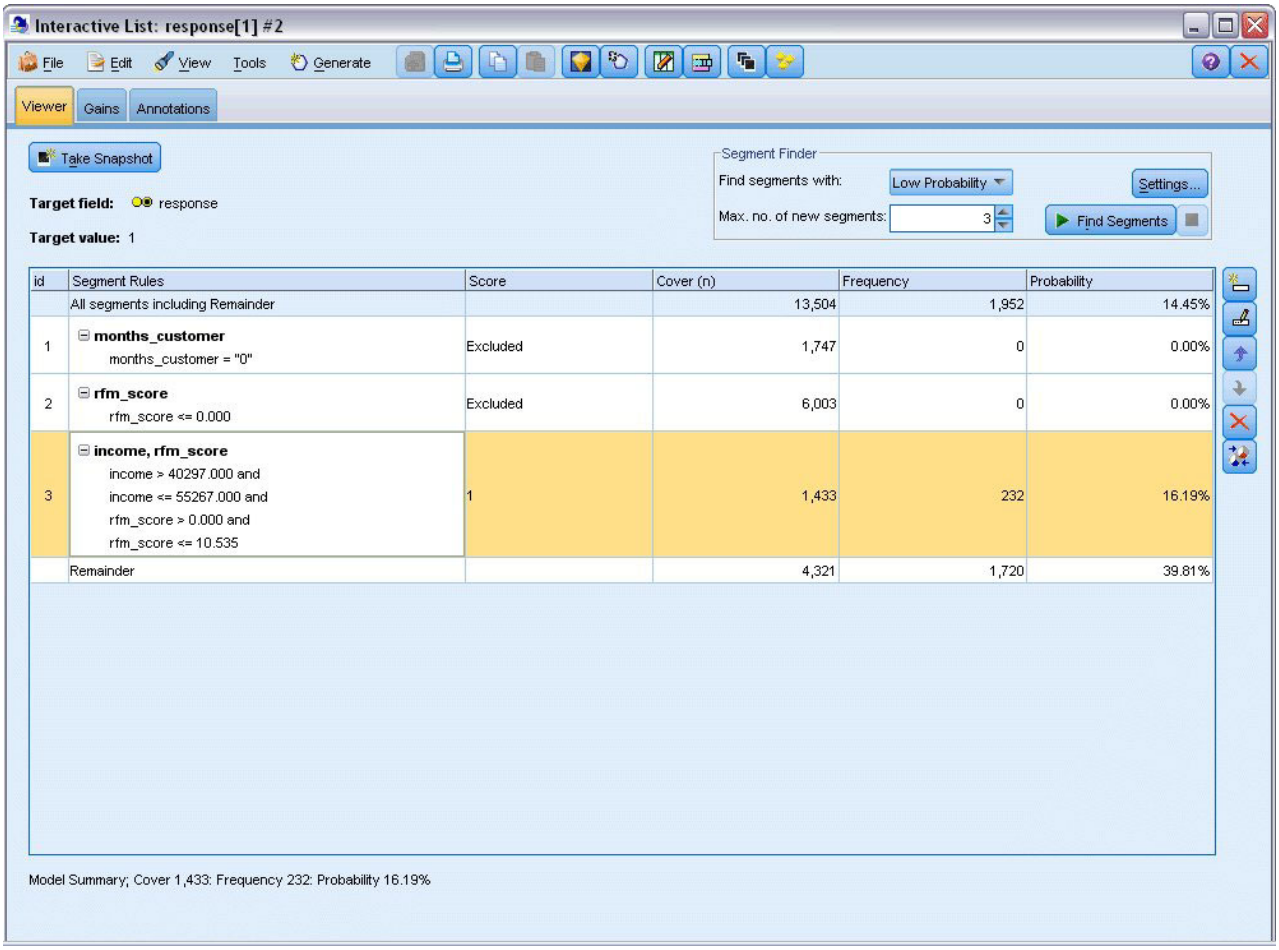


図 130. セグメントの除外

16. 最初の 2 つのセグメントのそれぞれを右クリックし、「**セグメントの除外**」を選択します。これらのセグメントでは、合わせて、ヒットがゼロの 8,000 件近くのレコードが取得されるため、将来のオファーからそれらのセグメントを除外するのは理に適っています。(除外されたセグメントは、それを示すために Null とスコアリングされます)。
17. 3 番目のセグメントを右クリックして、「**セグメントの削除**」を選択します。16.19% という、このセグメントのヒット率は、14.45% の基準率と差がないために、これを留めておくことを正当化するに足る情報は追加されません。

    注: セグメントの削除は、セグメントの除外と同じではありません。セグメントの除外では、単にそのスコアリング方法が変更されるだけですが、セグメントの削除では、そのセグメントはモデルから完全に削除されます。

    成績が最低のセグメントを除外したため、残りで成績の良いセグメントを検索できるようになりました。
18. 次のマイニング・タスクが残りの部分にだけ適用されるように、テーブルの残りの行をクリックして選択します。

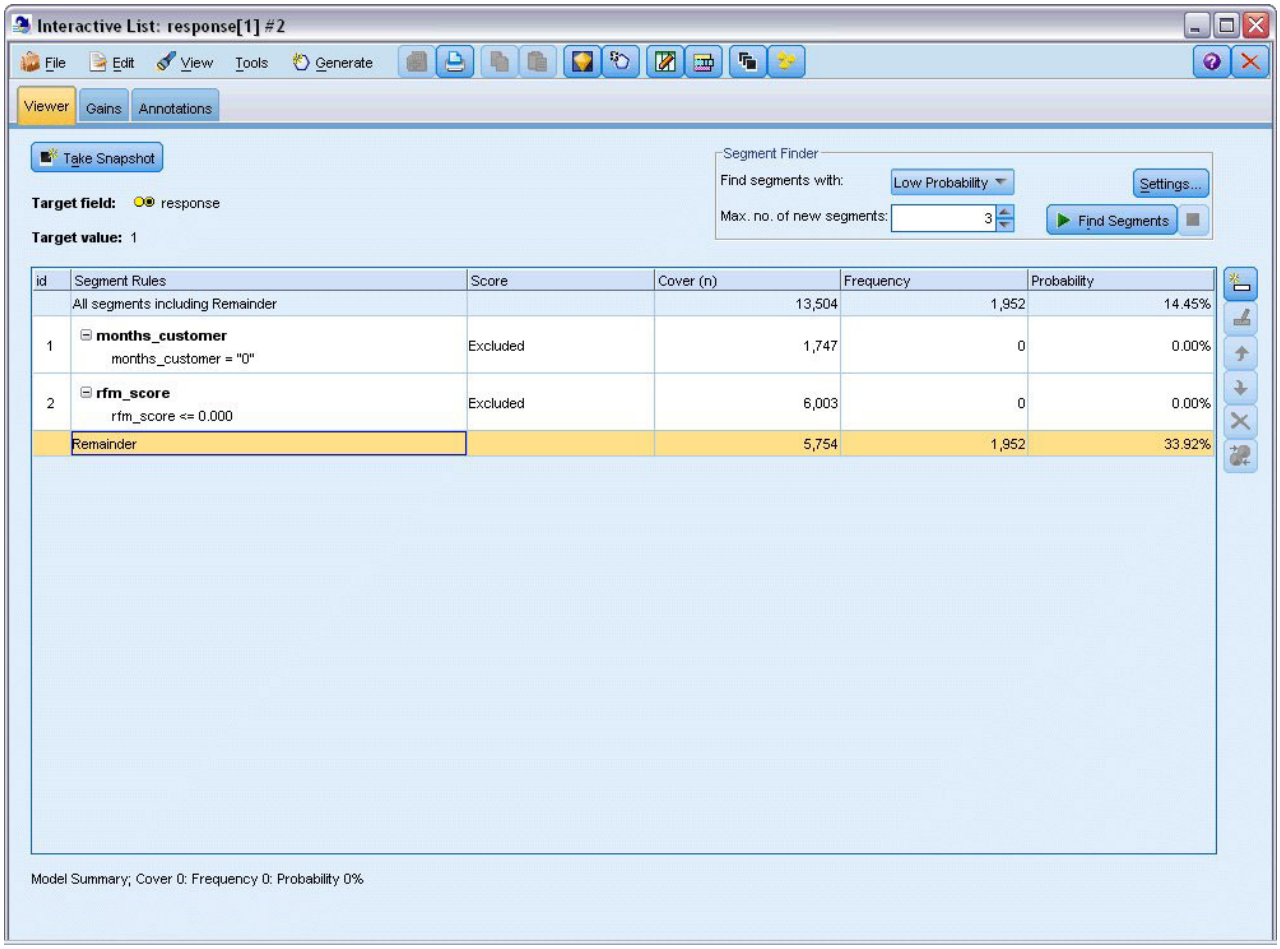


図 131. セグメントの選択

19. 残りの部分を選択し、「**設定**」をクリックして「マイニング・タスクの作成/編集」ダイアログ・ボックスを再度開きます。
20. 最上部の「**ロード設定**」で、デフォルトのマイニング・タスク **response[1]** を選択します。
21. 「**シンプル設定**」を編集して、新しいセグメント数を 5 に増やし、最小セグメント・サイズを 500 に増やします。
22. 「**OK**」をクリックして、Interactive List Viewer に戻ります。

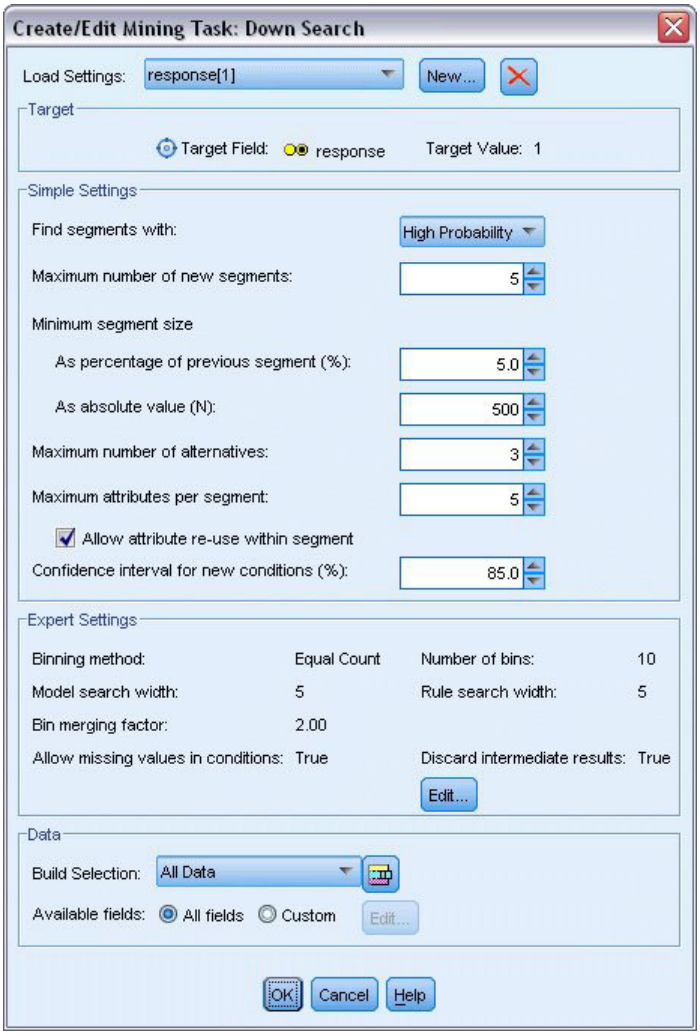


図 132. デフォルトのマイニング・タスクの選択

23. 「**セグメントの検索**」をクリックします。

これにより、また別の代替モデルのセットが表示されます。1 つのマイニング・タスクの結果を他の結果にフィードすることで、これらの最新のモデルに成績の高いセグメントと成績の低いセグメントが混じり合うようになります。回答率が低いセグメントは除外され、Null としてスコアリングされます。一方、含まれるセグメントは 1 としてスコアリングされます。全体の統計はこれらの除外を反映し、最初の代替モデルはヒット率 45.63% を示し、前のどのモデルよりも高いカバー率 (3,456 レコードのうち 1,577 ヒット) となっています。

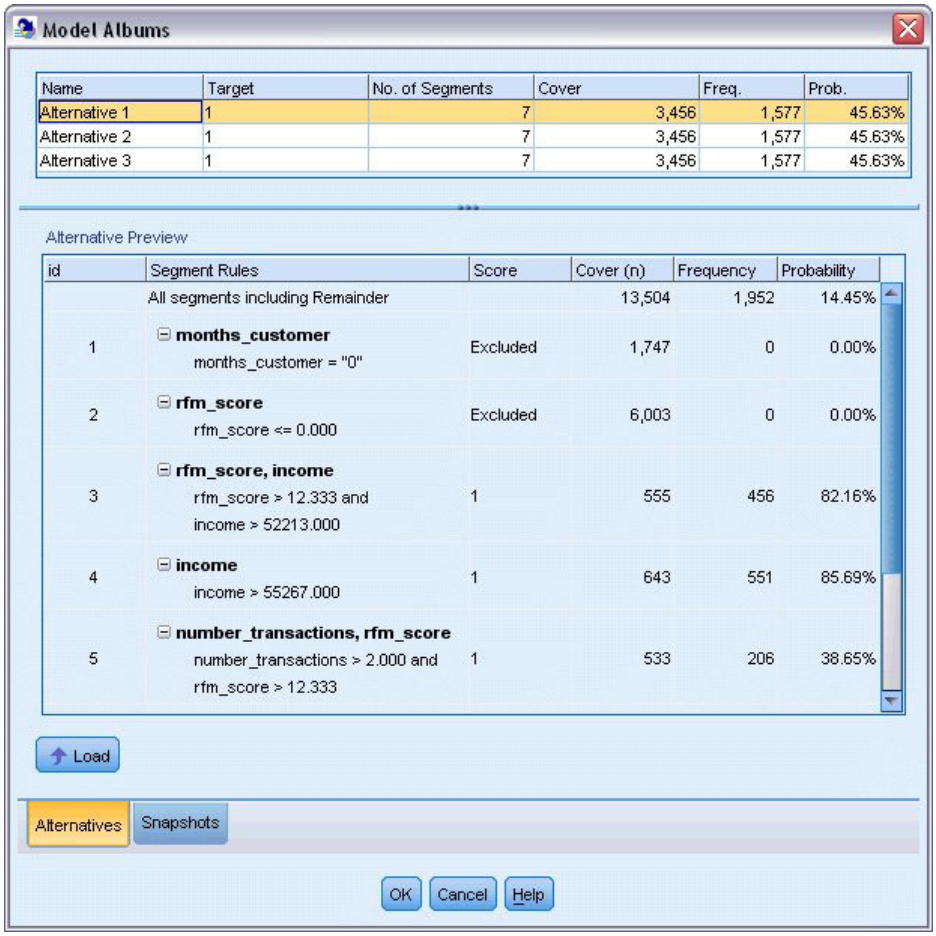


図 133. 結合されたモデルの代替

24. 最初の代替モデルをプレビューしてから、「**ロード**」をクリックし、そのモデルを作業モデルにします。

Excel を使用したカスタム指標の計算

1. 実践的観点からモデルがどのように実行されているかについて、より詳細に理解するには、ツール・メニューから「**モデル指標の編成**」を選択します。

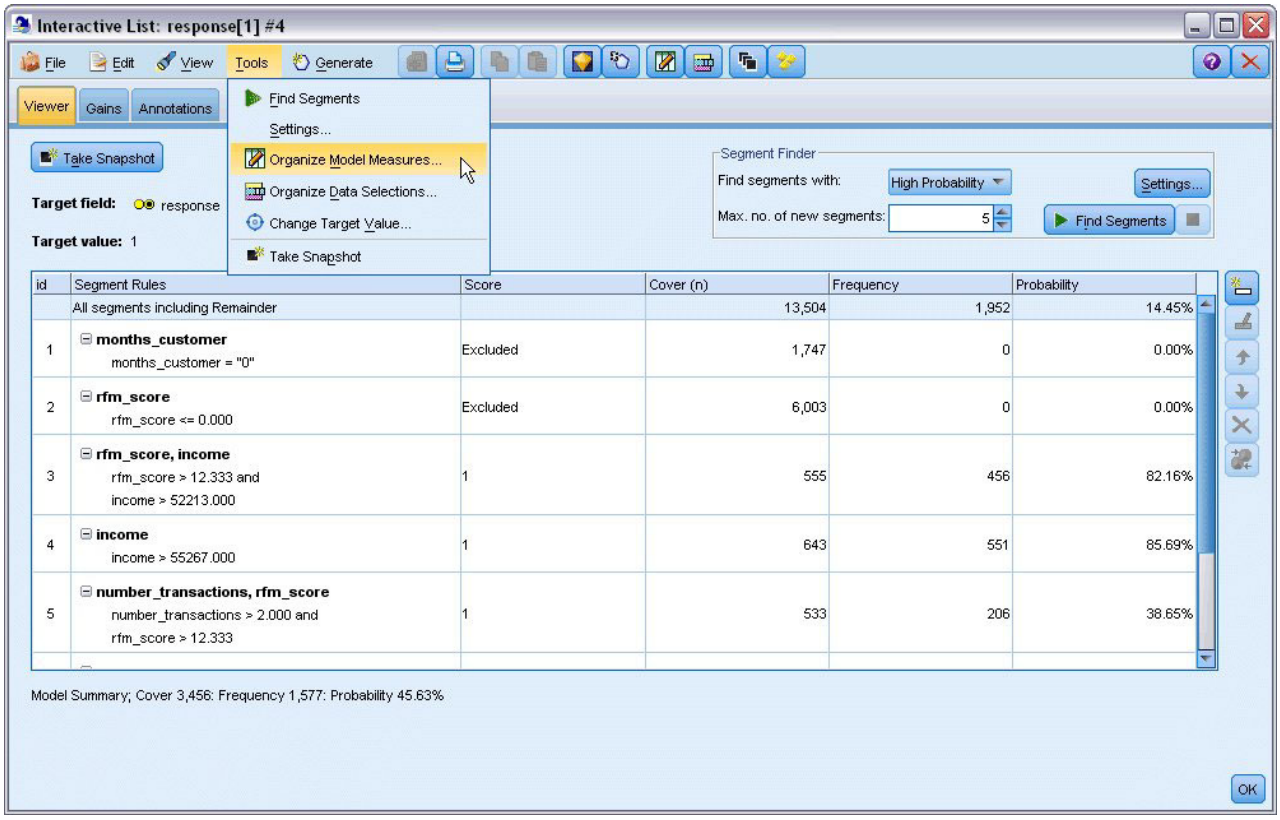


図 134. モデル指標の編成

「モデル指標の編成」ダイアログ・ボックスで、指標 (または列) を選択し、Interactive List Viewer で表示することができます。また、すべてのレコードまたは選択したサブセットに対して指標を計算するかどうかを指定することができます。さらに状況に応じて、数ではなく円グラフを表示するように選択できます。

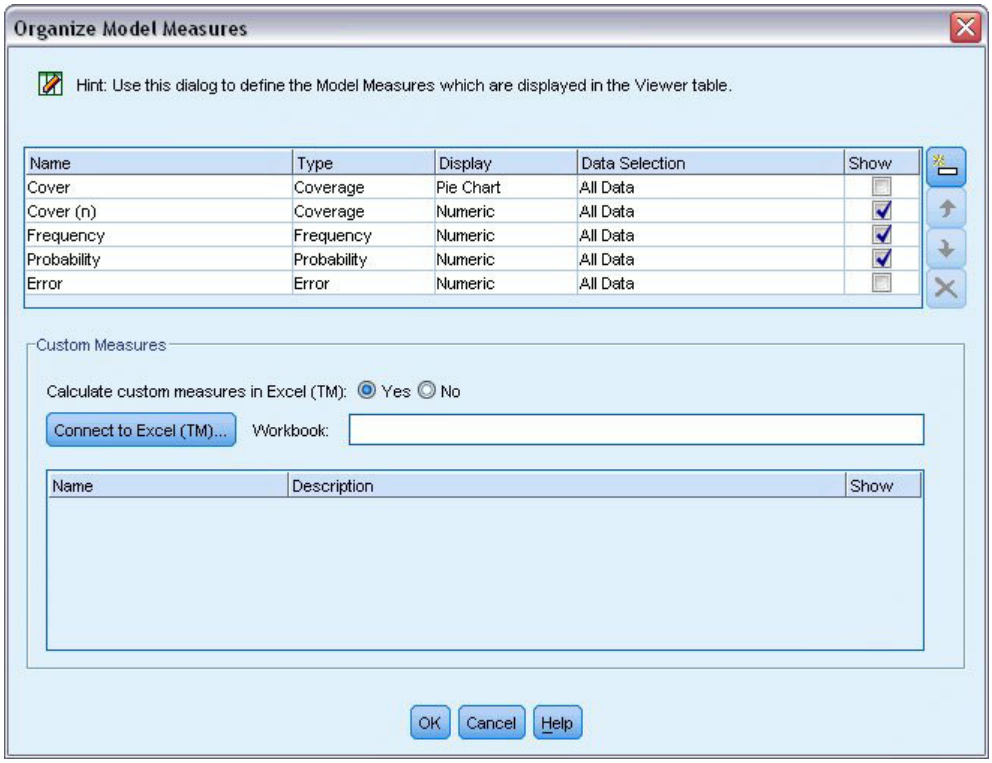


図 135. 「モデル指標の編成」ダイアログ・ボックス

また、Microsoft Excel がインストールされている場合、Excel のテンプレートとリンクして、カスタム指標を計算し、それらをインタラクティブ表示に追加することができます。

2. 「モデル指標の編成」ダイアログ・ボックスで、「**Excel (TM) 内のカスタム指標を計算**」を「**はい**」に設定します。
3. 「**Excel (TM) へ接続**」をクリックします。
4. IBM SPSS Modeler インストール環境の *Demos* フォルダーの *streams* の下にある *template_profit.xlt* ワークブックを選択し、「**開く**」をクリックしてスプレッドシートを開始します。

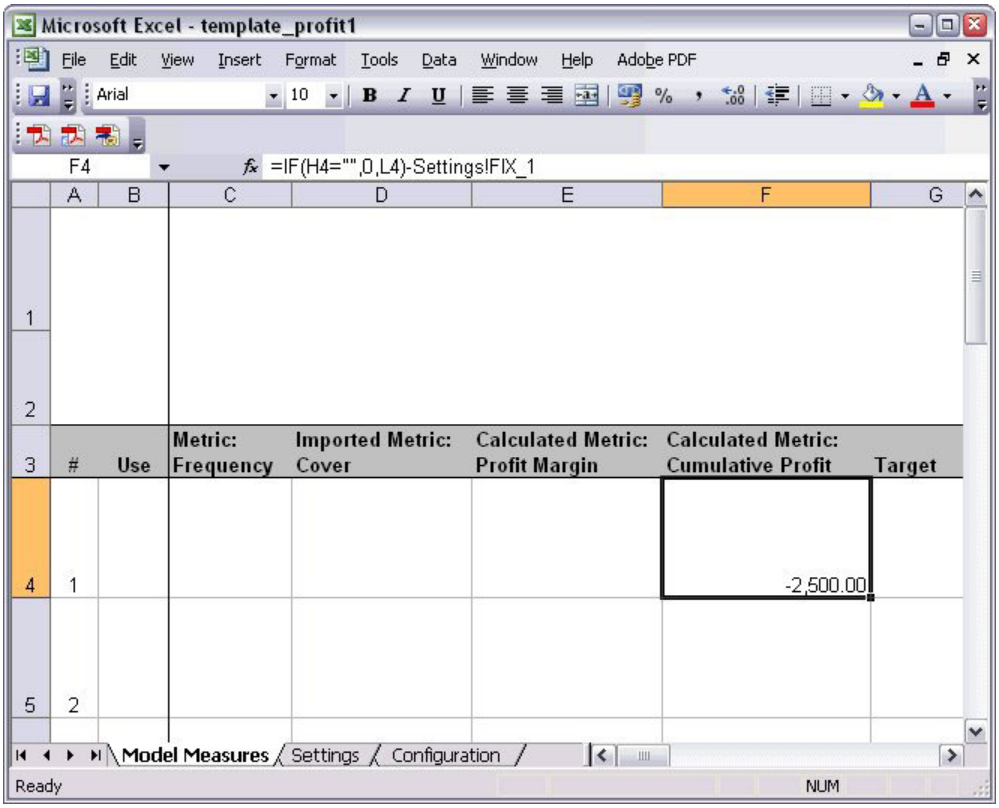


図 136. Excel の「モデル指標」ワークシート

Excel テンプレートには、次の 3 つのワークシートが含まれています。

- 「**モデル指標**」は、モデルからインポートされたモデル指標を表示し、モデルに再びエクスポートするためにカスタム指標を計算します。
- 「**設定**」には、カスタム指標の計算で使用されるパラメーターが含まれています。
- 「**構成**」では、モデルとの間でインポートおよびエクスポートする指標を定義します。

モデルに再びエクスポートされる測定基準は次のとおりです。

- 「**利益幅**」。セグメントからの純利益。
- 「**累積利益**」。キャンペーンの総利益。

次の式で定義されています。

```
Profit Margin = Frequency * Revenue per respondent - Cover * Variable cost

Cumulative Profit = Total Profit Margin - Fixed cost
```

なお、度数およびカバーはモデルからインポートされます。

コストおよび利益パラメーターは、「設定」ワークシートでユーザーによって指定されます。

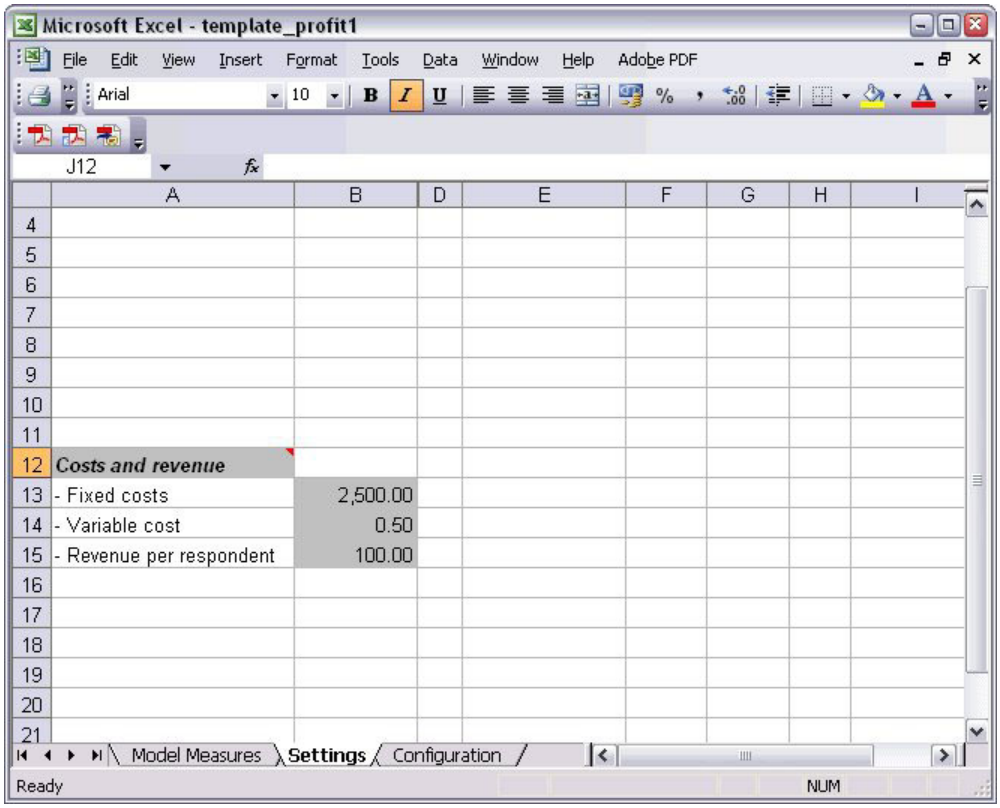


図 137. Excel の「設定」ワークシート

「**固定コスト**」は、設計や計画など、キャンペーンに設定されたコストです。

「**変動コスト**」は、封筒や切手など、各顧客にオファーを展開するコストです。

「**回答者あたりの利益**」は、オファーに回答した顧客からの純利益です。

5. モデルへのリンクを再度実行するには、Windows タスクバーを使用して (または Alt キーと Tab キーを押して) Interactive List Viewer に再度移動します。

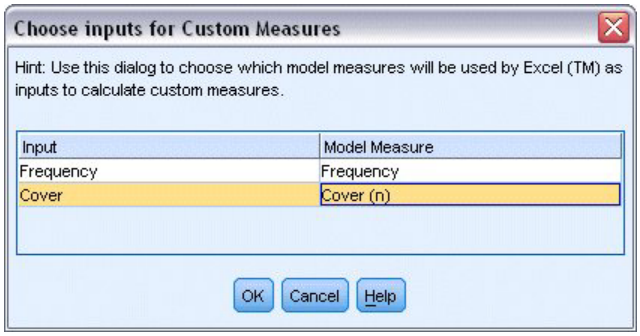


図 138. カスタム指標のための入力の選択

「カスタム指標のための選択入力」ダイアログ・ボックスが表示され、モデルからの入力をテンプレートに定義された特定のパラメーターにマッピングすることができます。左側の列では利用可能な指標がリストされています。右側の列では、「構成」ワークシートで定義されたスプレッドシート・パラメーターにこれらの指標をマッピングします。

6. 「**モデル指標**」列で、それぞれの入力に対して「**度数**」および 「**カバー (n)**」を選択し、「**OK**」をクリックします。

   この場合、テンプレート内のパラメーター名「度数」および「カバー (n)」はたまたま入力に一致していますが、異なる名前を使用することもできます。

7. 「モデル指標の編成」ダイアログ・ボックスで「**OK**」をクリックし、Interactive List Viewer を更新します。

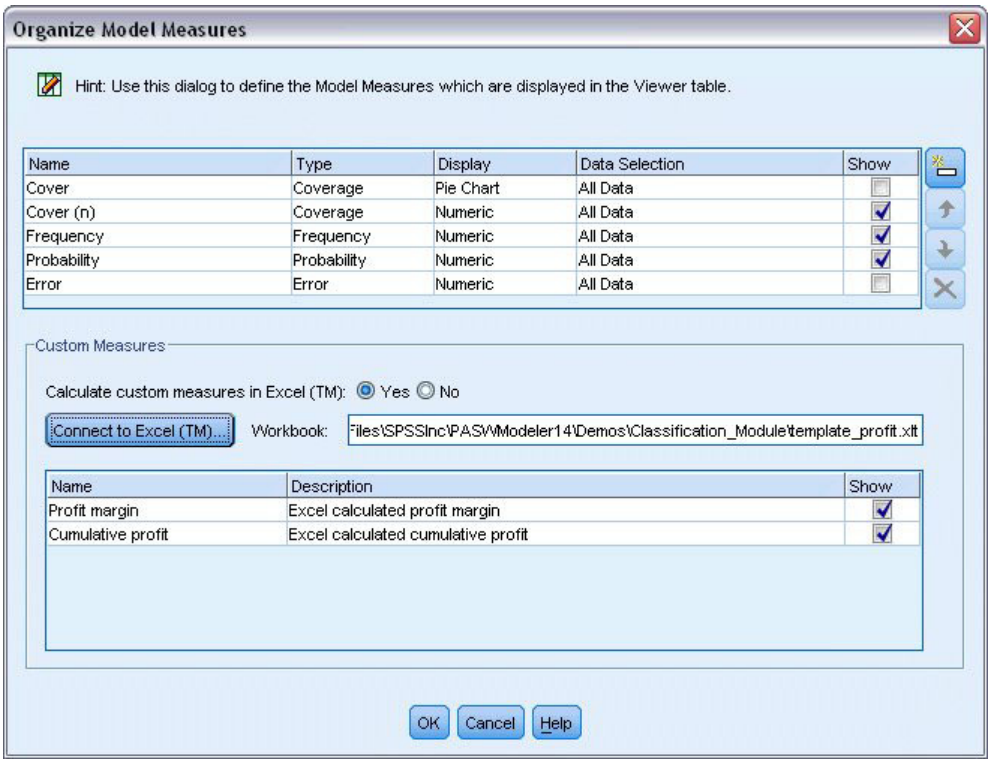


図 139. Excel のカスタム指標を表示する「モデル指標の編成」ダイアログ・ボックス

ウィンドウに新しい指標が新しい列として追加され、モデルが更新されるたびに再計算されます。

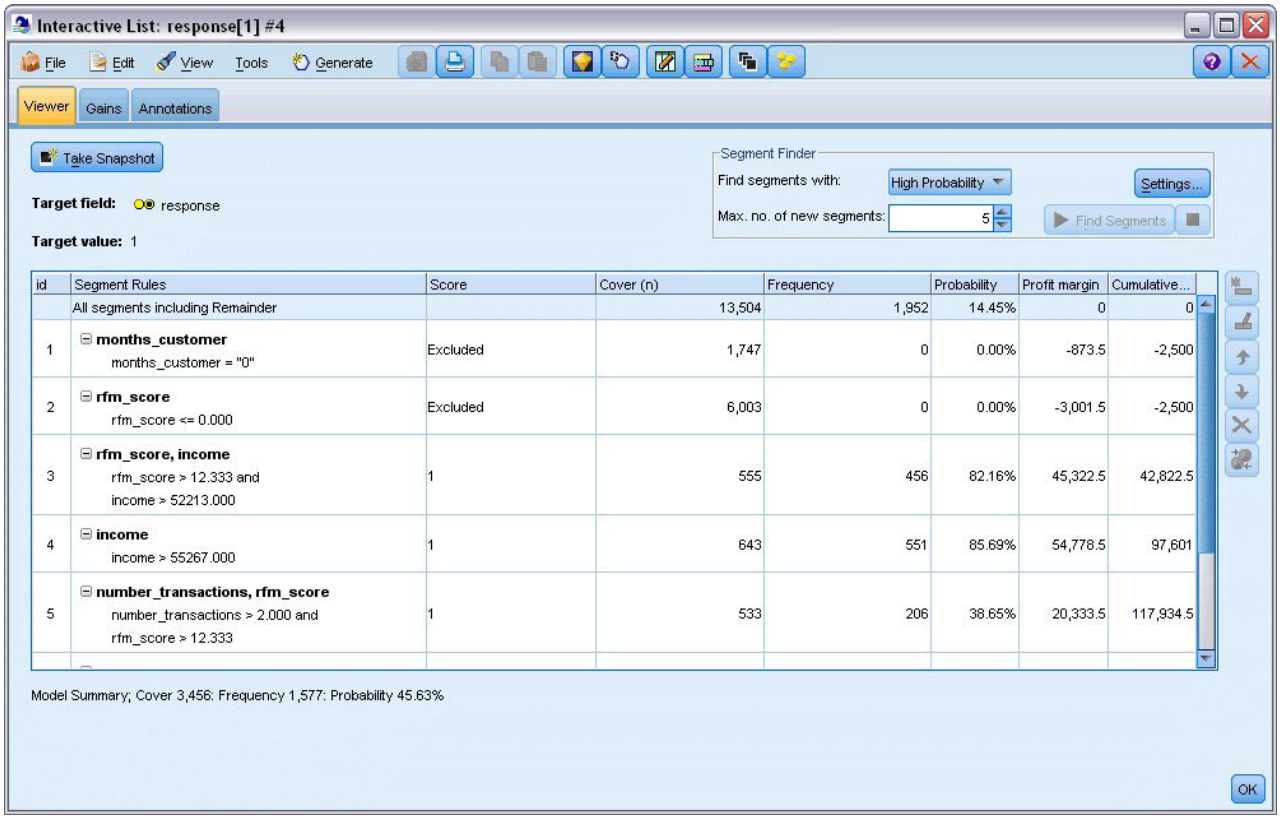


図 140. Interactive List Viewer に表示される Excel のカスタム指標

Excel テンプレートを編集することで、任意の数のカスタム指標を作成できます。

Excel テンプレートの変更

IBM SPSS Modeler では、Interactive List Viewer で使用するためのデフォルトの Excel テンプレートが提供されていますが、設定を変更したり、独自のテンプレートを追加したりすることもできます。例えば、テンプレートのコストが、組織にとって不適切で、修正の必要が生じることがあります。

注: 既存のテンプレートを変更した場合、または独自のテンプレートを作成した場合、必ず、Excel 2003 の拡張子 *.xlt* を使用してファイルを保存してください。

デフォルトのテンプレートを新しいコストおよび収益の詳細で変更し、新しい数値で Interactive List Viewer を更新する手順は、次のとおりです。

1. Interactive List Viewer で、「ツール」メニューから**「モデル指標の編成」**を選択します。
2. 「モデル指標の編成」ダイアログ・ボックスで、**「Excel™ への接続」**をクリックします。
3. *template_profit.xlt* ワークブックを選択し、**「開く」**をクリックしてスプレッドシートを起動します。
4. 「設定」ワークシートを選択します。
5. **「固定コスト」**を 3,250.00 に、**「回答者あたりの利益」**を 150.00 に編集します。

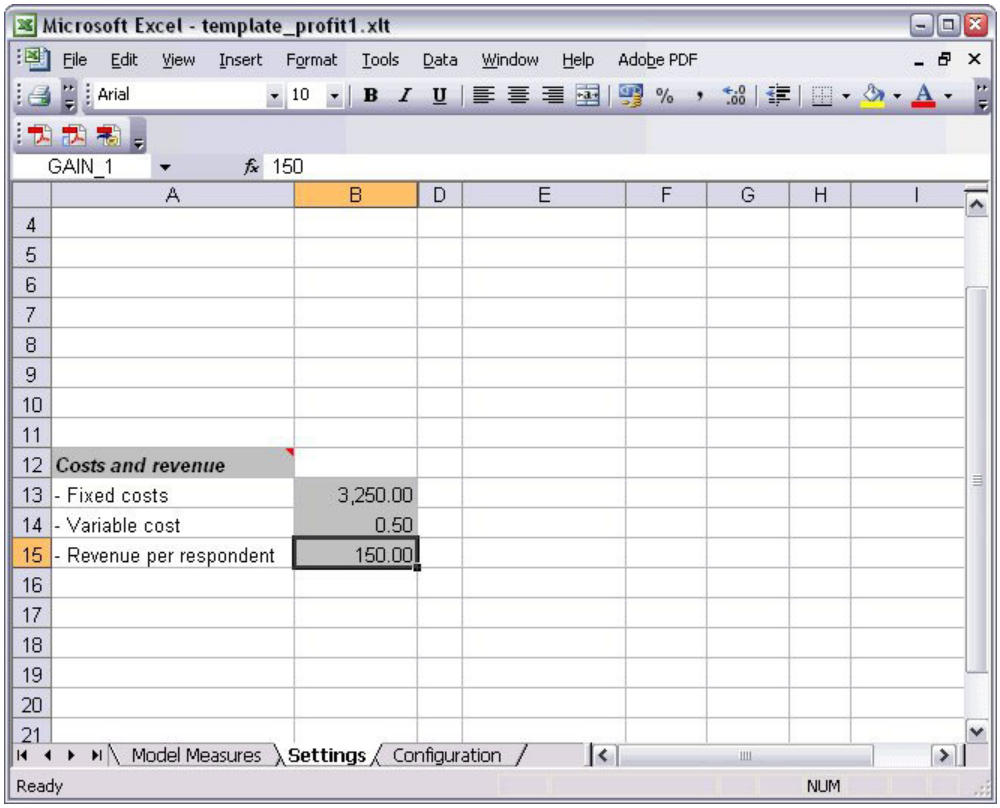


図 141. Excel の「設定」ワークシートの変更された値

6. 一意の関連するファイル名で、変更済みテンプレートを保存します。ファイルの拡張子が Excel 2003 の拡張子 *.xlt* であることを確認してください。

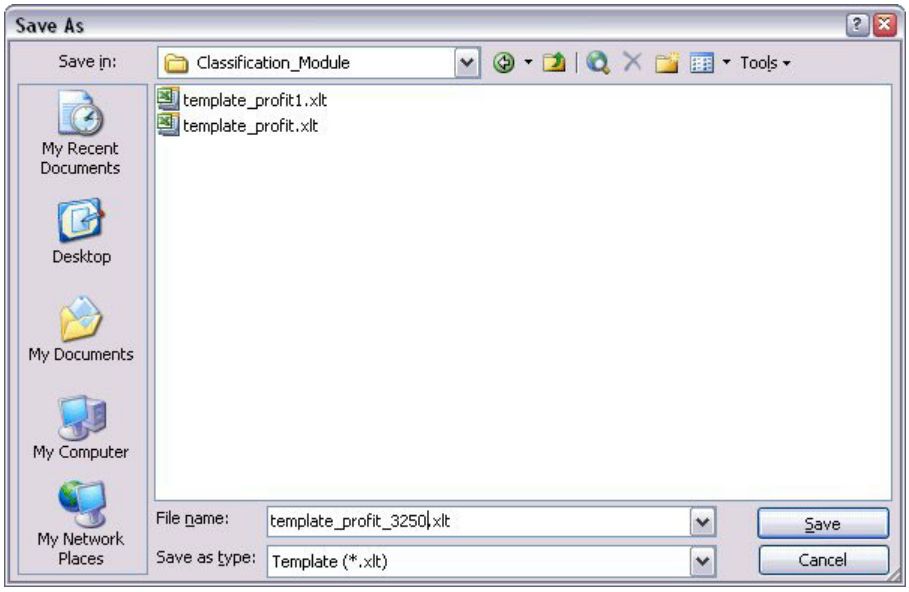


図 142. 変更された Excel テンプレートの保存

7. Windows タスクバーを使用して (または Alt + Tab キーを押して)、Interactive List Viewer に戻ります。

「カスタム指標のための入力の選択」ダイアログ・ボックスで、表示する指標を選択し、「**OK**」をクリックします。

8. 「モデル指標の編成」ダイアログ・ボックスで「**OK**」をクリックし、Interactive List Viewer を更新します。

言うまでもありませんが、この例では Excel テンプレートを変更するシンプルな 1 つの方法のみ示しています。Interactive List Viewer との間でデータをやり取りしたり、あるいは Excel 内で作業してグラフなどの他の出力を作成したり、より詳細な変更を行うことができます。

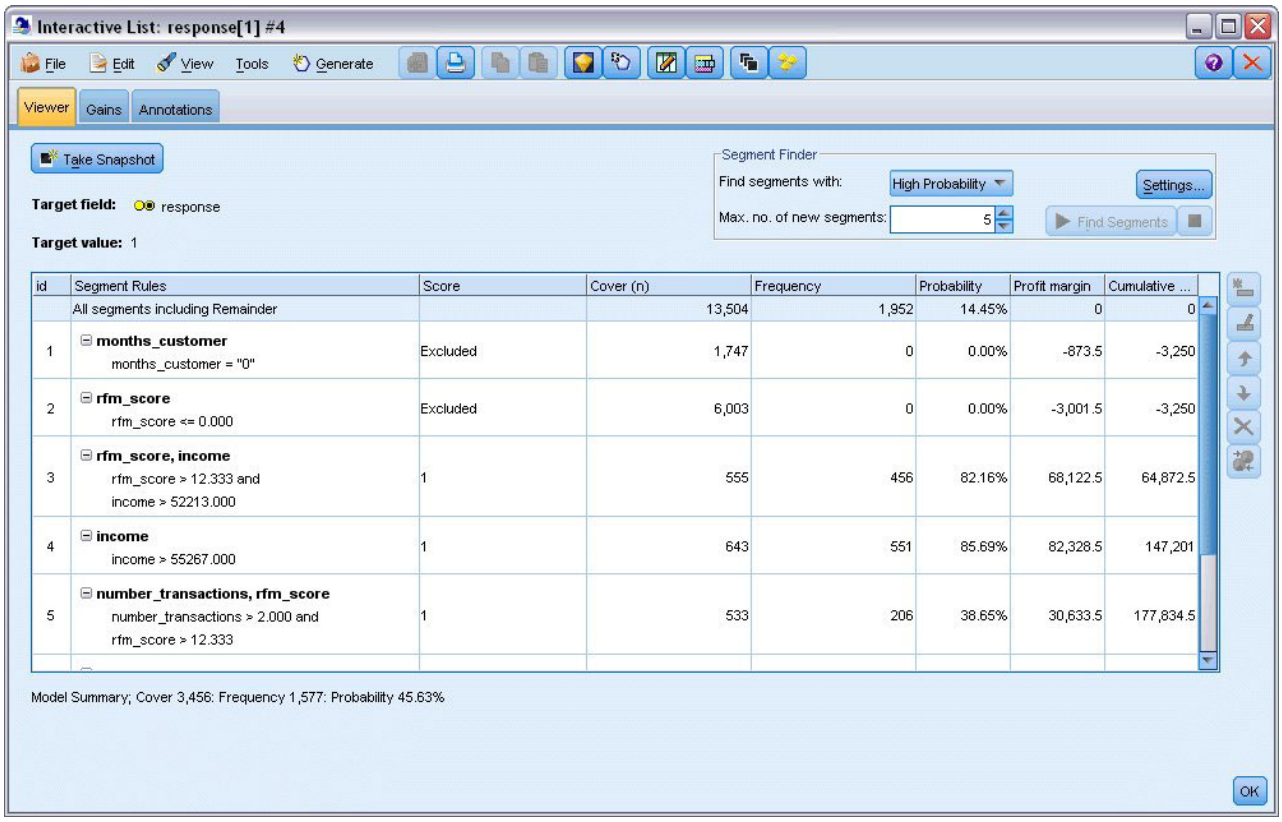


図 143. Interactive List Viewer に表示される、Excel で変更されたカスタム指標

結果の保存

インタラクティブ・セッション中に後から使用するためにモデルを保存するには、モデルのスナップショットを撮ります。スナップショットは、「スナップショット」タブにリストされます。インタラクティブ・セッション中は、いつでも保存したスナップショットに戻ることができます。

この方法を続行することで、追加のマイニング・タスクを実験して、追加のセグメントを検索することができます。また、既存のセグメントを編集し、独自のビジネス・ルールに基づいてカスタム・セグメントを挿入し、データ選択を作成して特定のグループのためにモデルを最適化し、その他のさまざまな方法でモデルをカスタマイズできます。最後に、必要に応じて各セグメントを明示的に含めるか除外して、それぞれのスコアリング方法を指定できます。

結果に満足したら、「生成」メニューを使用して、ストリームに追加するかスコアリング用に展開することができるモデルを生成できます。

あるいは、後日使用するため、インタラクティブ・セッションの現在の状態を保存するには、「ファイル」メニューから「**モデル作成ノードの更新**」を選択します。これにより、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、ディシジョン・リストのモデリング・ノードが現在の設定で更新されます。次回ストリームを実行するときは、現在の状態にセッションを回復するために、単にディシジョン・リスト・モデル作成ノードで「**保存済みセッション情報の使用**」が選択されていることを確認してください。

第 12 章 電気通信会社の顧客の分類 (多項ロジスティック回帰)

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。線型回帰と似ていますが、数値型対象フィールドではなくカテゴリー対象フィールドを取ります。

例えば、電気通信プロバイダーがその顧客ベースを、サービス使用パターンによって区分しており、顧客を 4 つのグループにカテゴリー化しているとします。人口統計データを使用して顧客がどのグループに所属するかを予測できれば、個々の見込み客にあわせてサービスをカスタマイズすることができます。

この例では、*telco.sav* というデータ・ファイルを参照する *telco_custcat.str* というストリームを使用します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*telco_custcat.str* ファイルは、*streams* ディレクトリー内にあります。

この例は、使用パターンを予測するための人口統計データの使用方法に注目します。以下のように、対象フィールド *custcat* には、4 つの顧客グループに対応する 4 つの可能な値があります。

値	ラベル
1	基本サービス
2	E-サービス
3	プラス・サービス
4	トータル・サービス

対象に複数のカテゴリーがあるために、多項モデルを使用します。はい/いいえ、真/偽、解約/解約しないなどの 2 つの明確なカテゴリーのある対象の場合は、代わりに 2 項モデルを作成できます。詳細については、141 ページの『第 13 章 電気通信会社の解約 (2 項検定ロジスティック回帰)』を参照してください。

ストリームの構築

1. *Demos* フォルダーの *telco.sav* を指し示す Statistics ファイル入力ノードを追加します。

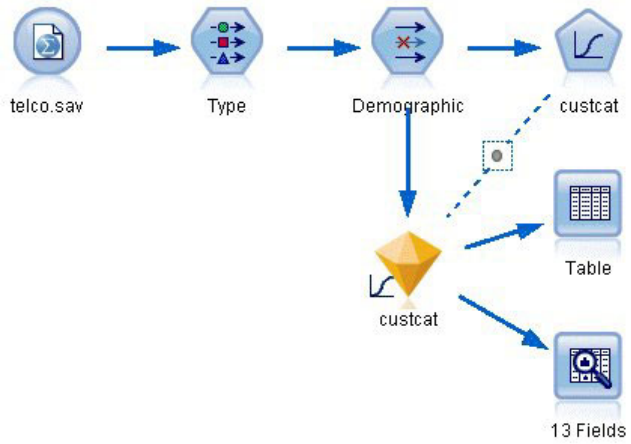


図 144. 多項検定ロジスティック回帰を使用して顧客を分類するためのサンプル・ストリーム

a. データ型ノードを追加し、「**値の読み込み**」をクリックして、すべての測定の尺度が正しく設定されていることを確認します。例えば、値 0 および 1 を持つほとんどのフィールドはフラグと見なすことができます。

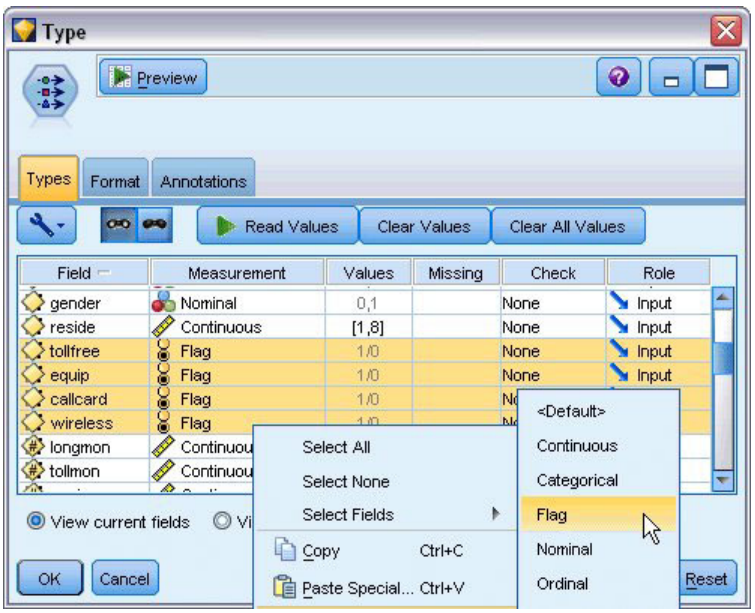


図 145. 複数のフィールドの測定の尺度の設定

ヒント: 類似の値 (0/1 など) を持つ複数のフィールドのプロパティーを変更するには、「*値*」列見出しをクリックしてフィールドを値の順序でソートし、Shift キーを押しながらマウスまたは矢印キーを使用して、変更するフィールドすべてを選択します。選択範囲を右クリックして、測定の尺度を変更するか、選択したフィールドの他の属性を変更します。

「*性別*」は、フラグではなく、2 つの値セットを持つフィールドと見なす方がより適切であるため、「尺度」の値は 「**名義型**」 のままにします。

b. 「*custcat*」フィールドの役割を「**対象**」に設定します。その他のフィールドの役割は、すべて「**入力**」に設定します。

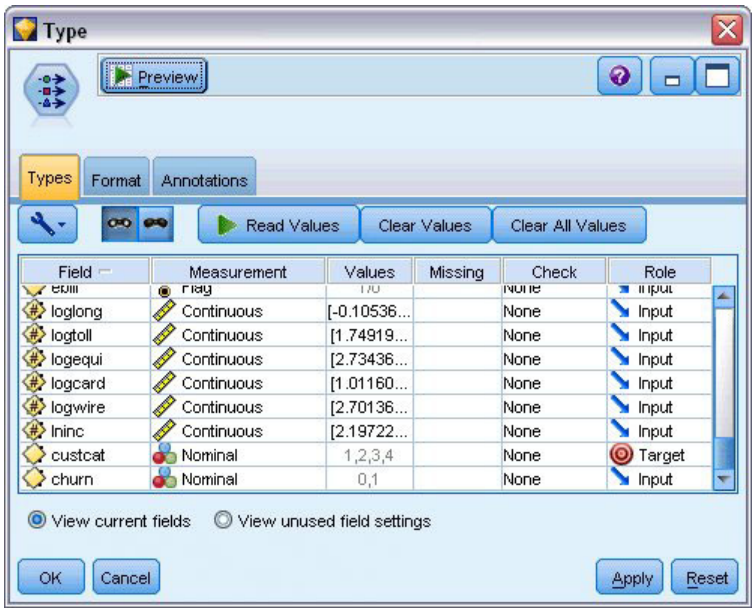


図 146. フィールドの役割の設定

この例はデモグラフィックに注目するため、フィルター・ノードを使用して関連するフィールド (*地域*、*年齢*、*結婚*、*住所*、*収入*、*学歴*、*雇用*、*退職*、*性別*、*居住*、および *custcat*) のみを含めます。この分析の目的上、その他のフィールドは除外してもかまいません。

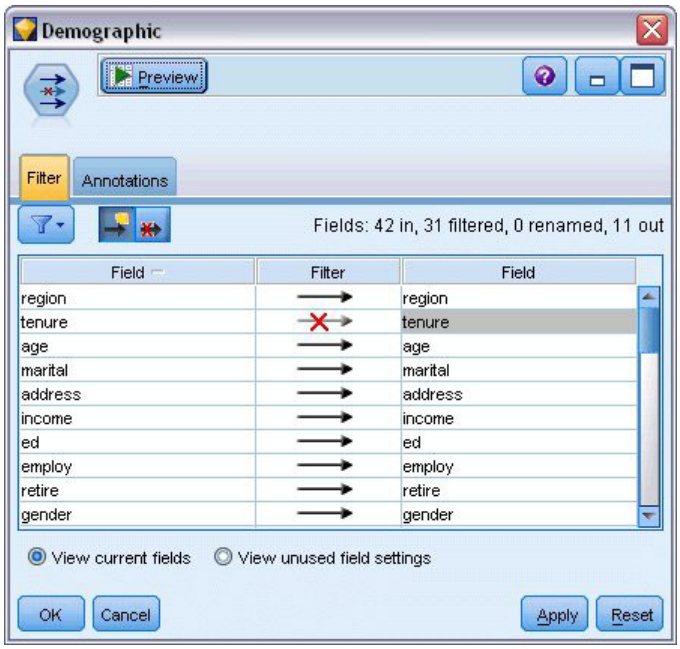


図 147. デモグラフィック・フィールドのフィルタリング

(または、これらのフィールドを除外するのではなく役割を「**なし**」に変更するか、モデル作成ノードで使用するフィールドを選択することもできます。)

2. ロジスティック・ノードで、「**モデル**」タブをクリックし、「**ステップワイズ法**」を選択します。「**多項式**」、「**主効果**」、および「**方程式に定数を含む**」も選択します。

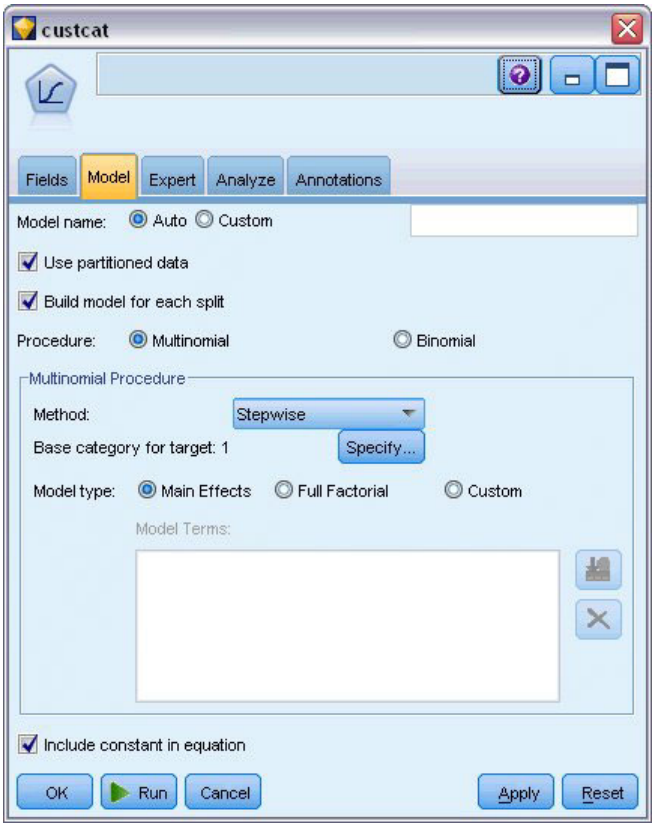


図 148. モデル・オプションの選択

対象のベース・カテゴリーを 1 のままにします。モデルは他の顧客を基本サービスに加入する顧客と比較します。

3. 「エキスパート」タブで、「**エキスパート**」モードを選択し、「**出力**」を選択し、「詳細出力」ダイアログ・ボックスで、「**分類**」を選択します。

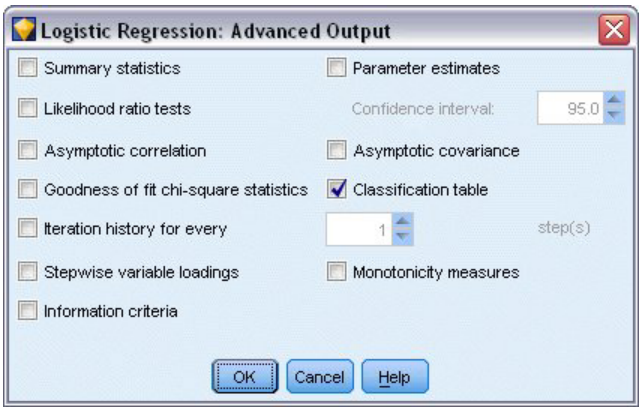


図 149. 出力オプションの選択

モデルの参照

1. ノードを実行してモデルを生成します。これは右上隅の「モデル」パレットに追加されます。詳細を表示するには、生成されたモデル・ノードを右クリックして、「**ブラウズ**」を選択します。

「モデル」タブに、対象フィールドの各カテゴリーにレコードを割り当てるために使用される方程式が表示されます。4 つの有効なカテゴリーがあり、そのうちの 1 つは方程式の詳細が表示されていないベース・カテゴリーです。カテゴリー 3 がプラス・サービスを表すなど、残りの 3 つの方程式についての詳細が表示されています。

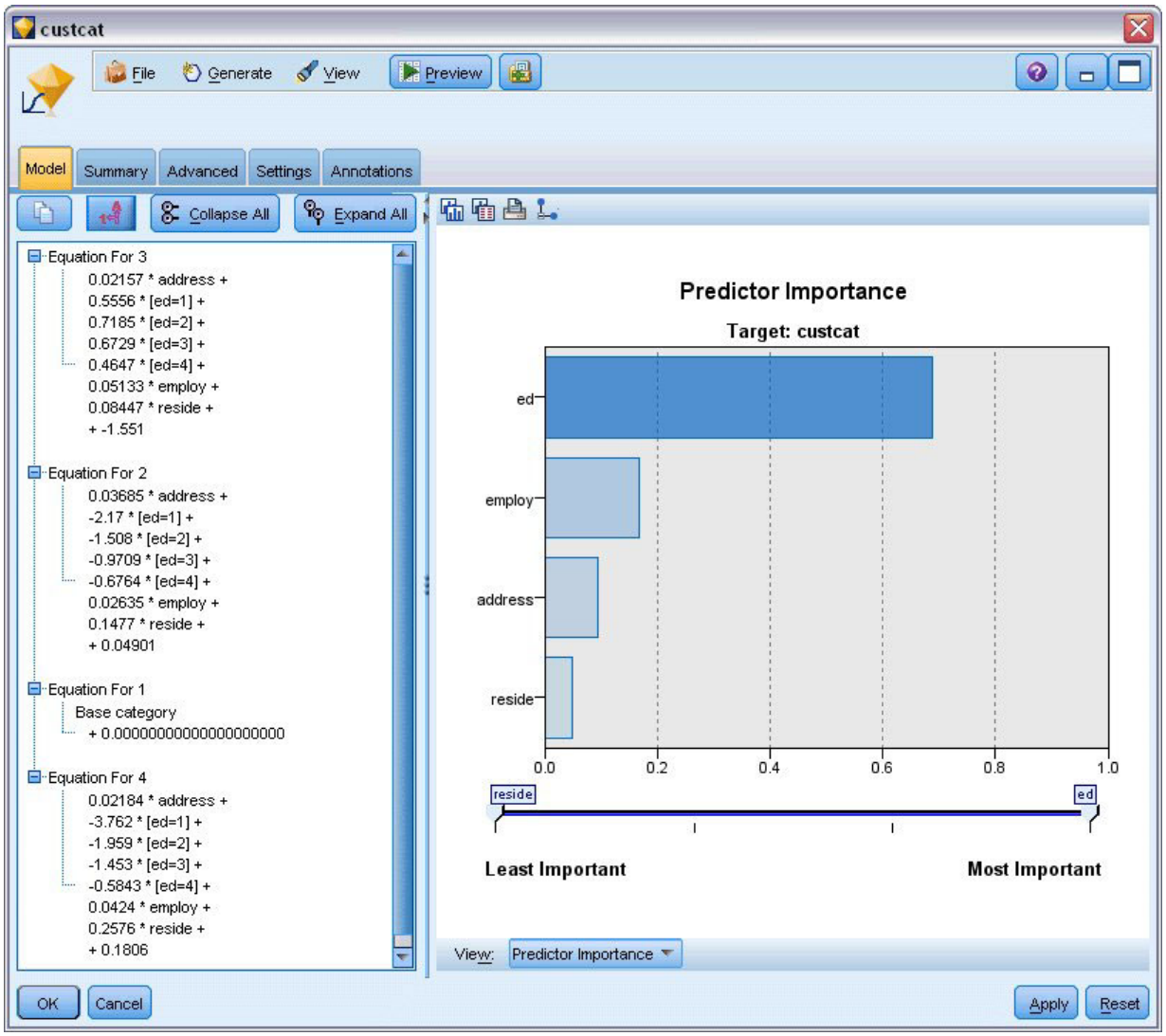


図 150. モデル結果の参照

「要約」タブに、モデルで使用された対象および入力 (予測値フィールド) が (他の項目とともに) 表示されます。ただし、これらは検討するために表示された完全なリストではなく、ステップワイズ法に基づいて実際に選択されたフィールドです。

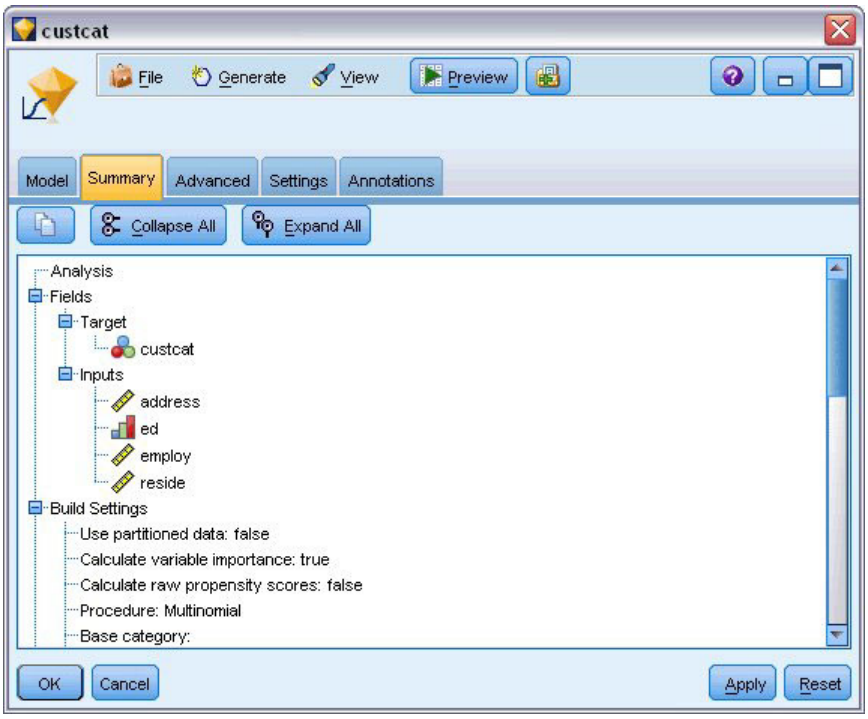


図 151. 対象および入力フィールドが表示されたモデルの要約

「詳細」タブに表示される項目は、モデル作成ノードの「詳細出力」ダイアログ・ボックスで選択されたオプションによって異なります。

常に表示される 1 つの項目は、「処理したケースの要約」です。これは、対象フィールドの各カテゴリーに該当するレコードの割合を表します。これにより、比較の基礎として使用する帰無仮説モデルが得られます。

予測値を使用したモデルを構築しない場合、最も一般的なグループ (プラス・サービス用のもの) にすべての顧客を割り当てるのが、最良の推測になります。

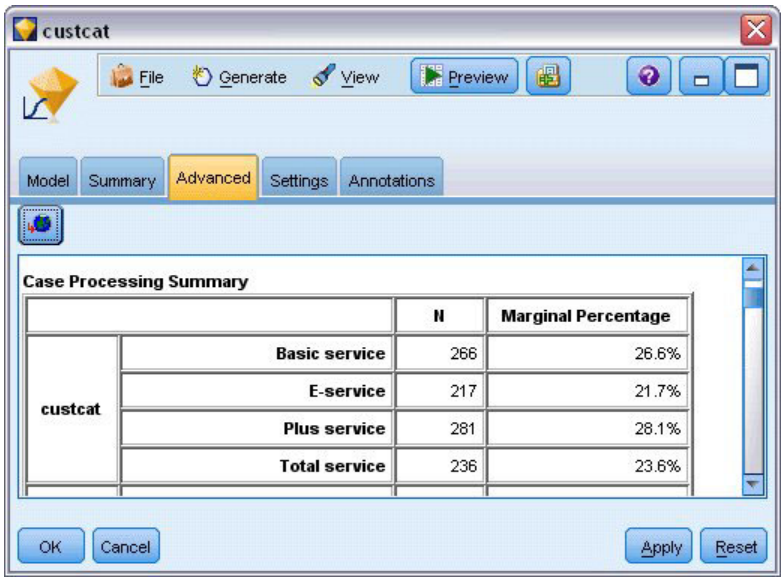


図 152. ケース処理の要約

学習データに基づいて、すべての顧客を帰無仮説モデルに割り当てた場合は、281/1000 = 28.1% の確率で正しくなります。「詳細」タブには、モデルの予測を検討できる情報がさらに表示されます。次に、予測を帰無仮説モデルの結果と比較して、モデルがデータでどれほどうまく機能するかを確認できます。

「詳細」タブの下部の分類テーブルに、モデルの結果が表示されます。これは 39.9% の確率で正確です。

特に、モデルはトータル・サービスの顧客 (カテゴリー 4) を識別する上で優れていますが、E-サービスの顧客 (カテゴリー 2) を識別する場合には非常に劣っています。カテゴリー 2 の顧客に関する精度を向上させる場合は、それらを識別するための別の予測値を見つける必要があると考えられます。

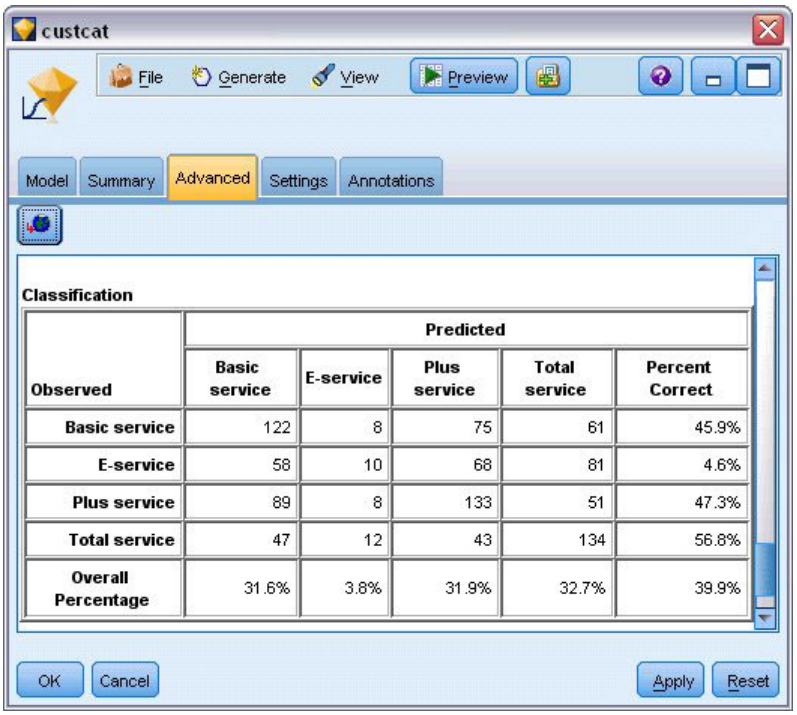


図 153. 分類表

予測するものによっては、このモデルがニーズに完全に合っていることがあります。例えば、カテゴリー 2 の顧客を識別することを重視しない場合は、このモデルは十分です。これは、E-サービスがほとんど利益をもたらさない特売サービスである場合です。

例えば、カテゴリー 3 または 4 に当てはまる顧客から最大の投資収益率を得ている場合は、このモデルから必要な情報を得ることができます。

モデルが実際にデータにどれほどうまく適合するかを評価するため、モデルを構築するときに、「詳細出力」ダイアログ・ボックスでさまざまな診断方法を使用することができます。IBM SPSS Modeler で使用されるモデル作成方法の数学的な基礎の説明については、インストール・ディスクの *¥Documentation* ディレクトリーにある「*IBM SPSS Modeler アルゴリズム・ガイド*」を参照してください。

これらの結果は学習データのみに基づくことに注意してください。モデルが実際の世界の他のデータにどれだけうまく一般化されるかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。

第 13 章 電気通信会社の解約 (2 項検定ロジスティック回帰)

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。線型回帰と似ていますが、数値型対象フィールドではなくカテゴリー対象フィールドを取ります。

この例では、*telco.sav* というデータ・ファイルを参照する *telco_churn.str* というストリームを使用します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*telco_churn.str* ファイルは、*streams* ディレクトリー内にあります。

例えば、競合他社に奪われる顧客の数に関して懸念を抱いている電気通信プロバイダーを想定します。サービス使用量データを、どの顧客が他のプロバイダーに移りそうかを予測するために使用できれば、提案をカスタマイズして、できるだけ多くの顧客を維持することができます。

この例では、顧客消失 (解約) を予測するために使用量データを使用する方法に注目します。対象には 2 つの異なるカテゴリーがあるために、2 項検定モデルを使用します。複数のカテゴリーのある対象の場合は、代わりに多項検定モデルを作成できます。詳細については、133 ページの『第 12 章 電気通信会社の顧客の分類 (多項ロジスティック回帰)』を参照してください。

ストリームの構築

1. *Demos* フォルダーの *telco.sav* を指し示す Statistics ファイル入力ノードを追加します。

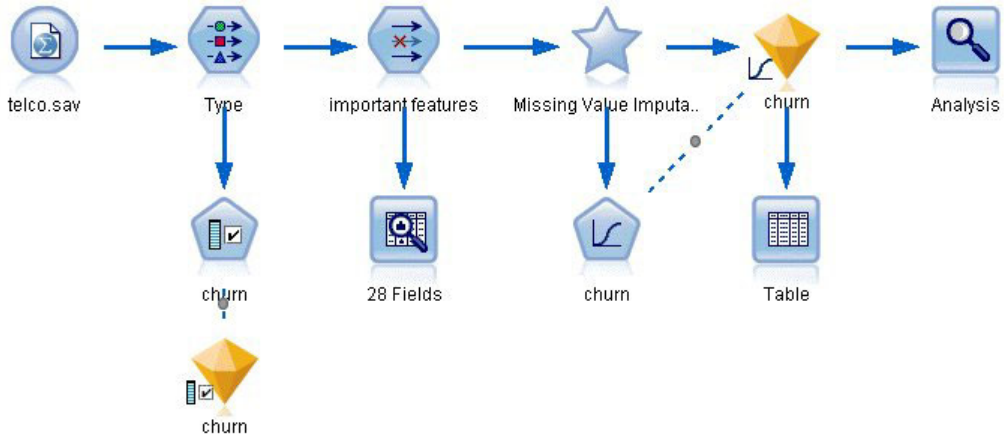


図 154. 2 項検定ロジスティック回帰を使用して顧客を分類するためのサンプル・ストリーム

2. データ型ノードを追加してフィールドを定義し、すべての測定の尺度が正しく設定されていることを確認します。例えば、値 0 および 1 を持つほとんどのフィールドはフラグ型と見なすことができますが、性別などの特定のフィールドは、2 つの値を持つ名義型フィールドの方がより正確に認識されます。

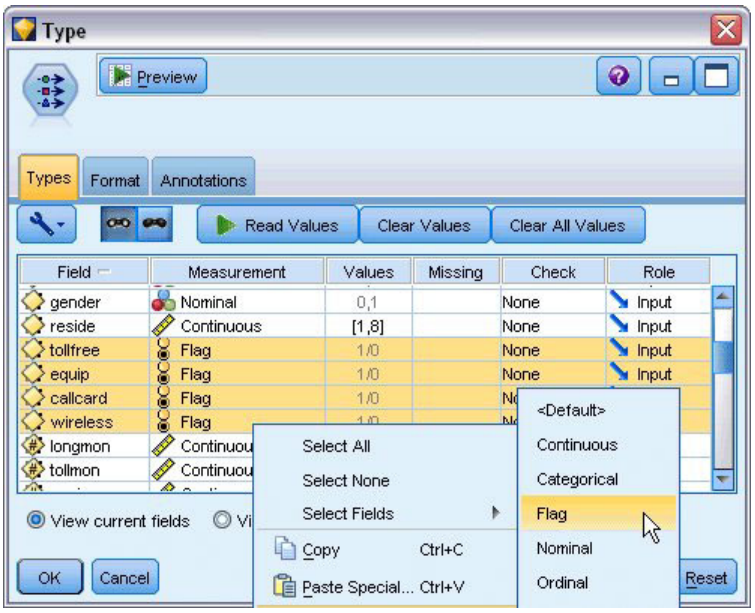


図 155. 複数のフィールドの測定の尺度の設定

ヒント: 類似した値 (0/1 など) を持つ複数のフィールドに対してプロパティーを変更するには、「*値*」列のヘッダーをクリックしてフィールドを値によってソートし、Shift キーを押しながらマウスまたは矢印キーを使用して、変更するフィールドをすべて選択します。選択範囲を右クリックして、測定の尺度を変更するか、選択したフィールドの他の属性を変更します。

3. *churn* フィールドの測定の尺度を「**フラグ型**」に設定し、役割を「**対象**」に設定します。その他のフィールドの役割は、すべて「**入力**」に設定します。

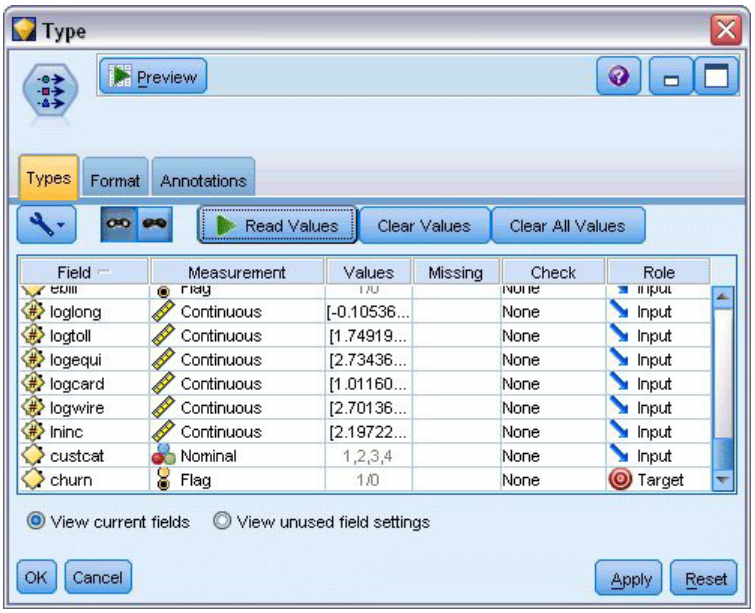


図 156. churn フィールドの測定の尺度および役割の設定

4. フィールド選択モデル作成ノードをデータ型ノードに追加します。

フィールド選択ノードを使用することで、予測値/対象の関係に関して有益な情報を追加しない予測値またはデータを削除することができます。

5. ストリームを実行します。
6. 結果のモデル・ナゲットを開き、「**生成**」メニューから「**フィルター**」を選択してフィルター・ノードを作成します。

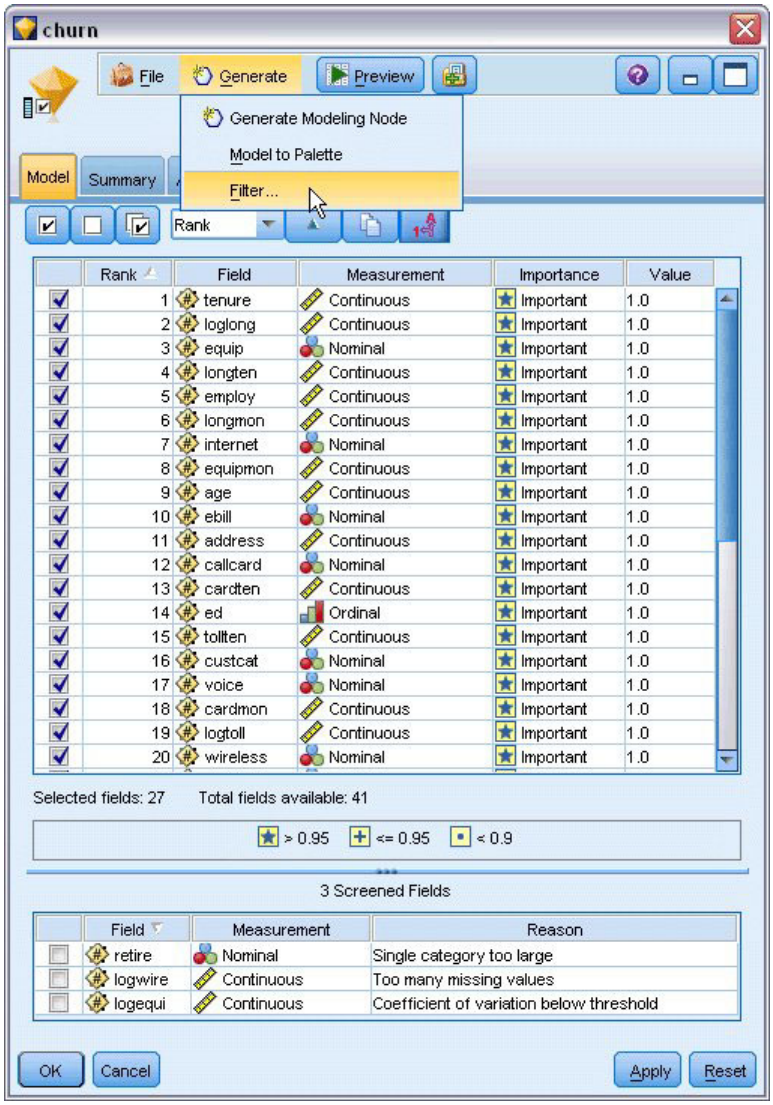


図 157. フィールド選択ノードからのフィルター・ノードの生成

telco.sav ファイル内のすべてのデータが解約の予測で役に立つわけではありません。予測値として使用するために重要と考えられるデータだけを選択するために、フィルターを使用します。

7. 「フィルターの生成」ダイアログ・ボックスで「**マークされているすべてのフィールド: 重要**」を選択し、「**OK**」をクリックします。
8. 生成されたフィルター・ノードをデータ型ノードに接続します。

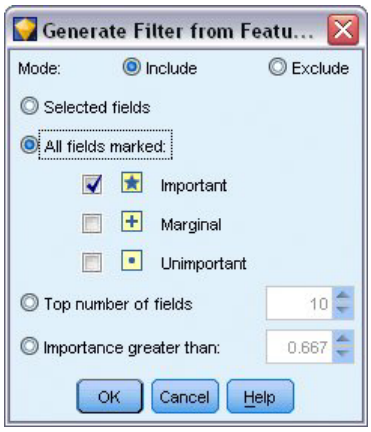


図 158. 重要なフィールドの選択

9. データ監査ノードを生成されたフィルター・ノードに接続します。

   データ監査ノードを開いて、「**実行**」をクリックします。

10. データ監査ブラウザーの「欠損値検査」タブで、「% 完了」列をクリックして、数値の昇順で列をソートします。こうすることで、欠損データの多いフィールドを特定できます。この場合は、修正する必要のあるフィールドは *logtoll* だけです。これは完了の割合が 50% を下回っています。
11. *logtoll* の「欠損値の代入」列で、「**指定**」をクリックします。

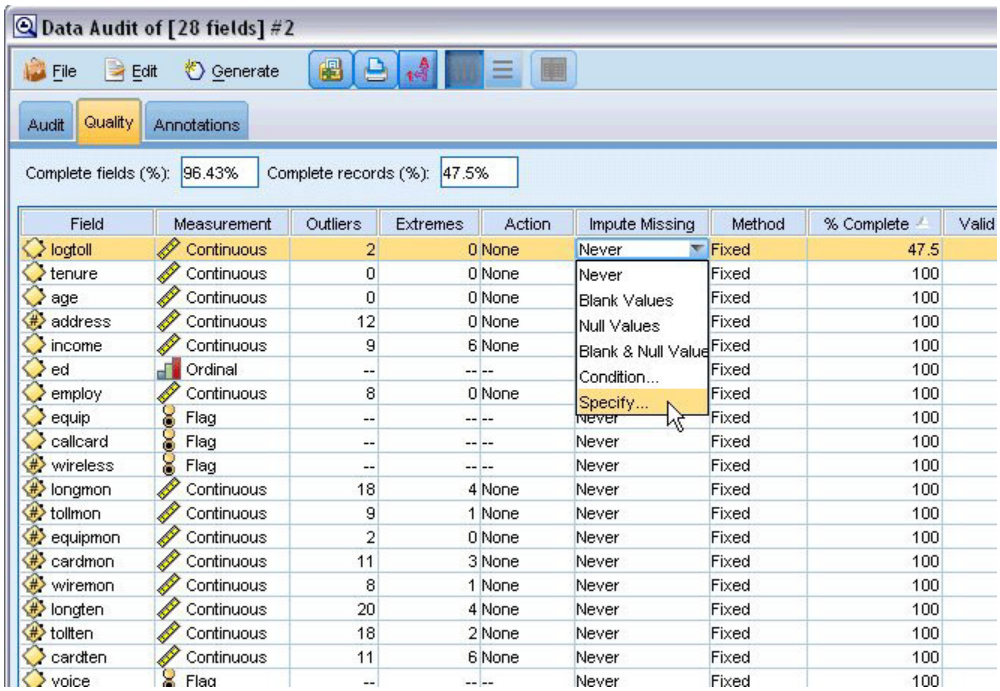


図 159. logtoll の欠損値の代入

12. 「**代入時**」で、「**空白値とヌル値**」を選択します。「**固定**」で、「**平均値**」を選択し、「**OK**」をクリックします。

「**平均値**」を選択すると、代入した値が、全体データの中のすべての値の平均に悪影響を及ぼしません。

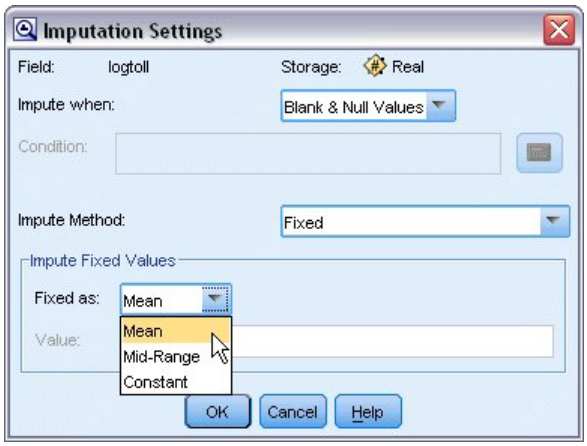


図 160. 代入設定の選択

13. データ監査ブラウザーの「欠損値検査」タブで、欠損値スーパーノードを生成します。これを行うには、メニューから次の項目を選択します。

「生成」 > 「欠損値スーパーノード」

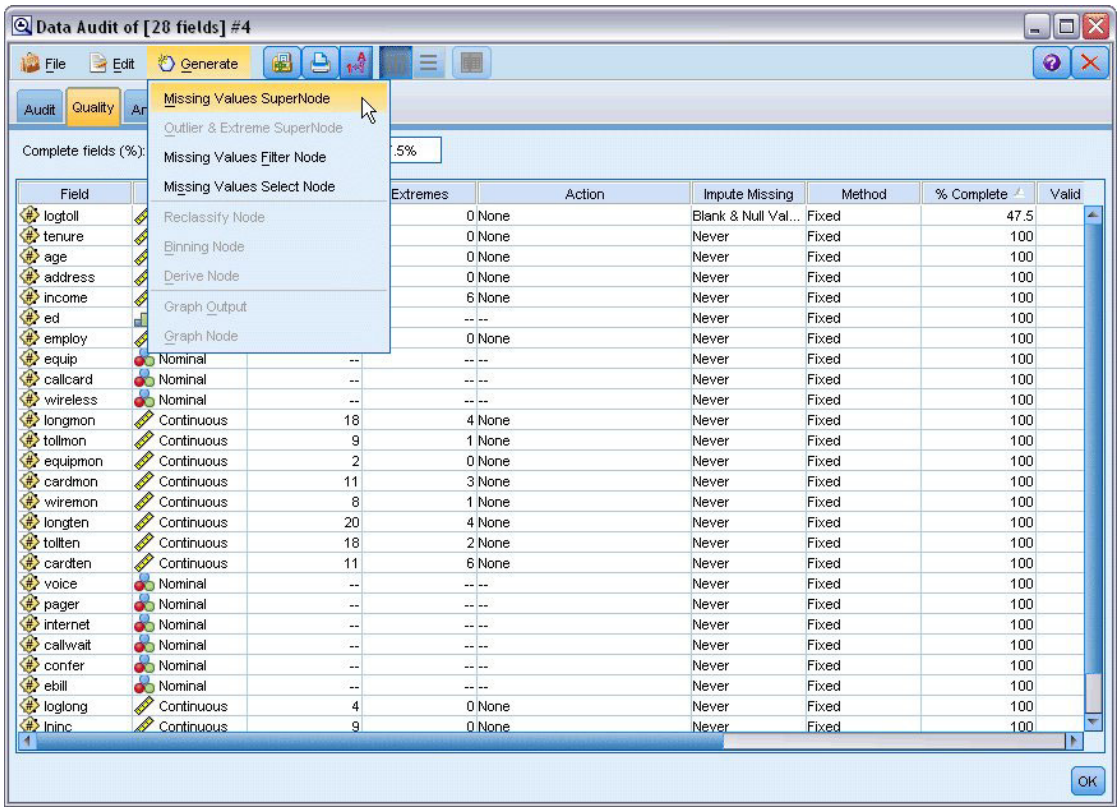


図 161. 欠損値スーパーノードの生成

「欠損値スーパーノード」ダイアログ・ボックスで、「**サンプル・サイズ**」を 50% に増加し、「**OK**」をクリックします。

スーパーノードが、「*欠損値の代入*」というタイトルで、ストリーム領域に表示されます。

14. スーパーノードをフィルター・ノードに接続します。

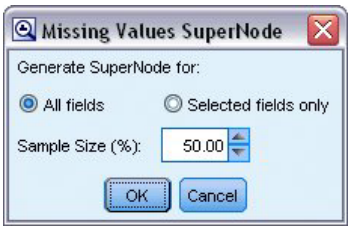


図 162. サンプル・サイズの指定

15. ロジスティック・ノードをスーパーノードに追加します。
16. ロジスティック・ノードで「モデル」タブをクリックし、「**2 項検定**」手続きを選択します。「*2 項検定手続き*」領域で、「**変数増加法**」を選択します。

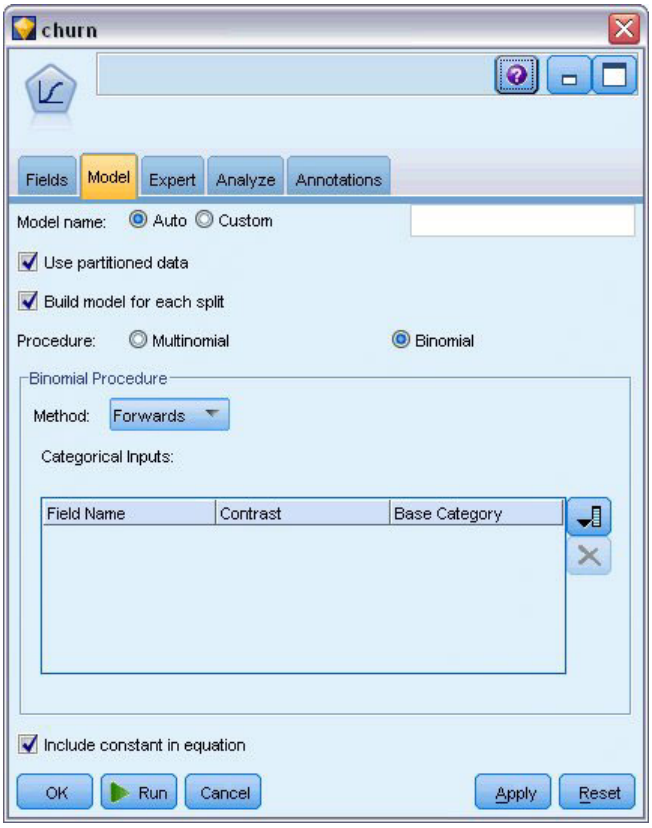


図 163. モデル・オプションの選択

17. 「エキスパート」タブで「**エキスパート**」モードを選択してから、「**出力**」をクリックします。「詳細出力」ダイアログ・ボックスが表示されます。
18. 「詳細出力」ダイアログで、「*表示*」タイプとして「**各ステップごと**」を選択します。「**反復の記述**」および「**パラメーター推定値**」を選択し、「**OK**」をクリックします。

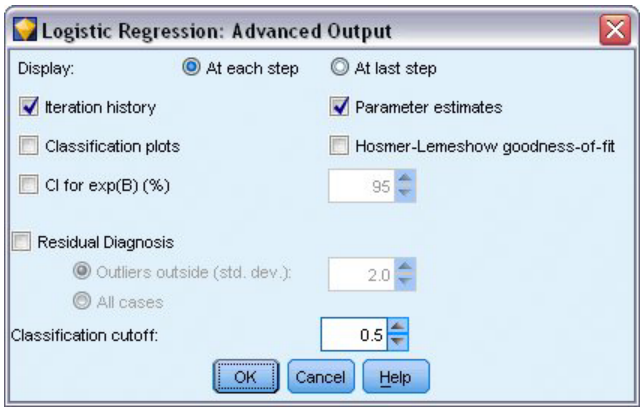


図 164. 出力オプションの選択

モデルの参照

1. ロジスティック・ノードで、「**実行**」をクリックしてモデルを作成します。

   モデル・ナゲットがストリーム領域、および右上隅の「モデル」パレットに追加されます。詳細を表示するには、モデル・ナゲットを右クリックして、「**編集**」または「**ブラウズ**」を選択します。

   「要約」タブに、モデルで使用された対象および入力 (予測値フィールド) が (他の項目とともに) 表示されます。ただし、これらは検討するために表示された完全なリストではなく、変数増加法に基づいて実際に選択されたフィールドです。

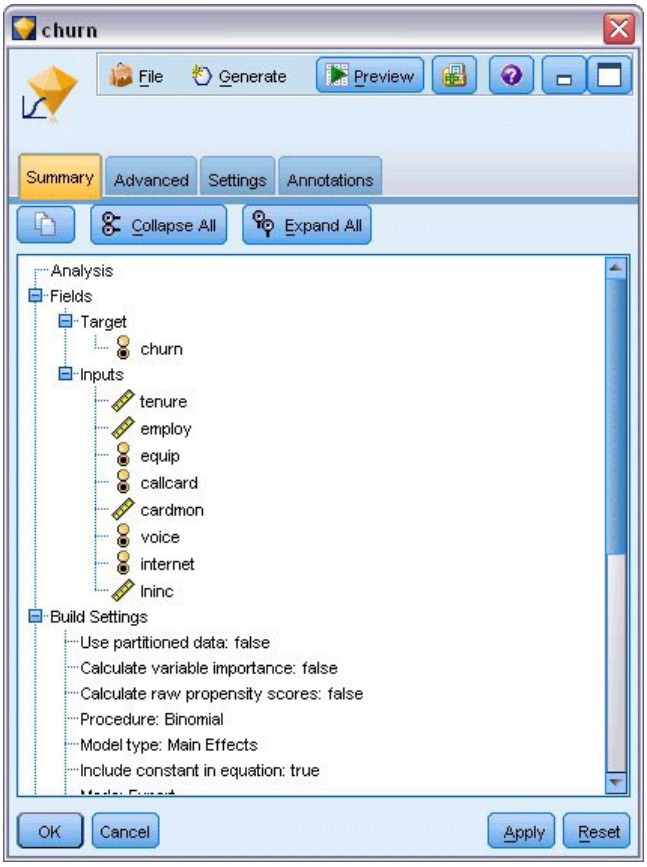


図 165. 対象および入力フィールドが表示されたモデルの要約

「詳細」タブに表示される項目は、ロジスティック・ノードの「詳細出力」ダイアログ・ボックスで選択されたオプションによって異なります。常に表示される 1 つの項目は、「処理したケースの要約」です。これは、分析に含まれているレコードの数および割合を表示します。さらに、1 つ以上の入力フィールドが利用不可の場合に、欠損したケースがある場合は、その数もリストします。また選択されなかったケースの数もリストします。

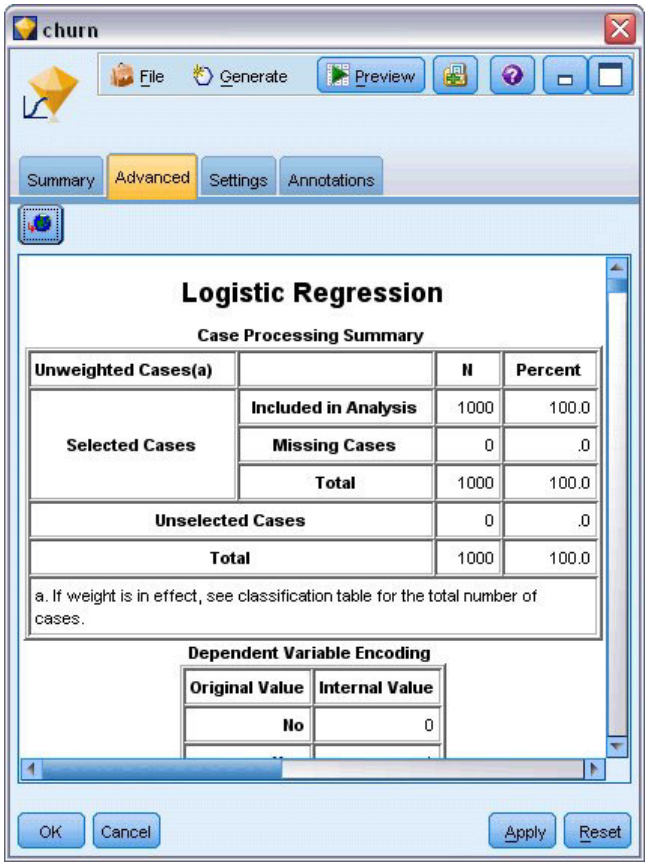


図 166. ケース処理の要約

2. 「処理したケースの要約」からスクロールダウンして、ブロック 0: 開始ブロックの下の分類表を表示します。

   変数増加ステップワイズ法は帰無仮説モデルから開始します。帰無仮説モデルは予測値のないモデルであり、最終ビルド・モデルの比較の基礎として使用できます。帰無仮説モデルは、規約により、すべてのものを 0 として予測するので、帰無仮説モデルは 72.6% の精度です。これは単純に、解約しなかった 726 人の顧客が正しく予測されるためです。ただし、解約した顧客については、まったく正しく予測されません。

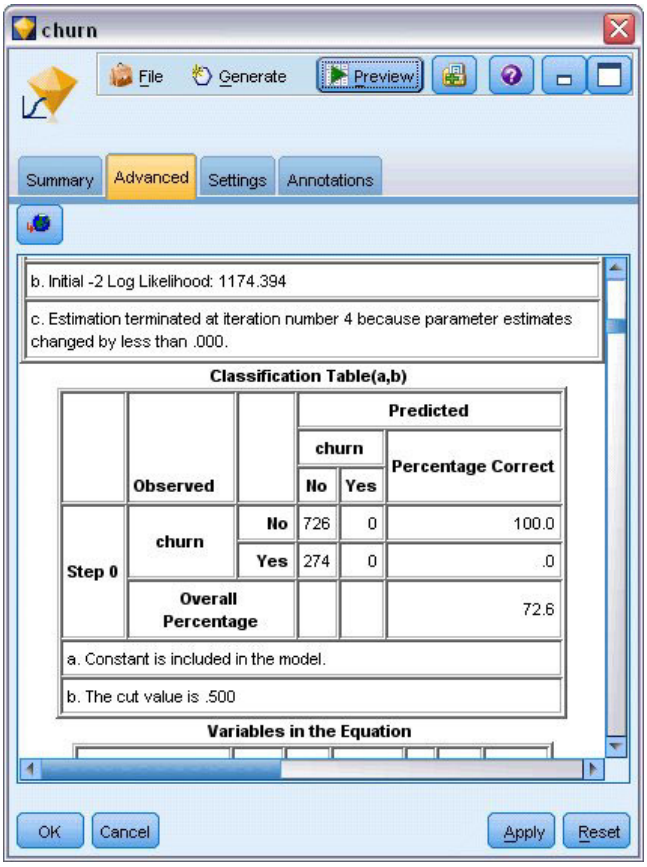


図 167. 分類表 - ブロック 0 の開始

3. それでは、スクロールダウンして、ブロック 1: 方法 = 変数増加ステップワイズ法の下の分類表を表示します。

   この分類表では、ステップごとに予測値が追加されたときのモデルの結果が示されます。既に、最初のステップで、1 つの予測値のみを使用した後で、モデルでは、解約予測の精度が 0.0% から 29.9% に上昇しています。

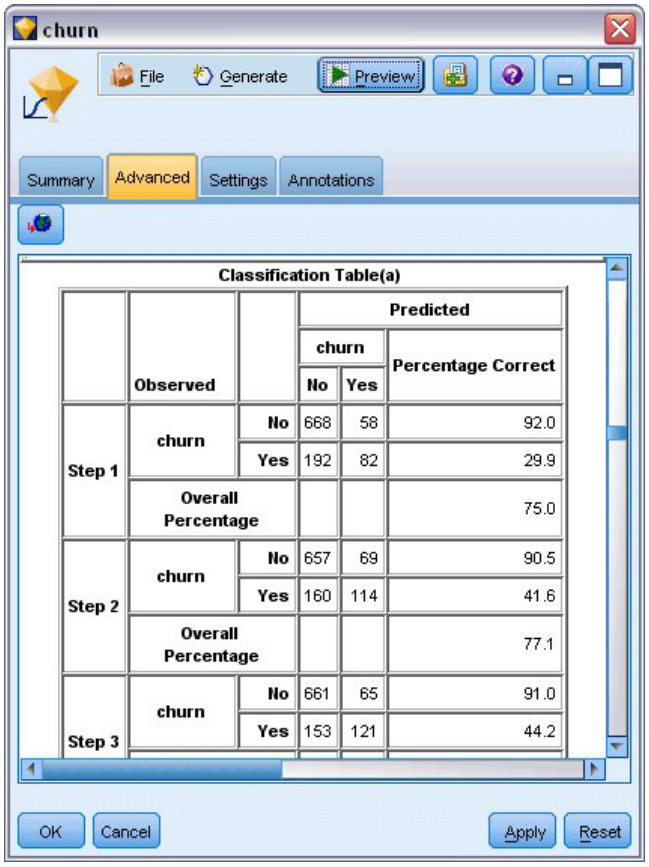


図 168. 分類表 - ブロック 1

4. この分類表の最下部までスクロールダウンします。

分類表は最後のステップがステップ 8 であることを示しています。この段階でアルゴリズムはモデルに予測値を追加することがもはや必要ないと判断しました。解約しない顧客の精度が少し減少して 91.2% になっていますが、解約した顧客の予測精度は、元の 0% から 47.1% に上昇しています。これは、予測値を使用しない元の帰無仮説モデルからの大幅な向上です。

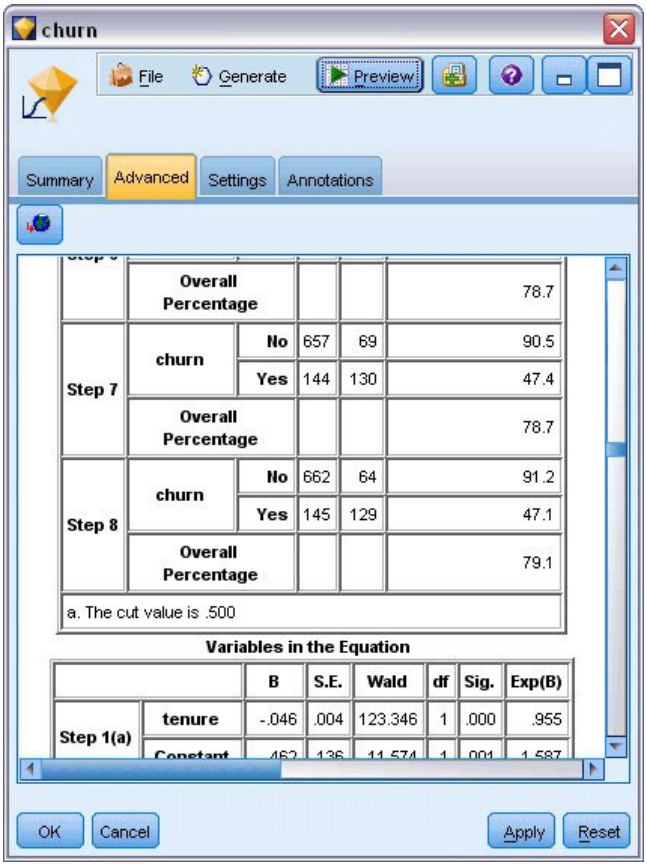


図 169. 分類表 - ブロック 1

解約を減らしたい顧客の場合、解約を約半分に減らすことができるのならば、収入ストリームを保護する上で大きなステップになります。

注: この例では、全体の割合をモデルの精度のガイドとして使用する方法が、一部のケースで、ミスリーディングになることがあることも示されています。元の帰無仮説モデルは 72.6% の全体精度である一方、最終予測モデルの全体精度は 79.1% でした。しかし、見てきたとおり、実際の個々のカテゴリー予測の精度は大きく違っていました。

モデルが実際にデータにどれほどうまく適合するかを評価するため、モデルを構築するときに、「詳細出力」ダイアログ・ボックスでさまざまな診断方法を使用することができます。IBM SPSS Modeler で使用されるモデル作成方法の数学的な基礎の説明については、インストール・ディスクの *¥Documentation* ディレクトリーにある「*IBM SPSS Modeler アルゴリズム・ガイド*」を参照してください。

これらの結果は学習データのみに基づくことに注意してください。モデルが実際の世界の他のデータにどれだけうまく一般化されるかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。

第 14 章 帯域幅の利用状況の予測 (時系列)

時系列ノードによる予測

全国規模のブロードバンド・プロバイダーのアナリストは、帯域幅の利用状況を予測するために、ユーザー加入数の予測を作成する必要があります。予測は、全国的な加入者ベースを構成する地方市場ごとに必要です。さまざまな地方市場の今後 3 カ月の予測を作成するために時系列モデルを使用します。2 番目の例では、ソース・データが時系列ノードに入力するための正しい形式ではない場合のデータの変換方法を示します。

これらの例では、*broadband_1.sav* というデータ・ファイルを参照する *broadband_create_models.str* というストリームを使用します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* フォルダーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*broadband_create_models.str* ファイルは、*streams* フォルダー内にあります。

最後の例では、予測をさらに 3 カ月延長するために、更新したデータ・セットに保存したモデルを適用する方法を示します。

IBM SPSS Modeler では、1 回の操作で複数の時系列モデルを作成できます。使用するソース・ファイルには 85 の異なる市場の時系列データが含まれています。ただし、単純化するために、すべての市場の合計と、これらの市場のうち 5 つだけをモデル化します。

broadband_1.sav データ・ファイルには、85 の地方市場のそれぞれの月次利用データが含まれています。この例では、最初の 5 つの系列のみを使用します。これらの 5 つの系列のそれぞれについての個別のモデルと全体のモデルを作成します。

このファイルには、それぞれのレコードの月と年を示すデータ・フィールドも含まれています。このフィールドは、レコードにラベルを付けるために時間区分ノードで使用します。日付フィールドは IBM SPSS Modeler に文字列で読み込まれますが、IBM SPSS Modeler でこのフィールドを使用するため、置換ノードを使用して、ストレージ・タイプを数値の日付の形式に変換します。

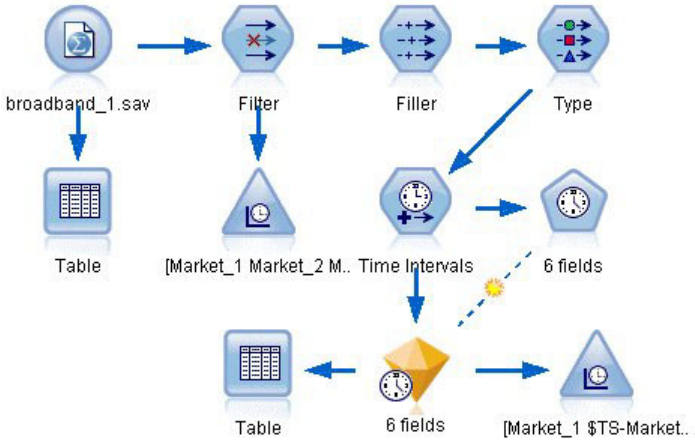
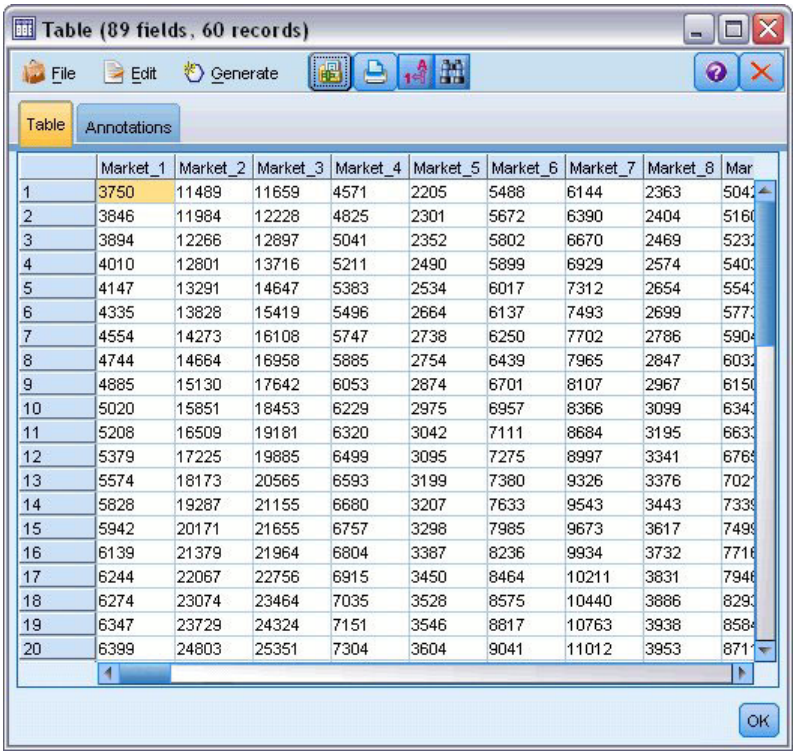


図 170. 時系列モデル作成を示すサンプル・ストリーム

時系列ノードでは、各系列は別々の列に配置され、区間ごとに 1 行が必要です。IBM SPSS Modeler は、この形式に適合するように、必要に応じてデータを変換する方法を提供します。



	Market_1	Market_2	Market_3	Market_4	Market_5	Market_6	Market_7	Market_8	Mar
1	3750	11489	11659	4571	2205	5488	6144	2363	5041
2	3846	11984	12228	4825	2301	5672	6390	2404	5160
3	3894	12266	12897	5041	2352	5802	6670	2469	5232
4	4010	12801	13716	5211	2490	5899	6929	2574	5403
5	4147	13291	14647	5383	2534	6017	7312	2654	5543
6	4335	13828	15419	5496	2664	6137	7493	2699	5773
7	4554	14273	16108	5747	2738	6250	7702	2786	5904
8	4744	14664	16958	5885	2754	6439	7965	2847	6032
9	4885	15130	17642	6053	2874	6701	8107	2967	6150
10	5020	15851	18453	6229	2975	6957	8366	3099	6343
11	5208	16509	19181	6320	3042	7111	8684	3195	6633
12	5379	17225	19885	6499	3095	7275	8997	3341	6765
13	5574	18173	20565	6593	3199	7380	9326	3376	7021
14	5828	19287	21155	6680	3207	7633	9543	3443	7339
15	5942	20171	21655	6757	3298	7985	9673	3617	7499
16	6139	21379	21964	6804	3387	8236	9934	3732	7716
17	6244	22067	22756	6915	3450	8464	10211	3831	7946
18	6274	23074	23464	7035	3528	8575	10440	3886	8293
19	6347	23729	24324	7151	3546	8817	10763	3938	8584
20	6399	24803	25351	7304	3604	9041	11012	3953	8711

図 171. ブロードバンド地方市場の月次加入数データ

ストリームの作成

1. 新規のストリームを作成し、*catalog_seasfac.sav* を指し示す Statistics ファイル入力ノードを追加します。
2. フィルター・ノードを使用して「*Market_6*」から「*Market_85*」までのフィールド、および「*MONTH_*」と「*YEAR_*」フィールドを除外して、モデルを単純化します。

ヒント: 複数の隣接するフィールドを 1 回の操作で選択するには、「*Market_6*」フィールドをクリックし、マウスの左ボタンを押しながら「*Market_85*」フィールドまでマウスをドラッグします。選択されたフィールドが青色に強調表示されます。その他のフィールドも追加するには、Ctrl キーを押しながら「*MONTH_*」および「*YEAR_*」フィールドをクリックします。

図 172. モデルの単純化

データの検証

モデルを構築する前に、使用するデータの性質を理解しておくことをお勧めします。そのデータには季節変動が見られるでしょうか。エキスパート・モデラーは自動的に各系列に最適な季節性モデルまたは非季節性モデルを見つけることができますが、使用するデータに季節性がない場合は検索を非季節性モデルに限定することで、多くの場合は、より速く結果を取得できます。各地方市場のデータを検証しなくても、5 つの市場すべての加入者総数をプロットすることで季節性の有無を大まかに把握できます。

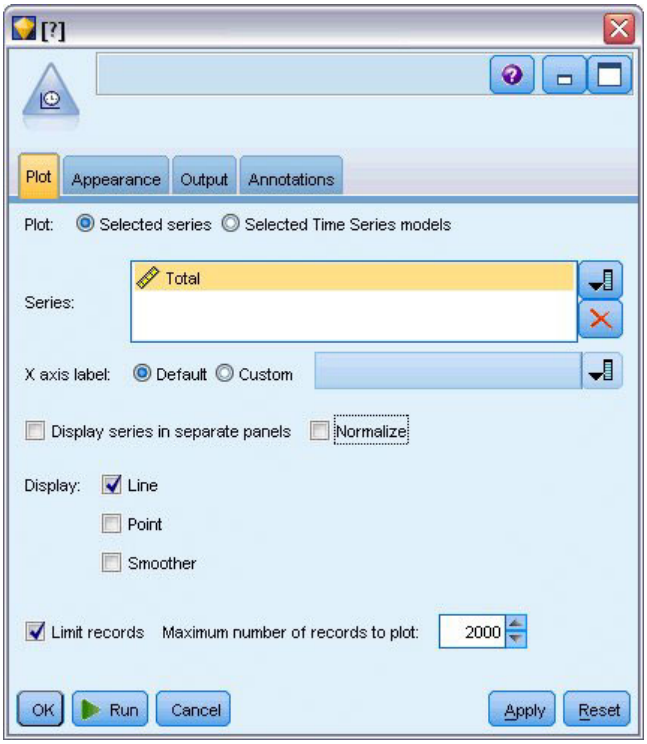


図 173. 加入者総数のプロット

1. 「グラフ」パレットから、時系列ノードをフィルター・ノードに接続します。
2. 「*合計*」フィールドを「系列」リストに追加します。
3. 「**別のパネルに時系列を表示**」および「**正規化**」チェック・ボックスの選択を解除します。
4. 「**実行**」をクリックします。

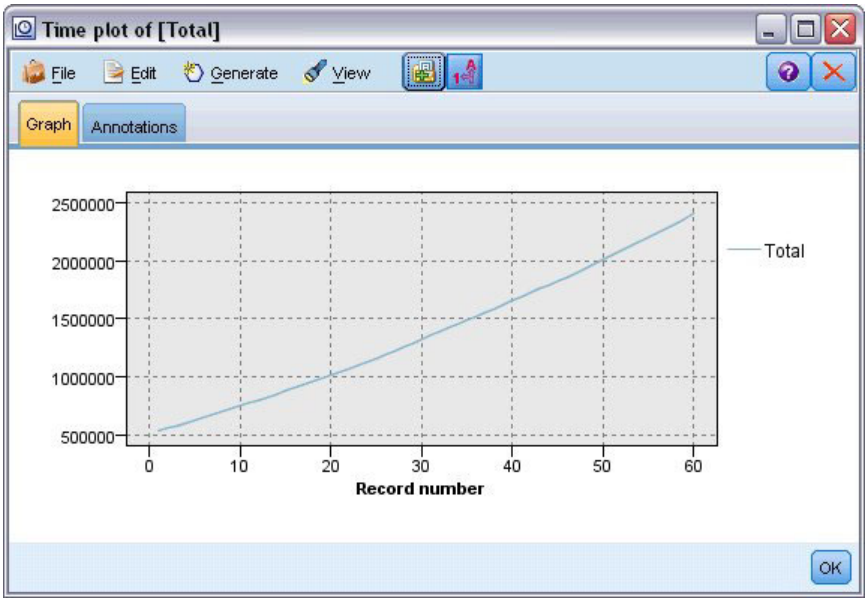


図 174. 「合計」フィールドの時系列

この系列は、非常に滑らかな上昇傾向を示しており、季節変動の存在を示すものはありません。個々の系列については、季節性を持つ系列も存在する可能性はありますが、全体的にはこのデータにおいて季節性は顕著な特徴ではないと考えられます。

もちろん、季節性モデルを除外する前に、個々の系列を調べることは必要です。そこで季節性を示す系列を取り出し、それを別個にモデル化するといいでしょう。

IBM SPSS Modeler を使用すると、複数の系列を一緒に簡単にプロットすることができます。

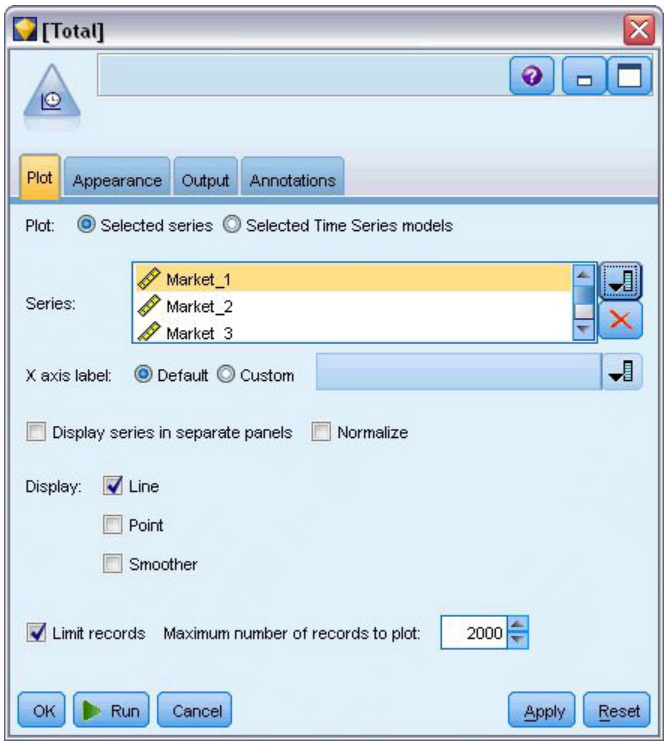


図 175. 複数の時系列のプロット

5. 時系列ノードを開き直します。
6. 「系列」リストから「*合計*」フィールドを削除します (選択して赤色の X ボタンをクリックします)。
7. 「*Market_1*」から「*Market_5*」までのフィールドをリストに追加します。
8. 「**実行**」をクリックします。

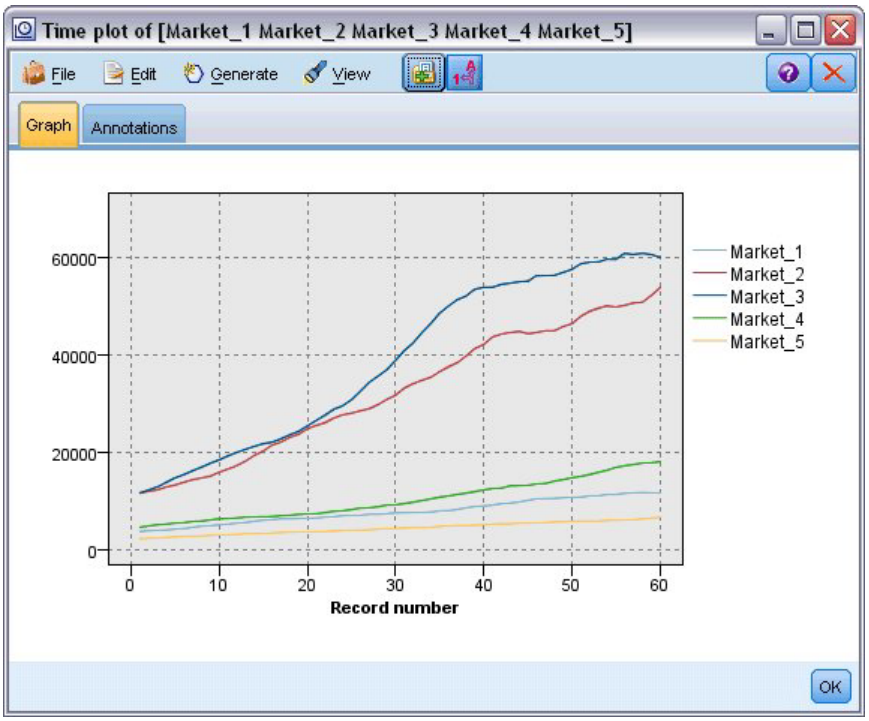


図 176. 複数のフィールドの時系列

それぞれの市場を調べると、各ケースにおける安定した上昇傾向が見られます。一部の市場では他の市場と比較してやや不安定ですが、季節性の兆候は見られません。

日付の定義

この段階で、「*DATE_*」フィールドのストレージ・タイプを日付形式に変更する必要があります。

1. 置換ノードをフィルター・ノードに接続します。
2. 置換ノードを開いてフィールド選択ボタンをクリックします。
3. 「**DATE_**」を選択して、それを「**対象フィールド**」に追加します。
4. 「**置換**」条件に「**常時**」を設定します。
5. 「**置換値**」の値を「**to_date(DATE_)**」に設定します。

図 177. 日付ストレージ・タイプの設定

デフォルトの日付形式を「日付」フィールドの形式と一致するように変更します。これは、「日付」フィールドの変換が期待どおりに機能するために必要です。

6. メニューで、「**ツール**」>「**ストリームのプロパティー**」>「**オプション**」を選択して、「ストリーム・オプション」ダイアログ・ボックスを表示します。
7. デフォルトの「**日付フォーマット**」を「**MON YYYY**」に設定します。

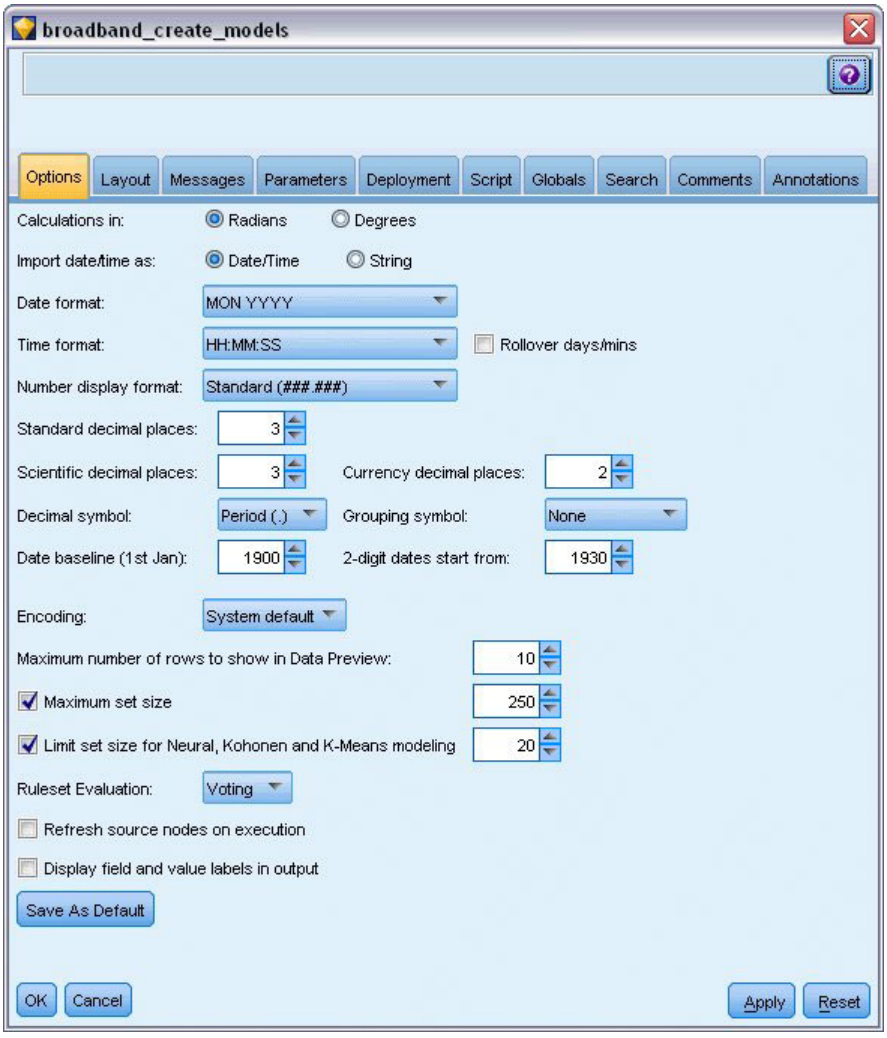


図 178. 日付形式の設定

対象の定義

1. データ型ノードを追加し、「*DATE_* 」フィールドの役割を「**なし**」に設定します。その他のすべてのフィールド (「*Market_n*」フィールドおよび「*合計*」フィールド) の役割を「**対象**」に設定します。
2. 「**値の読み込み**」ボタンをクリックして、「値」列にデータを取り込みます。

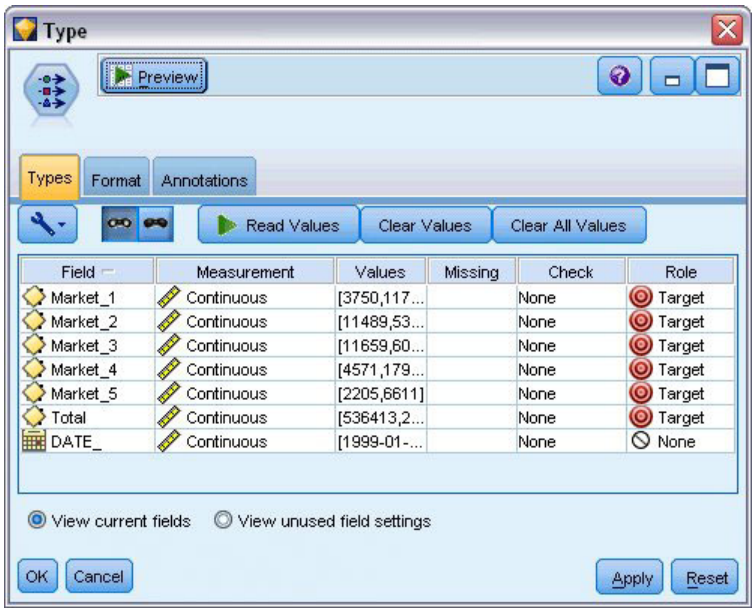


図 179. 複数のフィールドの役割の設定

時間区分の設定

1. 時間区分ノードを追加します (「フィールド設定」パレットから)。
2. 「区間」タブで、時間区分に「**月**」を選択します。
3. 「**データから構築**」オプションを選択します。
4. 構築フィールドとして「**DATE_**」を選択します。

図 180. 時間区分の設定

5. 「予測」タブで、「**レコードの将来への拡張**」チェック・ボックスを選択します。
6. 値を **3** に設定します。
7. 「**OK**」をクリックします。

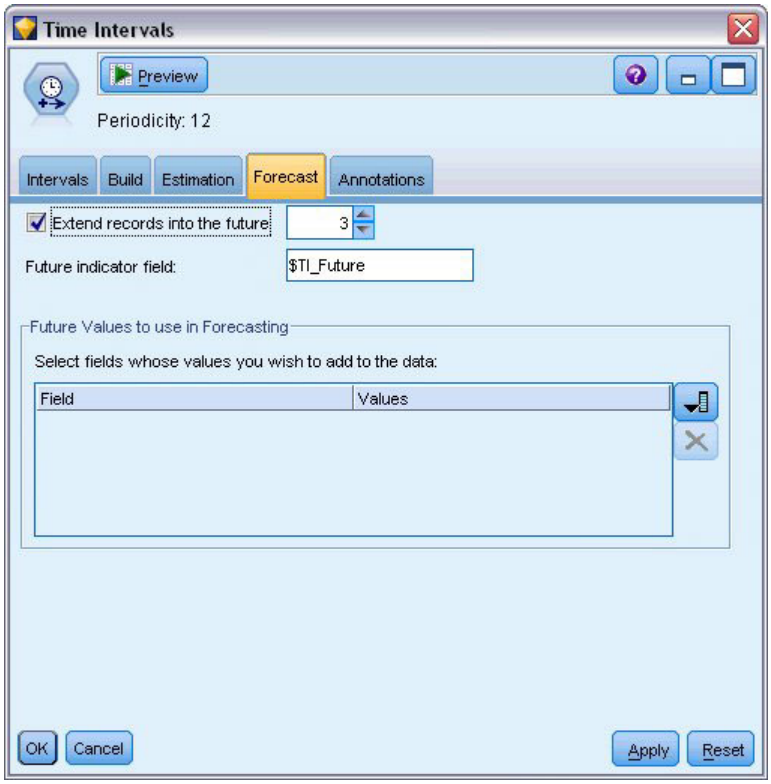


図 181. 予測期間の設定

モデルの作成

1. 「モデル作成」パレットから、時系列ノードをストリームに追加して時間区分ノードに接続します。
2. すべてのデフォルト設定を使用して時系列ノードで「**実行**」をクリックします。これにより、エキスパート・モデラーはそれぞれの時系列に使用する最適なモデルを決定できます。

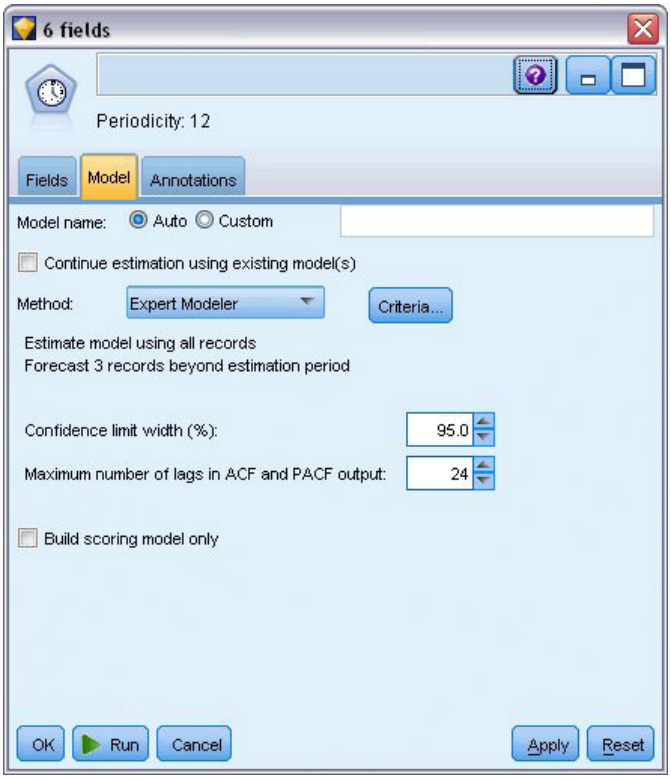


図 182. 時系列のエキスパート・モデラーの選択

3. 時系列モデル・ナゲットを時間区分ノードに接続します。
4. テーブル・ノードを時系列モデルに接続して「**実行**」をクリックします。

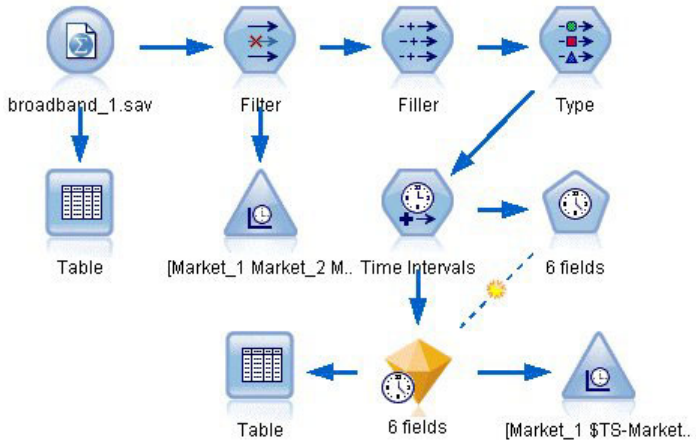


図 183. 時系列モデル作成を示すサンプル・ストリーム

これで 3 つの新しい行 (61 から 63) が元のデータに追加されます。これらは、予測期間 (この場合は 2004 年 1 月から 3 月) の行です。

いくつかの新しい列も追加されています。時間区分ノードによって追加された多くの *$TI_* 列と、時系列ノードによって追加された *$TS-* 列です。これらの列は、各行の以下の項目を示しています (つまり、時系列データにおける各区分)。

列(L)	説明
$TI_TimeIndex	この行の時間区分の指標値。
$TI_TimeLabel	この行の時間区分のラベル。
$TI_Year	この行で生成されたデータの年および月の指標。
$TI_Month	
$TI_Count	この行に対する新しいデータの確認に関係したレコードの数。
$TI_Future	この行に予測データが含まれるかどうかを示す。
$TS-*colname*	元のデータの各列に対して生成されたモデル・データ。
$TSLCI-*colname*	生成されたモデル・データの各列の信頼区間の下限値。
$TSUCI-*colname*	生成されたモデル・データの各列の信頼区間の上限値。
$TS-Total	この行の $TS-*colname* 値の合計。
$TSLCI-Total	この行の $TSLCI-*colname* 値の合計。
$TSUCI-Total	この行の $TSUCI-*colname* 値の合計。

予測操作で最も重要な列は、*$TS-Market_n*、*$TSLCI-Market_n*、および *$TSUCI-Market_n* 列です。特に、行 61 から 63 のこれらの列には、各地方市場のユーザー加入予測データおよび信頼区間が含まれています。

モデルの検証

1. 時系列モデル・ナゲットをダブルクリックして、市場ごとに生成されたモデルに関するデータを表示します。

   エキスパート・モデラーが、Market 5 に対して他の市場に生成したタイプとは異なるタイプのモデルを作成することを、どのように選択したかに注目してください。

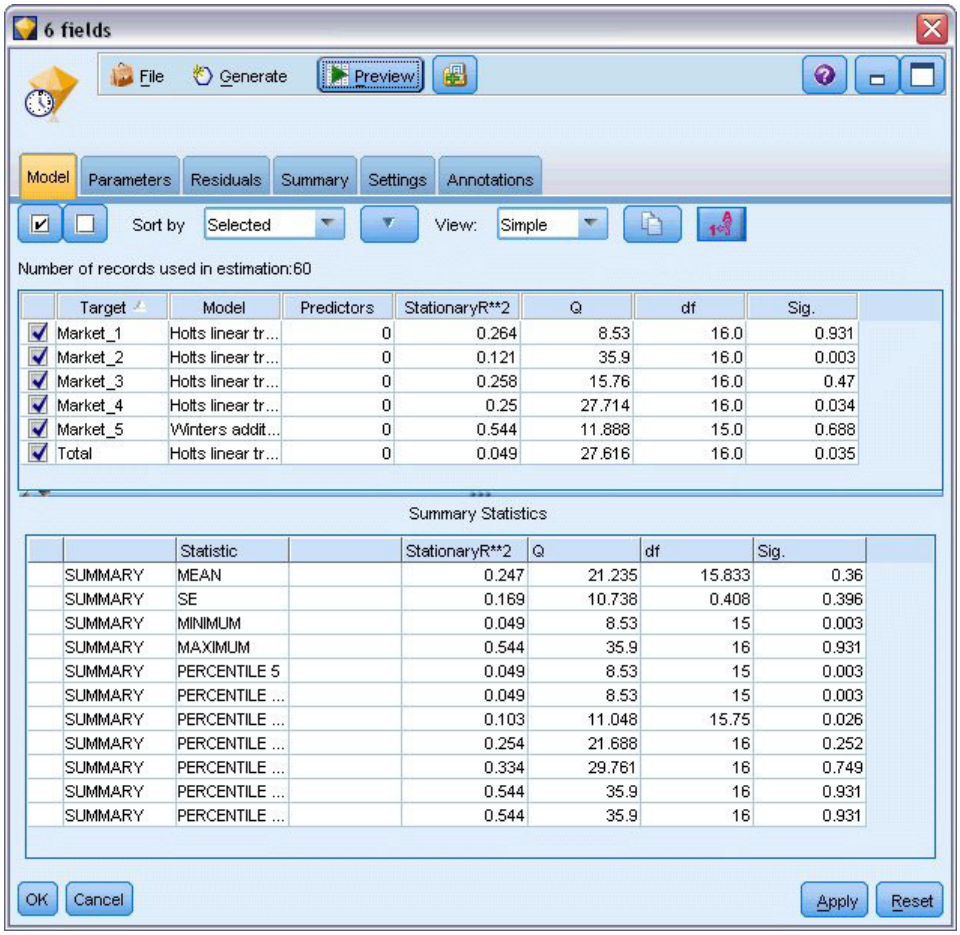


図 184. 市場に対して生成された時系列モデル

「予測値」列には、各対象に対して予測値として使用されたフィールド数が示されますが、今回は 0 です。

このビューのその他の列には、各モデルのさまざまな適合度の測定結果が示されます。「**定常 R-2 乗**」列には、定常 *R* 2 乗の値が示されます。この統計は、そのモデルによって説明される系列における総変動の比率の推定を示しています。値が大きいほど (最大 1.0)、モデルの適合度は良好です。

「**Q**」、「**自由度**」、および 「**有意確率**」 列は、モデルの残差エラーのランダム性の検定である Ljung-Box 統計に関連し 、エラーがランダムであるほど、そのモデルは良好です。「**Q**」は Ljung-Box 統計そのもので、「**自由度**」 (自由度の程度) は特定の対象を推定する際に自由に変化するモデル・パラメーターの数を示します。

「**有意確率**」列は、Ljung-Box 統計の有意確率値を示し、そのモデルが正しく指定されているかどうかに関するもう 1 つの指標です。 0.05 未満の有意確率値は、残差エラーがランダムではないことを示し、モデルでは考慮されていない構造が観測対象の系列にあるということを意味しています。

定常 *R* 2 乗値と有意確立値の両方を考慮すると、エキスパート・モデラーが *Market_1*、*Market_3*、および *Market_5* に対して選択したモデルは非常に良好です。 *Market_2* と *Market_4* の「**有意確立**」値は、どちらも 0.05 未満であり、これらのマーケットに対してより適合度が高いモデルを使用した実験が必要な可能性があることを示しています。

下部に表示された要約値からは、すべてのモデルの統計分布に関する情報が得られます。例えば、すべてのモデルの定常 *R* 2 乗値の平均は 0.247 であるのに対し、その最小値は 0.049 (*合計* モデルの値) で、最大値は 0.544 (*Market_5* の値) です。

SE は、統計ごとの、すべてのモデルの標準誤差を意味します。例えば、すべてのモデルの定常 *R* 2 乗の標準誤差は 0.169 です。

要約セクションには、モデル間の統計の分布に関する情報を示すパーセンタイル値も示されています。各パーセンタイルについて、そのモデルの割合は、表示されている値を下回る適合度統計量の値です。

したがって、例えば 25% のモデルのみに、0.121 未満の定常 *R* 2 乗値があります。

2. 「表示」ドロップダウン・リストをクリックし、「**詳細**」を選択します。

さまざまな適合度の追加測定結果が表示されます。「**R**2**」は *R* 2 乗値であり、そのモデルによって説明可能な時系列内の総変動の推定です。この統計の最大値は 1.0 で、この点で良好なモデルといえます。

図 185. 時系列モデルの詳細表示

2 乗平均平方根誤差 (**RMSE**) は、系列の実際の値がモデルで予測される値とどの程度異なるかの尺度であり、系列自体の単位と同じ単位で表されます。これは誤差の尺度であるため、この値はできるだけ低いことが期待されます。一見すると、*Market_2* および *Market_3* のモデルは、これまでに確認した統計に関しては妥当ですが、他の 3 つの市場のモデルと比較すると、それほど良好ではありません。

これらのその他の適合度測定結果には、平均絶対パーセント誤差 (**MAPE**) およびその最大値 (**MaxAPE**) が含まれます。絶対パーセント誤差は、モデル予測レベルから対象系列が変動する量に関する尺度で、パーセント値で表されます。すべてのモデルで平均値および最大値を検証することで、予測の不確定性についての目安を得ることができます。

MAPE 値を見ると、すべてのモデルが 1% を下回る非常に低い平均不確定性を示しています。MaxAPE 値は最大絶対パーセント誤差を表し、予測の最悪のシナリオを想定するために役立ちます。ここでは、それぞれのモデルの最大パーセント誤差がおよそ 1.8 から 2.5% の範囲内に収まる、非常に低い数値が再び示されています。

MAE (平均絶対誤差) 値は、予測の誤差の絶対値の平均を示します。RMSE 値と同様に、系列自体の単位と同じ単位で表されます。**MaxAE** は、最大予測誤差を同じ単位で示すもので、予測の最悪のシナリオを示します。

これらの絶対値は興味深いのですが、この場合は、対象系列が規模が変動する市場の加入者数を表すため、パーセント誤差の値 (MAPE および MaxAPE) の方が役立ちます。

MAPE 値および MaxAPE 値が表す不確定性は、そのモデルで許容される程度でしょうか。ここでは明らかに非常に低い値となっています。許容できるリスクは問題に応じて変化するため、これについてはビジネス・センスを活用する場面です。適合度統計は許容範囲内に収まると想定して、残差エラーの確認に進みます。

モデルの残差の自己相関関数 (ACF) および偏自己相関関数 (PACF) の値を検証することにより、単に適合度統計を表示するよりも、モデルに関してより数量的な洞察が得られます。

適切に指定された時系列モデルでは、季節性、トレンド、循環性などの重要な因子をはじめ、すべての無作為でない変動が取得されます。これに該当する場合、どの誤差についても、経時的にそれ自体と相関 (自己相関) させるべきではありません。自己相関関数のいずれかに有意な構造が見られる場合、それは基礎となるモデルが不完全であることを意味します。

3. 「残差」タブをクリックして、地方市場の最初のモデルにおける残差エラーの自己相関関数 (ACF) および偏自己相関関数 (PACF) の値を表示します。

図 186. 市場の ACF および PACF の値

これらのプロットでは、誤差変数の元の値を最大 24 期間遅延させて、元の値と比較することによって、経時的な相関があるかどうかを確認します。モデルとして許容されるには、上側のプロット (ACF) のどの棒も、正 (上) または負 (下) のいずれかの方向に、色の濃い領域からはみ出していない必要があります。

はみ出している場合は、下側のプロット (PACF) でその構造が確定されているかどうかを確認する必要があります。PACF プロットは、介入時点で系列値を制御した後の相関に注目しています。

Market_1 の値はすべて、色の濃い領域内にあるため、引き続き他の市場の値を確認できます。

4. **「モデル用プロットの表示」** ドロップダウン・リストをクリックして、他の市場および合計に対するこれらの値を表示します。

   Market_2 および *Market_4* にはわずかな懸念事項があるため、**「有意確立」**の値から、前に疑った事項を確認します。ある時点のこれらの市場に対してさまざまなモデルで実験を行い、より優れた適合度が得られるかを確認する必要がありますが、この例のその他の部分では、*Market_1* モデルから学習できるその他の内容について集中します。

5. 「グラフ」パレットから、時系列ノードを時系列モデル・ナゲットに接続します。
6. 「プロット」タブで、**「別のパネルに時系列を表示」**チェック・ボックスのチェック・マークを外します。

7. 「**系列**」リストでフィールド選択ボタンをクリックし、「*Market_1*」および「*$TS-Market_1*」フィールドを選択し、 「**OK**」をクリックしてリストに追加します。
8. 「**実行**」をクリックして、地方市場の 1 つ目の実際のデータおよび予測データの折れ線グラフを表示します。

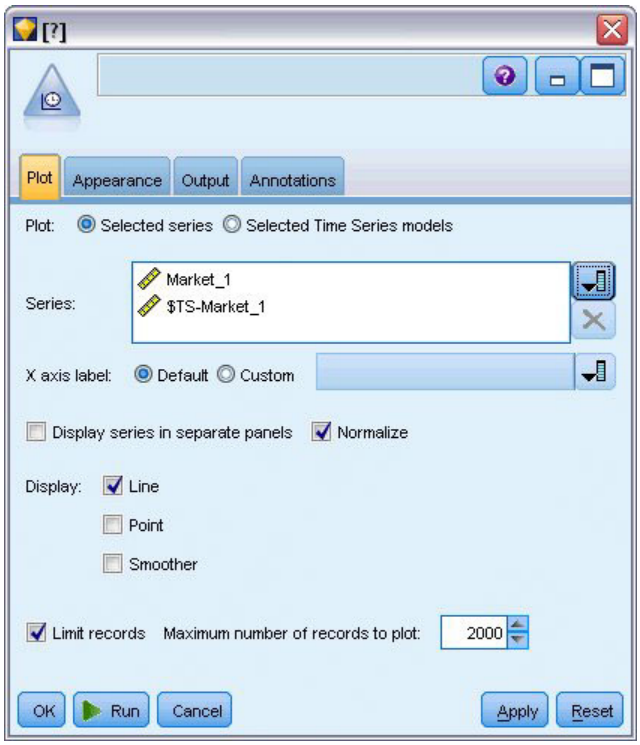


図 187. プロットするフィールドの選択

予測 (*$TS-Market_1*) の線が実際のデータの終端以降にどのように伸びるかに注目してください。これが、この市場の今後 3 カ月間の見込み需要の予測です。

時系列全体の実際のデータと予測データの線は、グラフ上で密接しており、モデルがこの特定の時系列に対しては信頼できることを示しています。

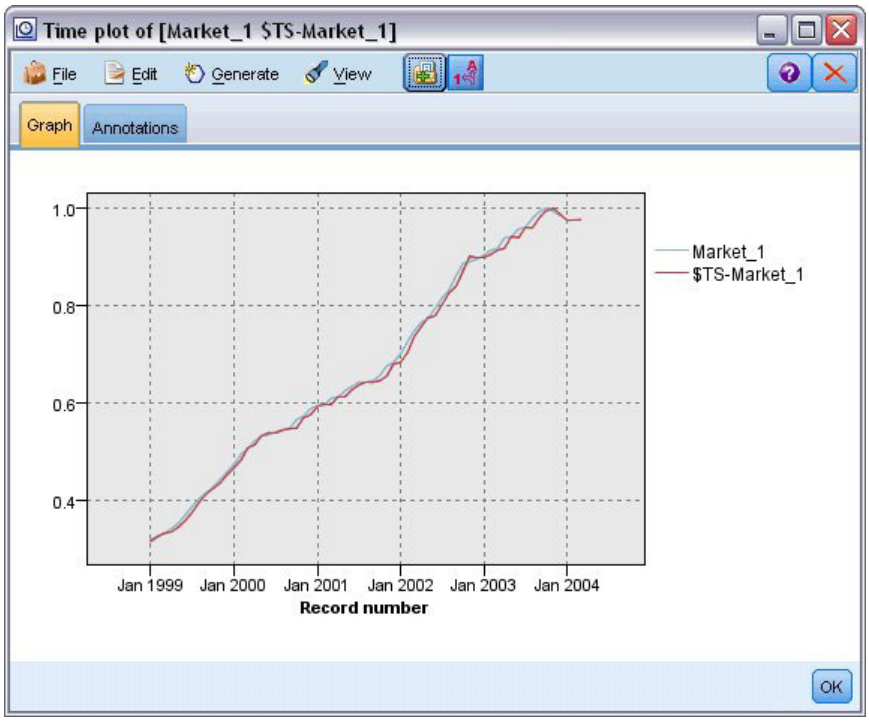


図 188. Market_1 の実際のデータと予測データの時系列

今後の例で使用できるように、モデルをファイルに保存します。

9. 「**OK**」をクリックして現在のグラフを閉じます。
10. 時系列モデル・ナゲットを開きます。
11. 「**ファイル**」>「**ノードの保存**」を選択し、ファイルの場所を指定します。
12. 「**保存**」をクリックします。

この特定の市場に対して信頼できるモデルが得られましたが、その予測はどのような誤差の許容範囲を持つでしょうか。信頼区間を検証することにより、この指標が得られます。

13. ストリームの最後の時系列ノード (ラベルは **Market_1 $TS-Market_1**) をダブルクリックして、ダイアログ・ボックスを再度開きます。
14. フィールド選択ボタンをクリックし、「*$TSLCI-Market_1*」および「*$TSUCI-Market_1*」フィールドを「**系列**」リストに追加します。
15. 「**実行**」をクリックします。

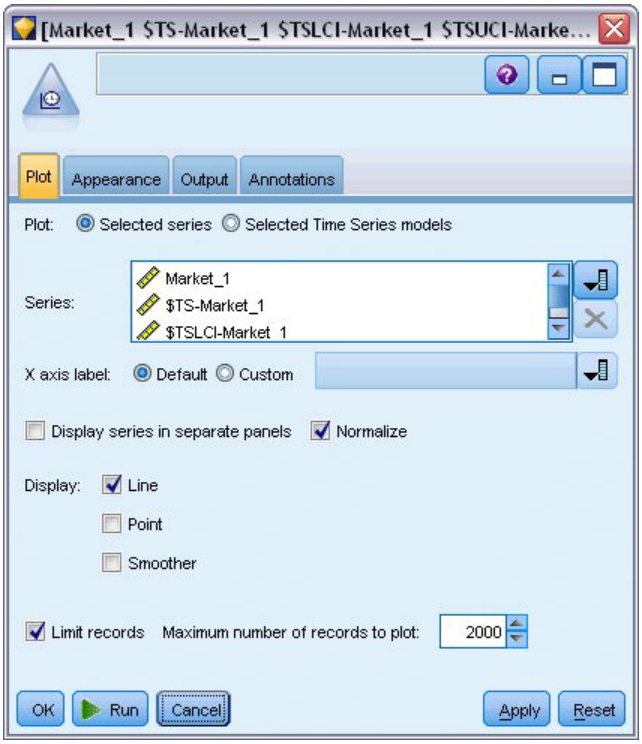


図 189. プロットへのフィールドの追加

前と同様のグラフが得られましたが、今回は信頼区間の上限 (*$TSUCI*) および下限 (*$TSLCI*) が追加されています。

信頼区分の境界が予測期間中にどのように分岐しているかに注目してください。予測が先の将来に進むほど不確定性が増大する様子を示しています。

ただし、期間が経過するたびに、予測の基礎となる 1 カ月分 (この場合) の実際の使用データが新たに得られます。この新しいデータをストリームに読み取り、信頼できると分かっているモデルを再適用することができます。詳細については、 173 ページの『時系列モデルの再適用』を参照してください。

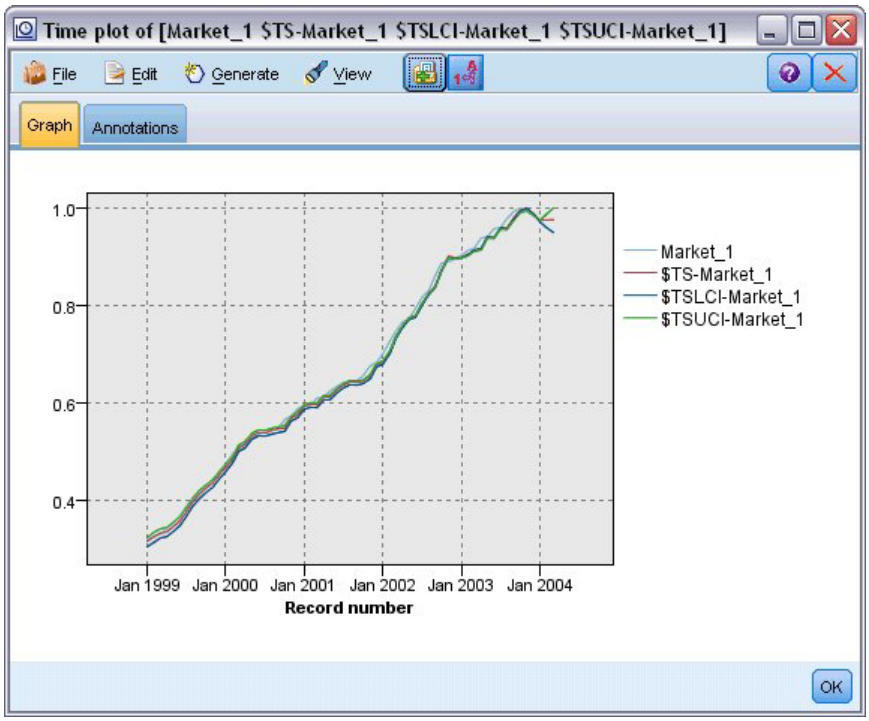


図 190. 信頼区分が追加された時系列

要約

エキスパート・モデラーを使用して複数の時系列の予測を生成する方法について説明し、結果として得られるモデルを外部のファイルに保存しました。

次の例では、非標準的な時系列データを時系列ノードに入力するのに適した形式に変換する方法について解説します。

時系列モデルの再適用

この例では、最初の時系列の例の時系列モデルを適用しますが、単独で使用することもできます。詳細については、 153 ページの『時系列ノードによる予測』を参照してください。

元のシナリオと同様に、全国的なブロードバンド・プロバイダーのアナリストは、帯域幅の要求基準を予測するためにさまざまな地方の市場のそれぞれについてユーザー加入数の月次予測を作成する必要があります。ただし、モデルは既にエキスパート・モデラーを使用して作成されており、今後 3 カ月の予測結果も出ています。

データウェアハウスは元の予測期間の実際のデータで更新されているため、そのデータを使用して予測期間をさらに 3 カ月延長します。

この例では、*broadband_2.sav* というデータ・ファイルを参照する *broadband_apply_models.str* というストリームを使用します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* フォルダーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*broadband_apply_models.str* ファイルは、*streams* フォルダー内にあります。

ストリームの取得

この例では、最初の例で保存した時系列モデルから時系列ノードを再作成します。このモデルを保存していない場合は、「*Demos*」フォルダーに用意されているモデルを使用できます。

1. 「*Demos*」の下の「*streams* 」フォルダーからストリーム *broadband_apply_models.str* を開きます。

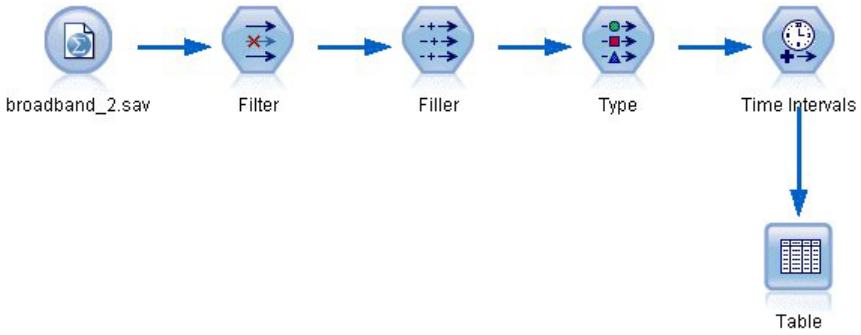
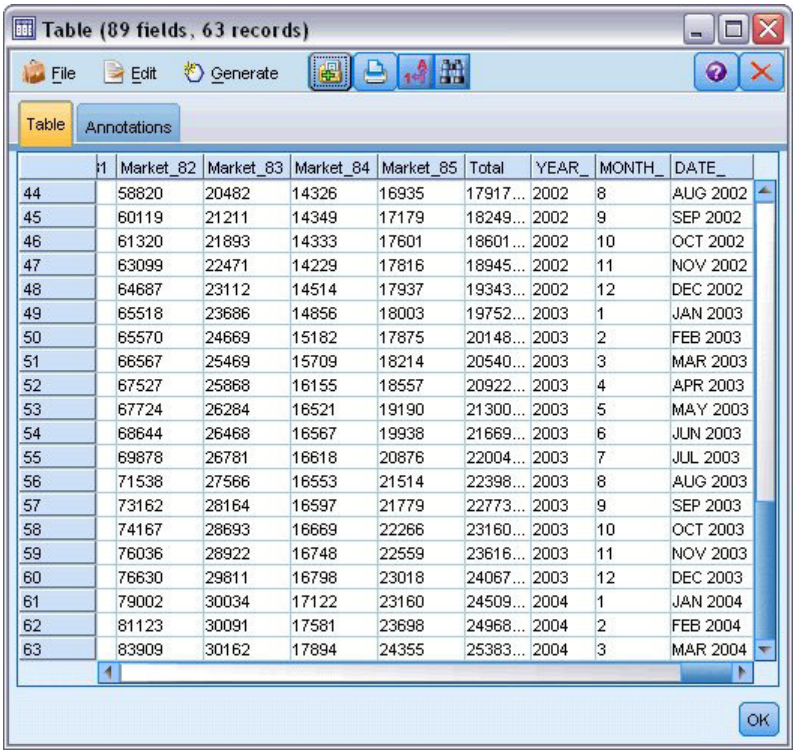


図 191. ストリームを開く



	Market_82	Market_83	Market_84	Market_85	Total	YEAR_	MONTH_	DATE_
44	58820	20482	14326	16935	17917...	2002	8	AUG 2002
45	60119	21211	14349	17179	18249...	2002	9	SEP 2002
46	61320	21893	14333	17601	18601...	2002	10	OCT 2002
47	63099	22471	14229	17816	18945...	2002	11	NOV 2002
48	64687	23112	14514	17937	19343...	2002	12	DEC 2002
49	65518	23686	14856	18003	19752...	2003	1	JAN 2003
50	65570	24669	15182	17875	20148...	2003	2	FEB 2003
51	66567	25469	15709	18214	20540...	2003	3	MAR 2003
52	67527	25868	16155	18557	20922...	2003	4	APR 2003
53	67724	26284	16521	19190	21300...	2003	5	MAY 2003
54	68644	26468	16567	19938	21669...	2003	6	JUN 2003
55	69878	26781	16618	20876	22004...	2003	7	JUL 2003
56	71538	27566	16553	21514	22398...	2003	8	AUG 2003
57	73162	28164	16597	21779	22773...	2003	9	SEP 2003
58	74167	28693	16669	22266	23160...	2003	10	OCT 2003
59	76036	28922	16748	22559	23616...	2003	11	NOV 2003
60	76630	29811	16798	23018	24067...	2003	12	DEC 2003
61	79002	30034	17122	23160	24509...	2004	1	JAN 2004
62	81123	30091	17581	23698	24968...	2004	2	FEB 2004
63	83909	30162	17894	24355	25383...	2004	3	MAR 2004

図 192. 更新された販売データ

更新済みの月次データは *broadband_2.sav* に収集されています。

2. テーブル・ノードを IBM SPSS Statistics ファイル入力ノードに接続し、テーブル・ノードを開いて「**実行**」をクリックします。

   注: データ・ファイルは、2004 年 1 月から 3 月の実際の販売データ (行 61 から 63) で更新されています。

3. ストリームの時間区分ノードを開きます。

4. 「**予測**」タブをクリックします。
5. 「**レコードの将来への拡張**」が **3** に設定されていることを確認します。

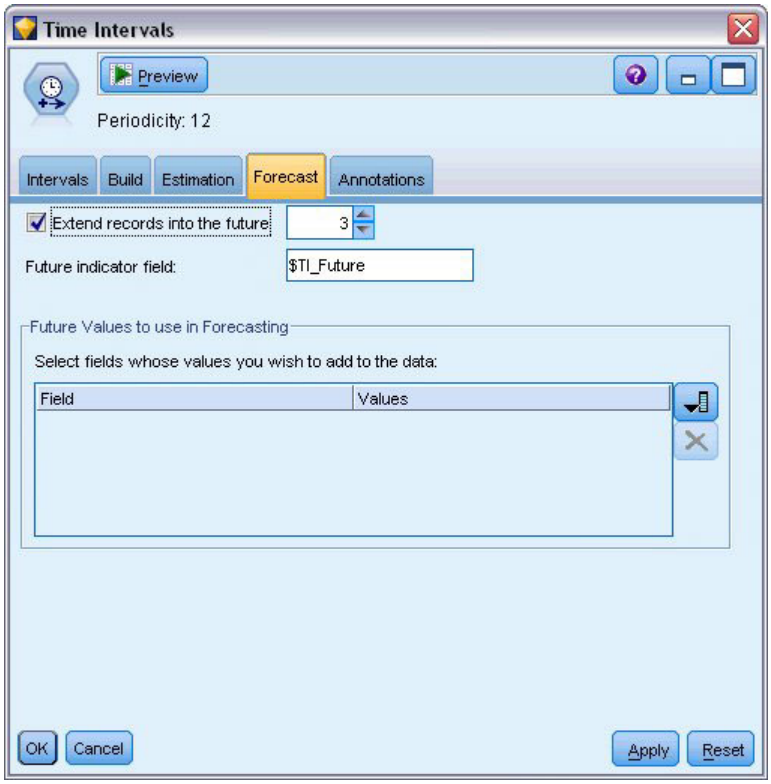


図 193. 予測期間の設定の確認

保存されたモデルの取得

1. IBM SPSS Modeler メニューで、「**挿入**」>「**ファイルからノード**」を選択し、「*Demos*」フォルダーから *TSmodel.nod* ファイルを選択します (または、最初の時系列の例で保存した時系列モデルを使用します)。

このファイルには、以前の例からの時系列モデルが含まれています。挿入操作を行うと、対応する時系列モデル・ナゲットが領域に配置されます。

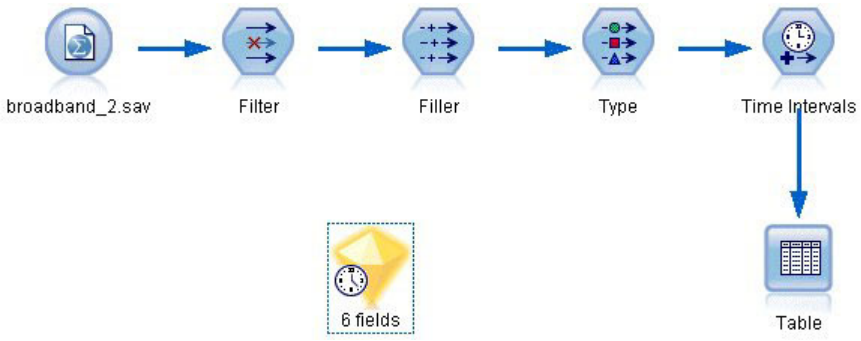


図 194. モデル・ナゲットの追加

モデル作成ノードの生成

1. 時系列モデル・ナゲットを開き、「**生成**」>「**モデル作成ノードを生成**」を選択します。

これにより、時系列モデル作成ノードが領域に配置されます。

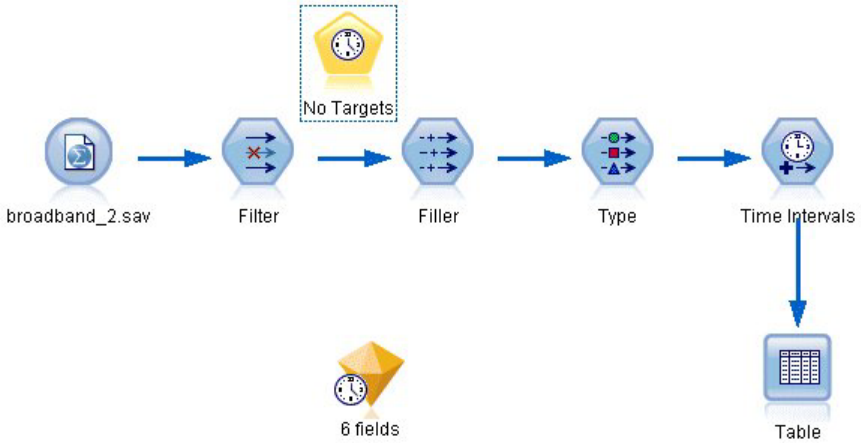


図 195. モデル・ナゲットからのモデル作成ノードの生成

新しいモデルの生成

1. 時系列モデル・ナゲットを閉じ、領域から削除します。

   古いモデルは、60 行のデータに基づいて構築されていました。更新済みの販売データ (63 行) に基づいて新しいモデルを生成する必要があります。
2. 新しく生成された時系列構築ノードをストリームに接続します。

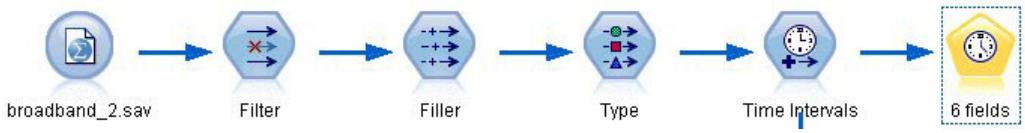


図 196. ストリームへのモデル作成ノードの接続

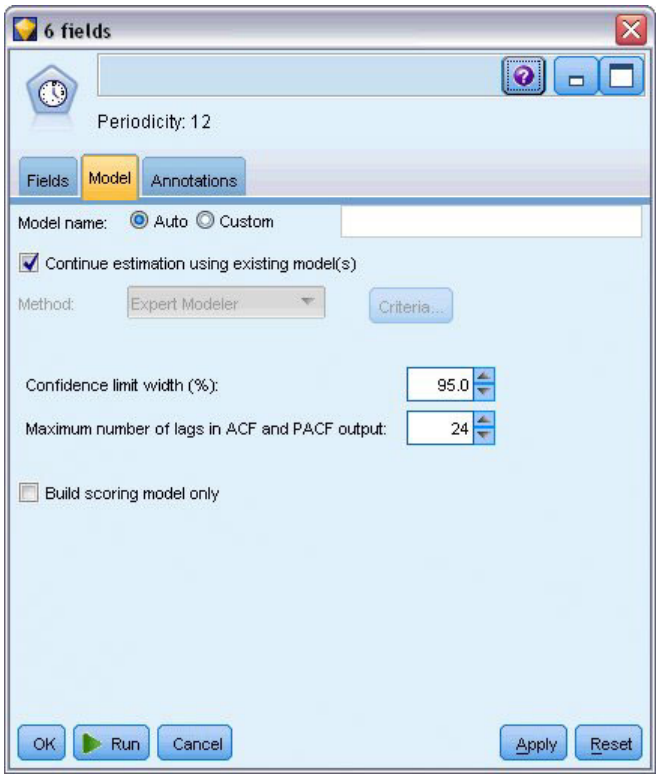
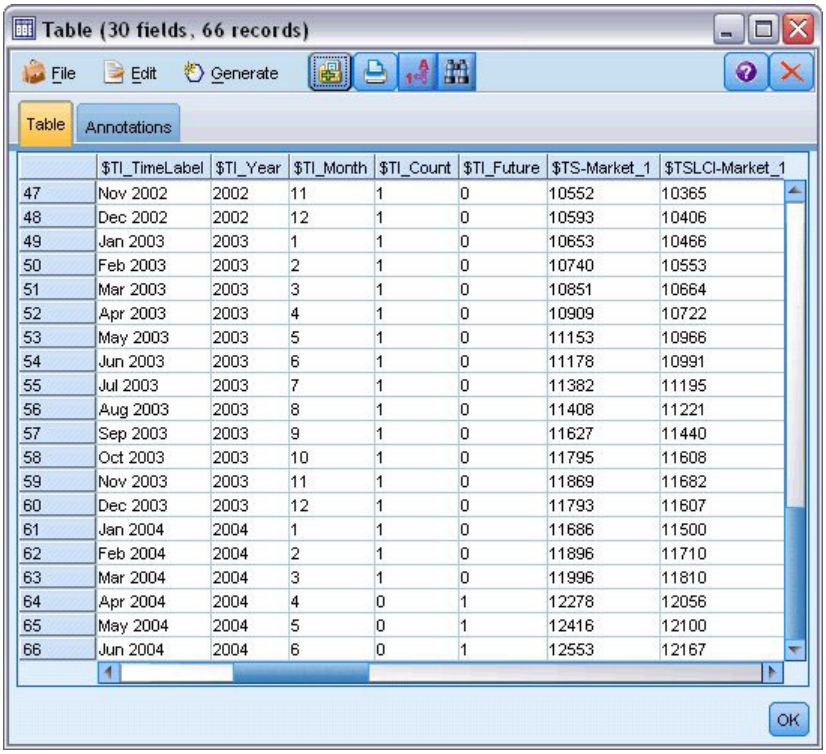


図 197. 時系列モデルに保存された設定の再利用

3. 時系列ノードを開きます。
4. 「**モデル**」タブで、「**既存のモデルを使用して推定を続行**」にチェック・マークが付いていることを確認します。
5. 「**実行**」をクリックして新しいモデル・ナゲットを領域および「モデル」パレットに配置します。

新しいモデルの検証



	$TI_TimeLabel	$TI_Year	$TI_Month	$TI_Count	$TI_Future	$TS-Market_1	$TSLCI-Market_1
47	Nov 2002	2002	11	1	0	10552	10365
48	Dec 2002	2002	12	1	0	10593	10406
49	Jan 2003	2003	1	1	0	10653	10466
50	Feb 2003	2003	2	1	0	10740	10553
51	Mar 2003	2003	3	1	0	10851	10664
52	Apr 2003	2003	4	1	0	10909	10722
53	May 2003	2003	5	1	0	11153	10966
54	Jun 2003	2003	6	1	0	11178	10991
55	Jul 2003	2003	7	1	0	11382	11195
56	Aug 2003	2003	8	1	0	11408	11221
57	Sep 2003	2003	9	1	0	11627	11440
58	Oct 2003	2003	10	1	0	11795	11608
59	Nov 2003	2003	11	1	0	11869	11682
60	Dec 2003	2003	12	1	0	11793	11607
61	Jan 2004	2004	1	1	0	11686	11500
62	Feb 2004	2004	2	1	0	11896	11710
63	Mar 2004	2004	3	1	0	11996	11810
64	Apr 2004	2004	4	0	1	12278	12056
65	May 2004	2004	5	0	1	12416	12100
66	Jun 2004	2004	6	0	1	12553	12167

図 198. 新しい予測を示しているテーブル

1. テーブル・ノードを、領域内の新しい時系列モデル・ナゲットに接続します。
2. テーブル・ノードを開いて、「**実行**」をクリックします。

   保存された設定を再利用しているため、新しいモデルでも 3 カ月先まで予測されます。ただし、推定期間 (時間区分ノードで指定) が 1 月ではなく 3 月に終了するため、今回は 4 月から 6 月までが予測されます。

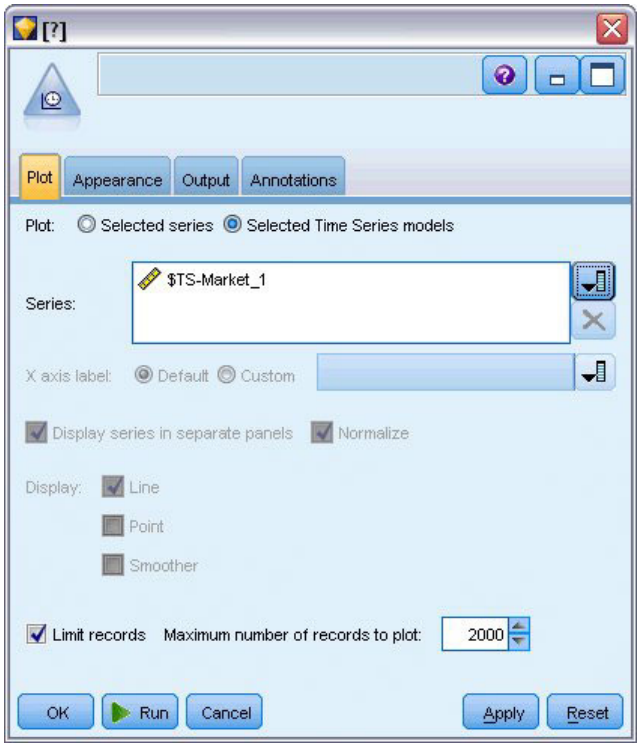


図 199. プロットするフィールドの指定

3. 時系列グラフ・ノードを時系列モデル・ナゲットに接続します。

   今回は、特に時系列モデル用に設計された時系列表示を使用します。

4. 「プロット」タブで、「**選択された時系列モデル**」オプションを選択します。
5. 「**系列**」リストで、フィールド選択ボタンをクリックし、「*$TS-Market_1*」フィールドを選択し、「**OK**」をクリックしてリストに追加します。
6. 「**実行**」をクリックします。

2004 年 3 月までの *Market_1* の実際の売り上げと、2004 年 6 月までの予測 (予想) 売り上げおよび信頼区分 (青色で網掛けされた領域) を示すグラフが作成されます。

最初の例と同様に、予測値はその期間中の実際のデータと密接しており、優れたモデルであることが示されています。

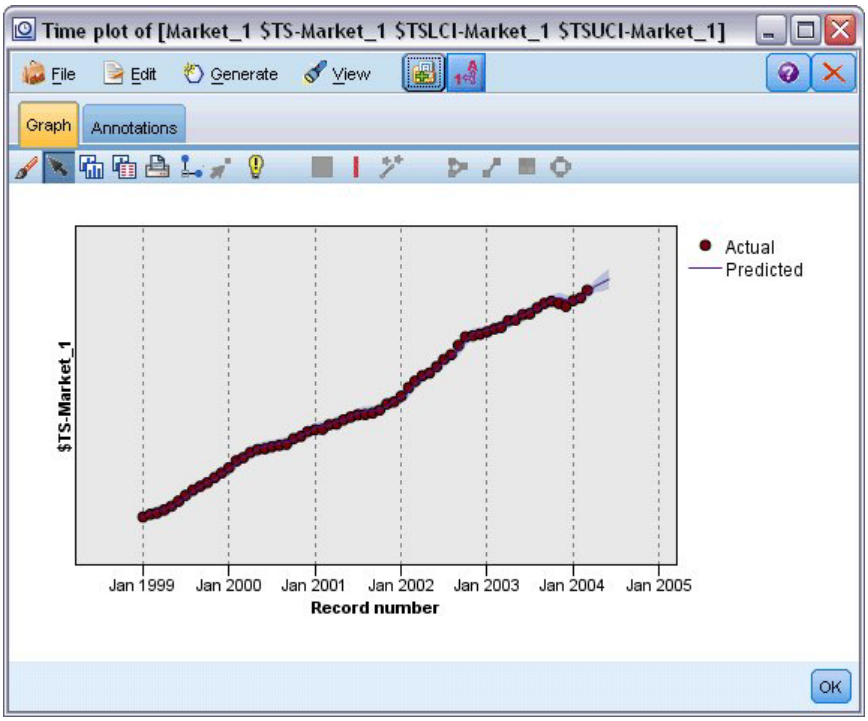


図 200. 6 月まで延長された予測

要約

より新しいデータが利用できるようになった場合に、保存したモデルを適用して以前の予測を拡張する方法について説明しました。さらに、モデルを再構築することなく実行することができました。もちろん、モデルが変化したと見なすべき理由があれば、そのモデルを再構築する必要があります。

第 15 章 カタログ販売の予測 (時系列)

カタログ会社は、過去 10 年間の販売データに基づく紳士服ラインの月間販売の予測に関心を示しています。

この例では、*catalog_seasfac.sav* というデータ・ファイルを参照する *catalog_forecast.str* というストリームを使用します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*catalog_forecast.str* ファイルは、*streams* ディレクトリー内にあります。

前述の例では、時系列に最適なモデルをエキスパート・モデルに決定させる方法を確認しました。ここでは、モデルを自分で選択する場合に利用できる 2 つの方法 (指数平滑法と ARIMA) について詳細に検討します。

適切なモデルを決定するために、まず時系列をプロットするのは良い考えです。時系列の目視検査は、多くの場合、選択を支援する強力な指針になります。特に、以下の点について自問する必要があります。

- 系列は全体的なトレンドを示しているか。当てはまる場合、そのトレンドは一定したものに見えるか、それとも時間の経過とともに減衰するように見えるか。
- 系列は季節性を示しているか。当てはまる場合、季節的変動は時間の経過とともに大きくなっているか、あるいは後続の期間でも一定しているように見えるか。

ストリームの作成

1. 新しいストリームを作成し、*catalog_seasfac.sav* を指し示す Statistics ファイル入力ノードを追加します。

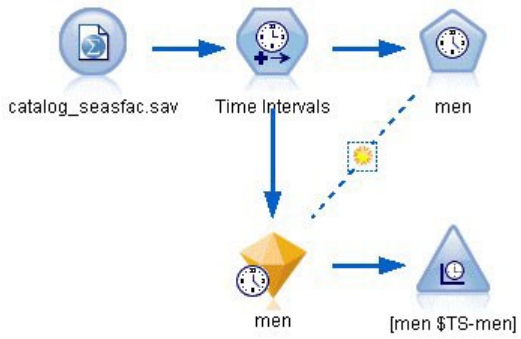


図 201. カタログ販売の予測

図 202. 対象フィールドの指定

2. IBM SPSS Statistics ファイル入力ノードを開き、「データ型」タブを選択します。
3. 「**値の読み込み**」をクリックしてから、「**OK**」をクリックします。
4. 「*男性*」フィールドの「*役割*」列をクリックし、役割を「**対象**」に設定します。
5. 他のすべてのフィールドの役割を「**なし**」に設定し、「**OK**」をクリックします。

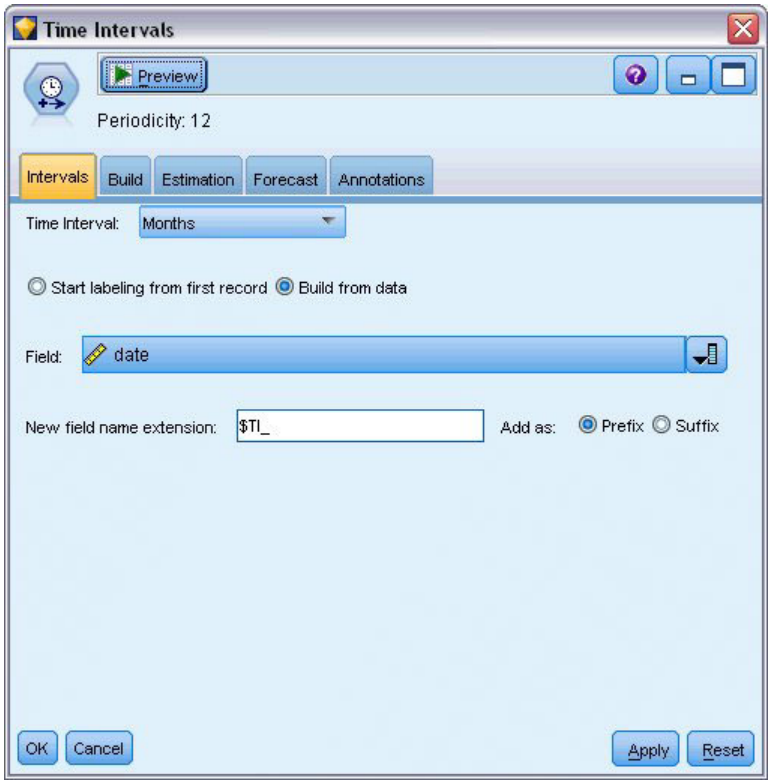


図 203. 時間区分の設定

6. IBM SPSS Statistics ファイル入力ノードに時間区分ノードを接続します。
7. 時間区分ノードを開いて、「**時間区分**」を「**月**」に設定します。
8. 「**データから構築**」を選択します。
9. 「**フィールド**」に「**日付**」を設定し、「**OK**」をクリックします。

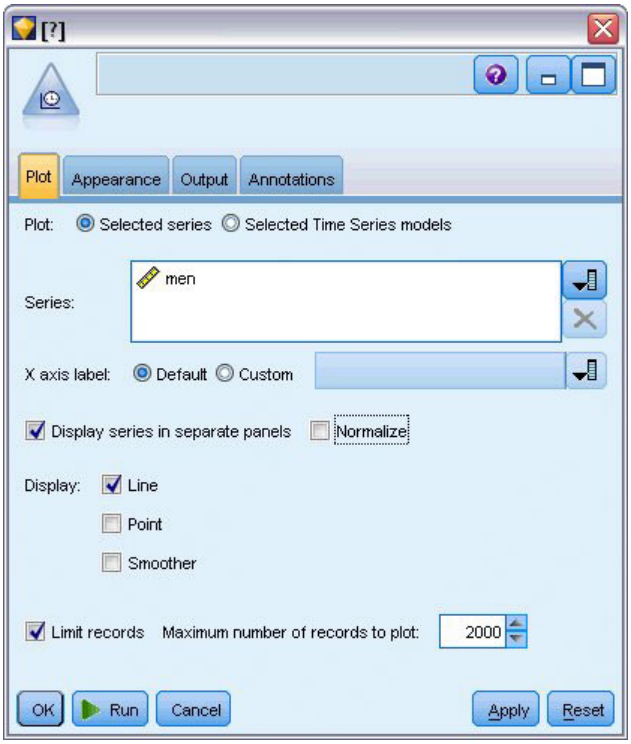


図 204. 時系列のプロット

10. 時系列ノードを時間区分ノードに接続します。
11. 「プロット」タブで、「**男性**」を「系列」リストに追加します。
12. 「**正規化**」チェック・ボックスの選択を解除します。
13. 「**実行**」をクリックします。

データの検証

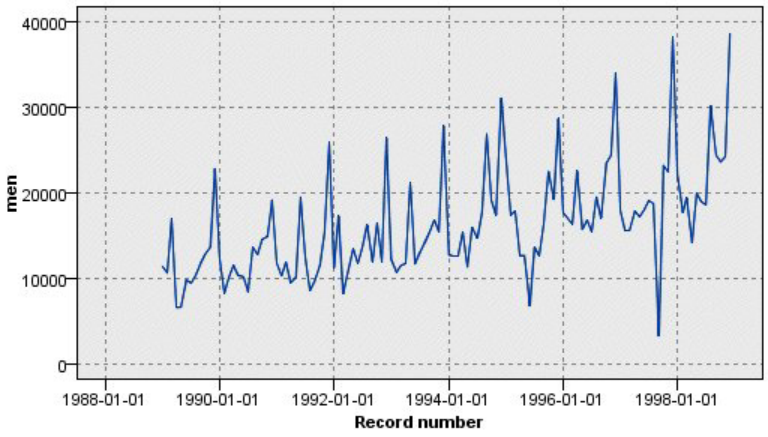


図 205. 紳士服の実際の売り上げ

この系列は全体的に上昇傾向を示しています。つまり、系列値は経時的に増加する傾向があります。上昇傾向は見たところ一定であり、線型トレンドを示しています。

また、この系列はグラフの垂直線で示されるように、毎年 12 月に売り上げが増えるという明確な季節性パターンも示しています。系列の上昇傾向とともに季節変動が大きくなる様子が見られ、相加的季節性ではなく相乗的季節性を示唆しています。

1. 「**OK**」をクリックしてプロットを閉じます。

系列の特性を確認したので、それをモデル化する準備ができました。トレンド、季節性、またはその両方を示す系列を予測する場合は、指数平滑法が役立ちます。既に確認したように、このデータは両方の特性を示しています。

指数平滑化

最適な指数平滑法モデルの構築では、モデル・タイプの判定 (そのモデルがトレンド、季節性、またはその両方を含む必要があるか) と、選択したモデルの最適なパラメーターの取得が行われます。

紳士服の経時的な売り上げのプロットでは、線型トレンド・コンポーネントおよび相乗的季節性コンポーネントの両方を持つモデルが示唆されました。これは Winters モデルを意味します。ただし、まず、単純モデル (トレンドおよび季節性なし) を調査し、次に Holt モデル(線型トレンドを導入、季節性なし) を調べます。これは、モデルがデータに適合しない場合を特定する練習になります (これはモデル構築に成功するには欠かせないスキルです)。

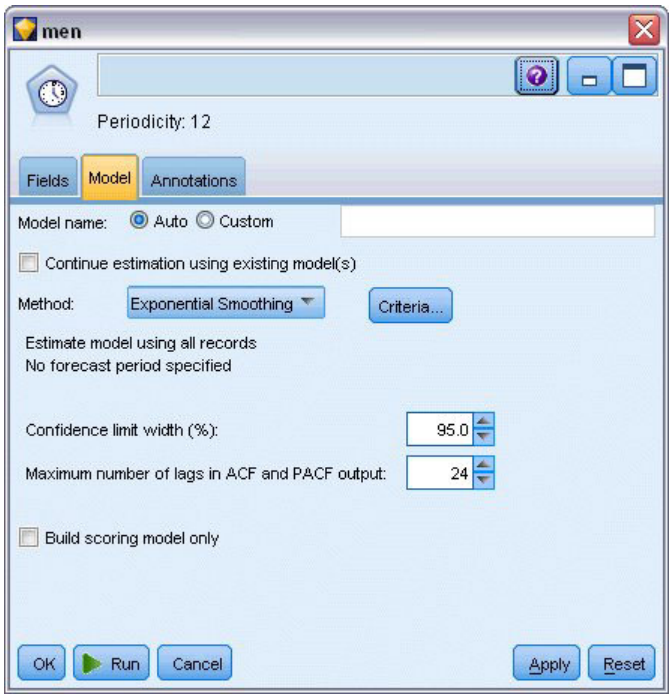


図 206. 指数平滑法の指定

単純な指数平滑法モデルから始めます。

1. 時系列ノードを時間区分ノードに接続します。
2. 「**モデル**」タブで、「**方法**」を「**指数平滑法**」に設定します。
3. 「**実行**」をクリックしてモデル・ナゲットを作成します。

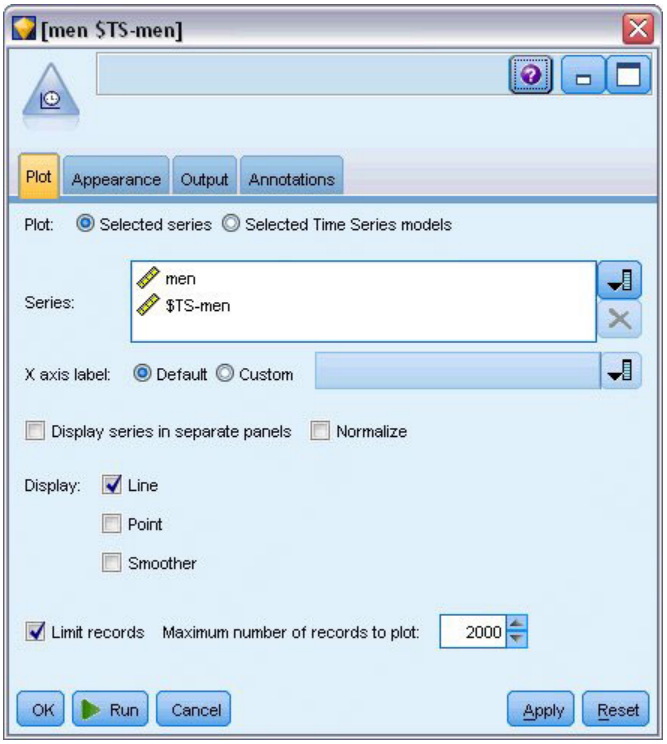


図 207. 時系列モデルのプロット

4. 時系列ノードをモデル・ナゲットに接続します。
5. 「**プロット**」タブで、「*男性*」および「*$TS-men*」を「**系列**」リストに追加します。
6. 「**別のパネルに時系列を表示**」および「**正規化**」チェック・ボックスの選択を解除します。
7. 「**実行**」をクリックします。

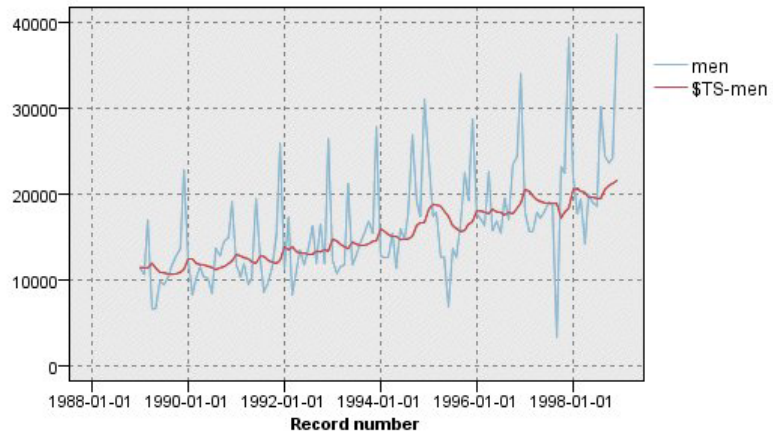


図 208. 単純な指数平滑法モデル

「**男性**」プロットは実際のデータを表し、「**$TS-men**」は時系列モデルです。

この単純モデルは、実際、漸進的な (かつ、やや重い) 上昇傾向を示しますが、季節性は考慮されていません。このモデルは問題なく却下できます。

8. 「**OK**」をクリックして時系列ウィンドウを閉じます。

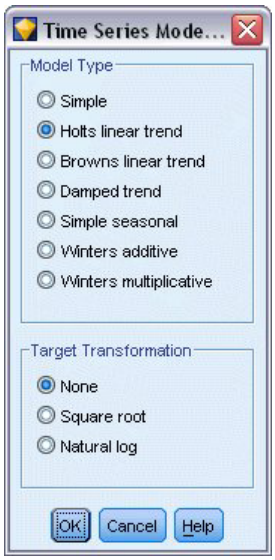


図 209. Holt のモデルの選択

Holt の線型モデルを試してみましょう。これも季節性を取得する可能性は低くなりますが、少なくとも単純モデルよりも適切にトレンドがモデル化されるはずです。

9. 時系列ノードを開き直します。
10. 「**モデル**」タブで、方法に「**指数平滑法**」を選択した状態で、「**基準**」をクリックします。
11. 「指数平滑法の基準」ダイアログ・ボックスで、「**Holts 線型トレンド**」を選択します。
12. 「**OK**」をクリックしてダイアログ・ボックスを閉じます。
13. 「**実行**」をクリックして、モデル・ナゲットを再作成します。
14. 時系列ノードを開き直し、「**実行**」をクリックします。

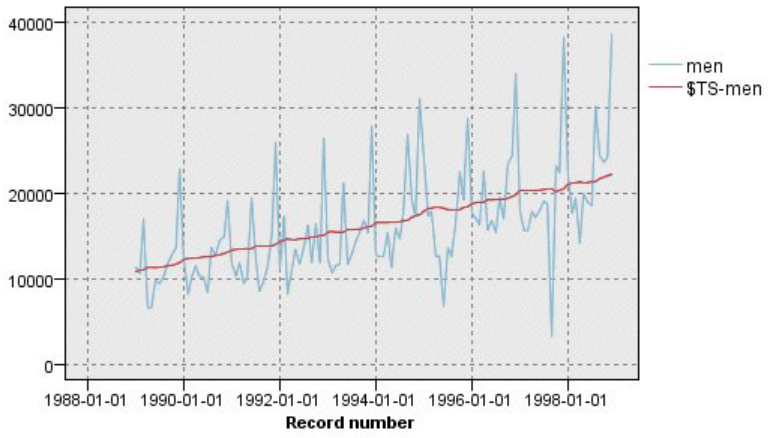


図 210. Holt の線型トレンド・モデル

Holt のモデルは単純モデルよりも滑らかな上昇傾向を示しますが、やはり季節性は考慮されないため、このモデルも破棄できます。

15. 時系列ウィンドウを閉じます。

紳士服の経時的な売り上げの初期プロットで、線型トレンドと相乗的季節性を統合したモデルが示唆されたことを思い出してください。したがって、Winters のモデルが候補として適していると考えられます。

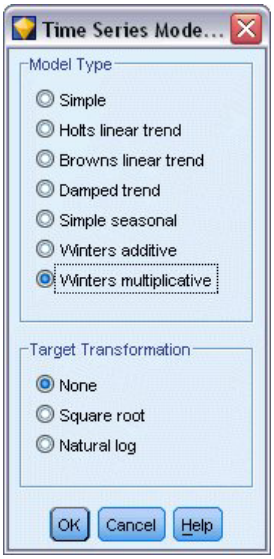


図 211. Winters のモデルの選択

16. 時系列ノードを開き直します。
17. 「**モデル**」タブで、方法に「**指数平滑法**」を選択した状態で、「**基準**」をクリックします。
18. 「指数平滑法の基準」ダイアログ・ボックスで、「**Winters 相乗モデル**」を選択します。
19. 「**OK**」をクリックしてダイアログ・ボックスを閉じます。
20. 「**実行**」をクリックして、モデル・ナゲットを再作成します。
21. 時系列ノードを開き、「**実行**」をクリックします。

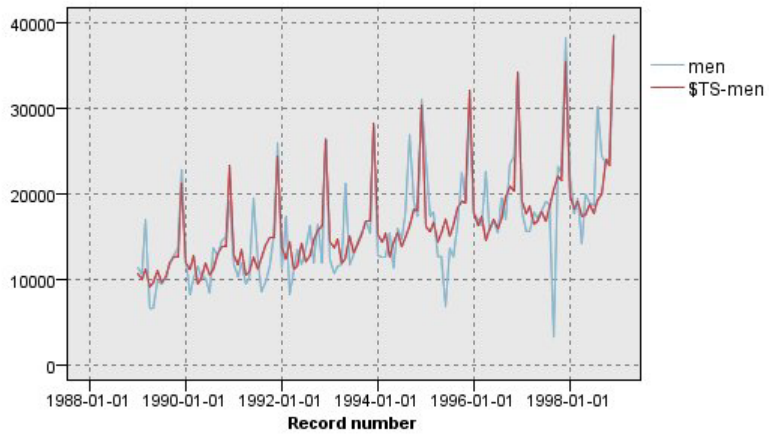


図 212. Winters の相乗モデル

前よりも良好になりました。モデルはデータのトレンドと季節性の両方を反映しています。

このデータ・セットは 10 年間のものです。10 の季節性ピークが見られ、それは毎年 12 月に発生しています。予測結果に示された 10 のピークは、実際のデータで年に一度発生する 10 のピークとよく合致しています。

ただし、この結果により、指数平滑法手順の限界も浮き彫りになっています。上昇と下降の両方の山形部分に注目すると、考慮されていない有意な構造があることが分かります。

主として季節変動による長期トレンドのモデル化に関心がある場合は、指数平滑法を選択するといいでしょう。このケースのように比較的複雑な構造をモデル化する場合は、ARIMA 手順の使用を検討する必要があります。

ARIMA

ARIMA 手順では、時系列の微調整されたモデル作成に適した自己回帰和分移動平均 (ARIMA) モデルを作成できます。ARIMA モデルには、トレンドおよび季節性コンポーネントのモデル作成用に、指数平滑法モデルよりも洗練された方法が提供されています。また、モデルに予測変数を含めることができる利点もあります。

引き続き予測モデルを開発する必要があるカタログ会社の例について検討します。この会社が紳士服の月次売上データと、売り上げの変動の説明に役立つ可能性があるいくつかの系列をどのように収集したかについては、既に説明しました。考えられる予測値には、郵送したカタログ数、カタログのページ数、受注用の電話回線の数、広告印刷物の費用、および顧客サービス担当者の数があります。

これらの中に、予測に有効な予測値はあるでしょうか。予測値を持つモデルは、予測値を持たないモデルよりも実際に優れているでしょうか。ARIMA 手順を使用すると、予測値を持つ予測モデルを作成して、予測機能において、予測値を持たない指数平滑法モデルを上回る有意差があるかどうかを確認できます。

ARIMA 法では、自己回帰、差分、および移動平均の順序と、これらのコンポーネントに対する季節性の同等物の順序も指定することによりモデルを微調整できます。これらのコンポーネントの最適値を手動で決定するのは、多大な試行錯誤を伴う時間のかかるプロセスになるため、この例については、ARIMA モデルの選択をエキスパート・モデラーにまかせます。

データ・セット内のその他の変数を予測変数として扱うことにより、より優れたモデルの構築を試みます。予測値として含めると最も役立つと思われるものは、郵送したカタログ数 (*郵送*)、カタログのページ数 (ページ) 、受注用の電話回線の数 (*電話*)、広告印刷物の費用 (*印刷*)、および顧客サービス担当者の数 (サービス) です。

図 213. 予測値フィールドの設定

1. IBM SPSS Statistics ファイル入力ノードを開きます。
2. 「データ型」タブで、「*郵送*」、「*ページ*」、「*電話*」、「*印刷*」、および「サービス」の「*役割*」を「**入力**」に設定します。
3. 「**男性**」の役割が「**対象**」に設定され、その他のフィールドが「**なし**」に設定されていることを確認します。
4. 「**OK**」をクリックします。

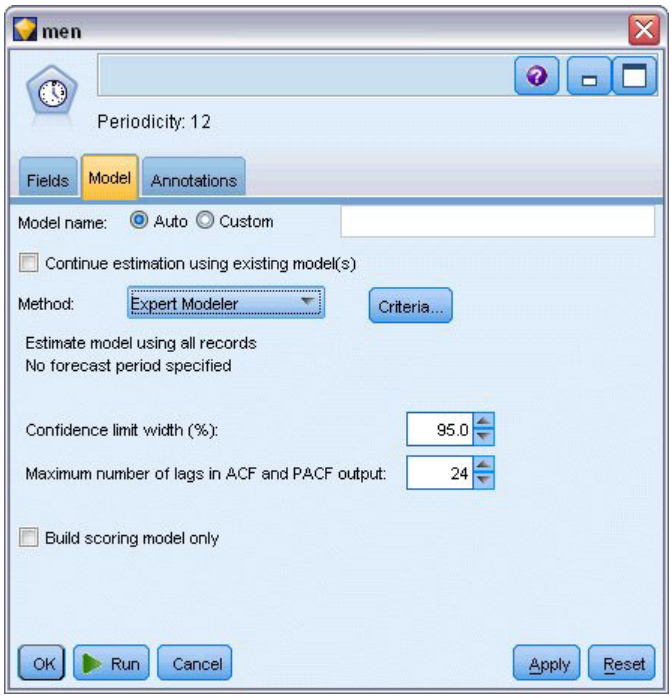


図 214. エキスパート・モデラーの選択

5. 時系列ノードを開きます。
6. 「モデル」タブで、「**方法**」を「**エキスパート・モデラー**」に設定し、「**基準**」をクリックします。

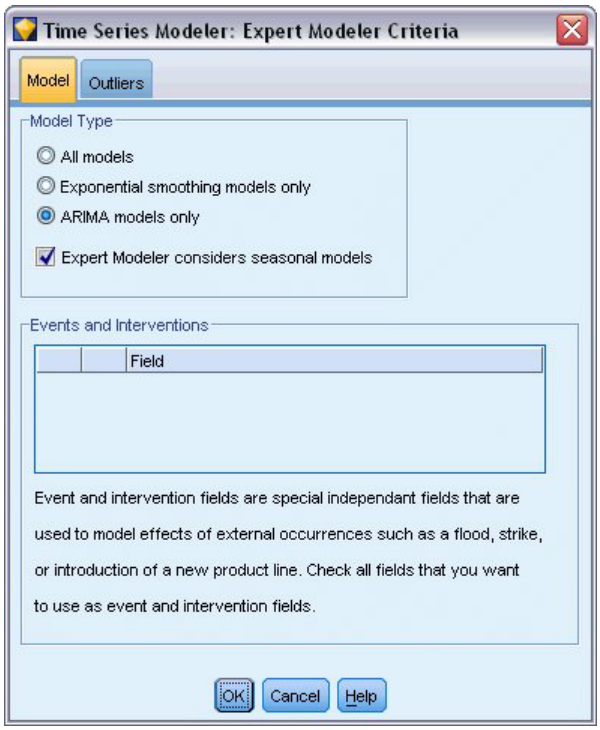


図 215. ARIMA モデルのみの選択

7. 「エキスパート・モデラーの基準」ダイアログ・ボックスで、「**ARIMA モデルのみ**」オプションを選択し、「**エキスパート・モデラーが季節性モデルを検討**」にチェック・マークが付いていることを確認します。
8. 「**OK**」をクリックしてダイアログ・ボックスを閉じます。
9. 「モデル」タブで「**実行**」をクリックしてモデル・ナゲットを再作成します。

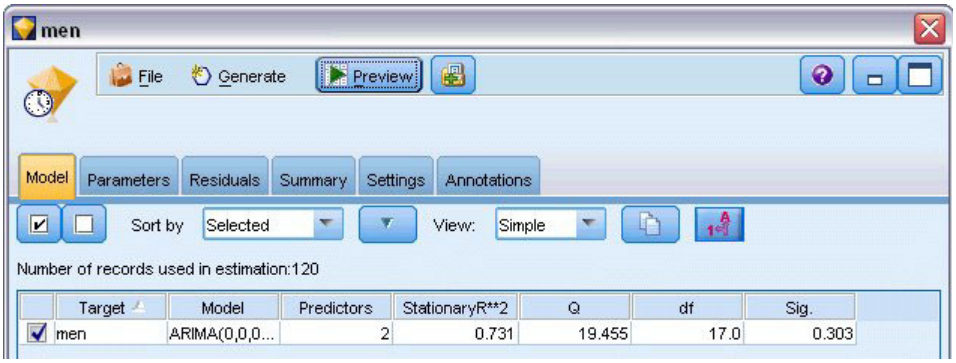


図 216. エキスパート・モデラーによる 2 種類の予測値の選択

10. モデル・ナゲットを開きます。

    エキスパート・モデラーが 5 種類の指定予測値のうち、どのように 2 種類のみをこのモデルに有意として選択したかに注目してください。
11. 「**OK**」をクリックしてモデル・ナゲットを閉じます。
12. 時系列ノードを開き、「**実行**」をクリックします。

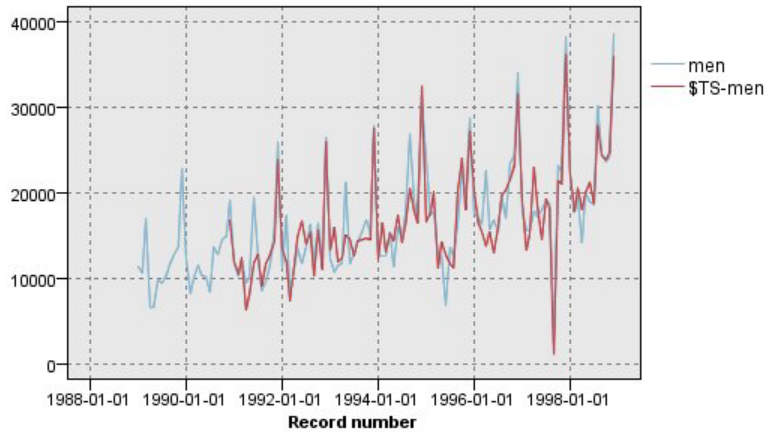


図 217. 指定予測値を使用した ARIMA モデル

このモデルでは、大きな下降の山形部分を取得することで以前のモデルが改善されており、これまでで最適な状態になっています。

このモデルをさらに洗練できますが、この時点からの改善は最小限になると予想されます。予測値を持つ ARIMA モデルが望ましいことは立証したので、構築したばかりのモデルを使用してみましょう。この例の目的である来年の売り上げを予測します。

13. 「**OK**」をクリックして時系列ウィンドウを閉じます。

14. 時間区分ノードを開き、「*予測*」タブを選択します。
15. 「*レコードの将来への拡張*」チェック・ボックスを選択し、値を 12 に設定します。

    予測する際に予測値を使用するには、モデラーが対象フィールドをより正確に予測できるように、予測期間内のそれらのフィールドに対して推定値を指定する必要があります。

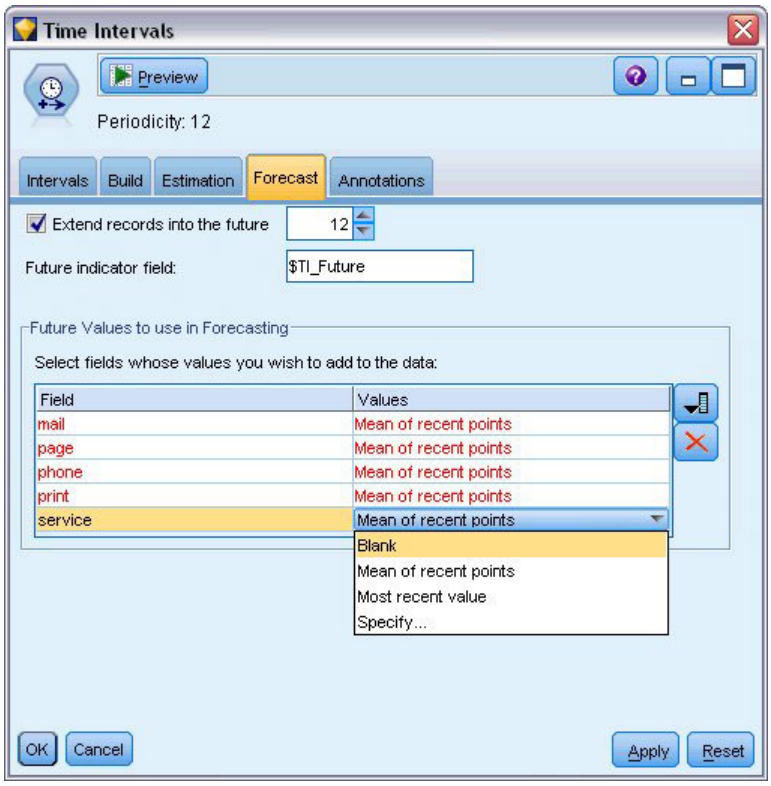


図 218. 予測フィールドに対する将来の値の指定

16. 「**予測で使用する将来の値**」グループで、「値」列の右側のフィールド選択ボタンをクリックします。
17. 「フィールドの選択」ダイアログ・ボックスで、「**郵送**」から「**サービス**」までを選択し、「**OK**」をクリックします。

    実世界では、これらの 5 種類の予測値はすべて管理された項目に関係しているため、この時点で将来の値を手動で指定します。この例の目的に合わせて、定義済みの関数の 1 つを使用し、それぞれの予測値に対して 12 個の値を指定する手間を省きます。(この例に慣れたら、さまざまな将来値で実験して、モデルに与える影響を調べるといいでしょう。)
18. それぞれのフィールドについて順番に、「**値**」フィールドをクリックして指定可能な値のリストを表示し、「**最近使用したポイントの平均**」を選択します。このオプションでは、このフィールドの最後の 3 個のデータ・ポイントの平均を計算し、それぞれの場合の推定値として使用します。
19. 「**OK**」をクリックします。
20. 時系列ノードを開き、「**実行**」をクリックしてモデル・ナゲットを再作成します。
21. 時系列ノードを開き、「**実行**」をクリックします。

1999 年の予測は良好のようです。期待どおり、12 月のピークの後は通常の売上レベルに戻り、その年の下半期では安定した上昇傾向となり、全般的に前年を大きく上回る売り上げとなりました。

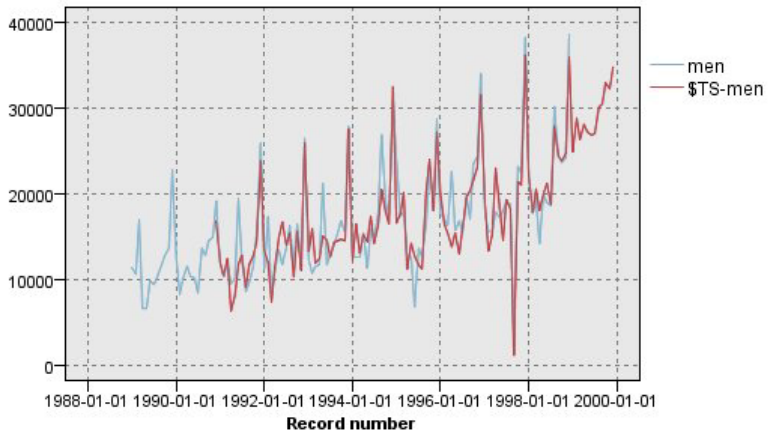


図 219. 指定予測値による売り上げ予測

要約

増加トレンドだけでなく季節変動やその他の変動も組み込まれている複雑な時系列を問題なくモデル化できました。また、試行錯誤により、正確なモデルに徐々に近づくことができることを確認し、将来の売り上げを予測するのに使用しました。

実際には、実際の販売データが更新された場合は (例えば、月ごとあるいは四半期ごと)、モデルを再適用して更新した予測を作成する必要があります。詳細については、 173 ページの『時系列モデルの再適用』を参照してください。

第 16 章 顧客への提案 (自己学習)

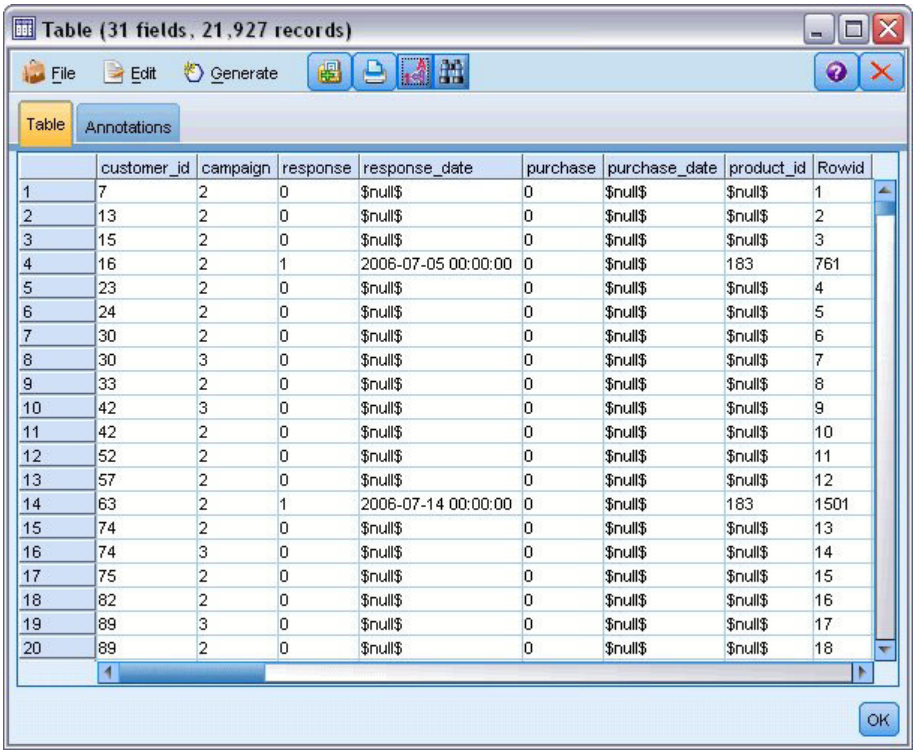
自己学習応答モデル (SLRM) ノードは、顧客に最も適している提案およびその提案が受け入れられる確率を予測できるモデルを生成し、そのモデルの更新を可能にします。このような種類のモデルは、マーケティング・アプリケーションやコール・センターなどのカスタマー・リレーションシップ・マネジメントで最も役立ちます。

この例は、架空の銀行に基づいています。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金融サービスを提案することで、今後のキャンペーンでさらに収益の高い結果を実現することを望んでいます。具体的には、この例では、自己学習応答モデルを使用して、以前の提案および応答を基に、好意的な反応を示す可能性が最も高い顧客の特徴を識別し、その結果に基づいて現在の最良の提案を促進します。

この例では、データ・ファイル *pm_customer_train1.sav*、*pm_customer_train2.sav*、および *pm_customer_train3.sav* を参照するストリーム *pm_selflearn.str* を使用します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* フォルダーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*pm_selflearn.str* ファイルは *streams* フォルダー内にあります。

既存データ

この会社には、過去のキャンペーンで顧客に行った提案、およびそれらの提案に対する応答を追跡する履歴データがあります。これらのデータには、さまざまな顧客の応答率を予測するために使用できる人口統計および財務情報も含まれています。



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

図 220. 以前の提案に対する応答

ストリームの構築

1. IBM SPSS Modeler インストール環境の *Demos* フォルダーにある *pm_customer_train1.sav* を指し示す Statistics ファイル入力ノードを追加します。

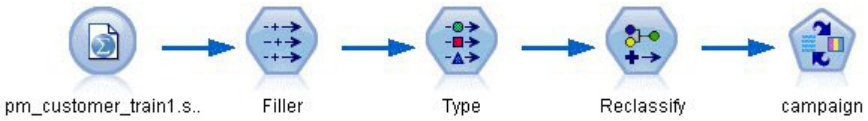


図 221. SLRM サンプル・ストリーム

2. 置換ノードを追加し、対象フィールドとして `campaign` を選択します。
3. 置換タイプ「**常時**」を選択します。
4. 「置換」テキスト・ボックスに、`to_string(campaign)` と入力し、「**OK**」をクリックします。

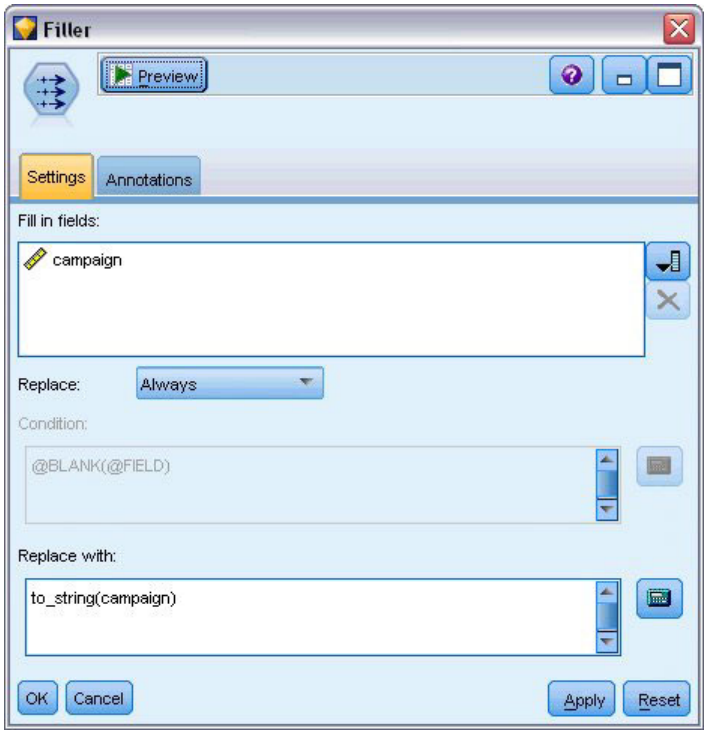


図 222. キャンペーン・フィールドの派生

5. データ型ノードを追加し、*customer_id*、*response_date*、*purchase_date*、*product_id*、*Rowid*、および *X_random* の各フィールドの「*役割*」を「**なし**」に設定します。

図 223. データ型ノードの設定変更

6. *campaign* フィールドおよび *response*フィールドの「*役割*」を「**対象**」に設定します。これらは、予測の基本となるフィールドです。

response フィールドの「**測定**」を「**フラグ型**」に設定します。

7. 「**値の読み込み**」をクリックしてから、「**OK**」をクリックします。

   キャンペーン・フィールドのデータが数字のリスト (1、2、3、4) として表示されるため、フィールドを再分類してより分かりやすいタイトルにすることができます。

8. データ分類ノードをデータ型ノードに追加します。
9. 「**データ分類先**」フィールドで、「**既存フィールド**」を選択します。
10. 「**データ分類フィールド**」リストで、**campaign** を選択します。
11. 「**取得**」ボタンをクリックします。キャンペーンの値が「*元の値*」列に追加されます。
12. 「*新しい値*」列の最初の 4 行に、次のキャンペーン名を入力します。
    - **住宅ローン**
    - **カー・ローン**
    - **貯金**
    - **年金**
13. 「**OK**」をクリックします。

図 224. キャンペーン名のデータ分類

14. SLRM モデル作成ノードをデータ分類ノードに接続します。「フィールド」タブで、対象フィールドに **campaign** を選択し、対象応答フィールドに **response** を選択します。

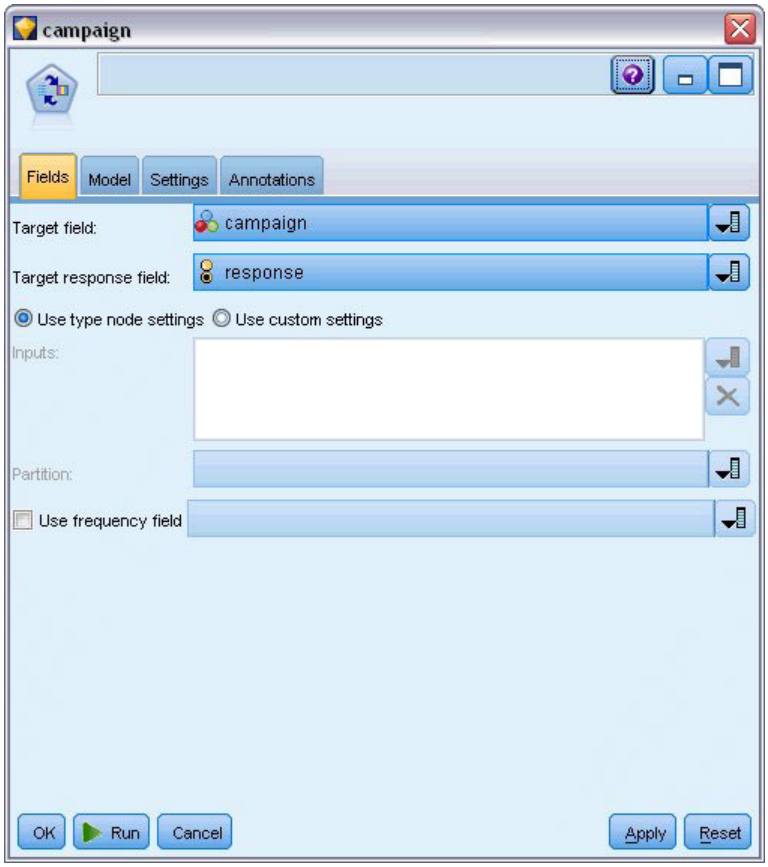


図 225. 対象および対象応答の選択

15. 「設定」タブの「レコードあたりの最大予測数」フィールドで数を 2 に減少させます。

    これにより、各顧客に対し、受け入れられる確率が最も高いと特定された 2 つの提案が存在することになります。

16. 「**モデルの信頼性を考慮**」が選択された状態にし、「**実行**」をクリックします。

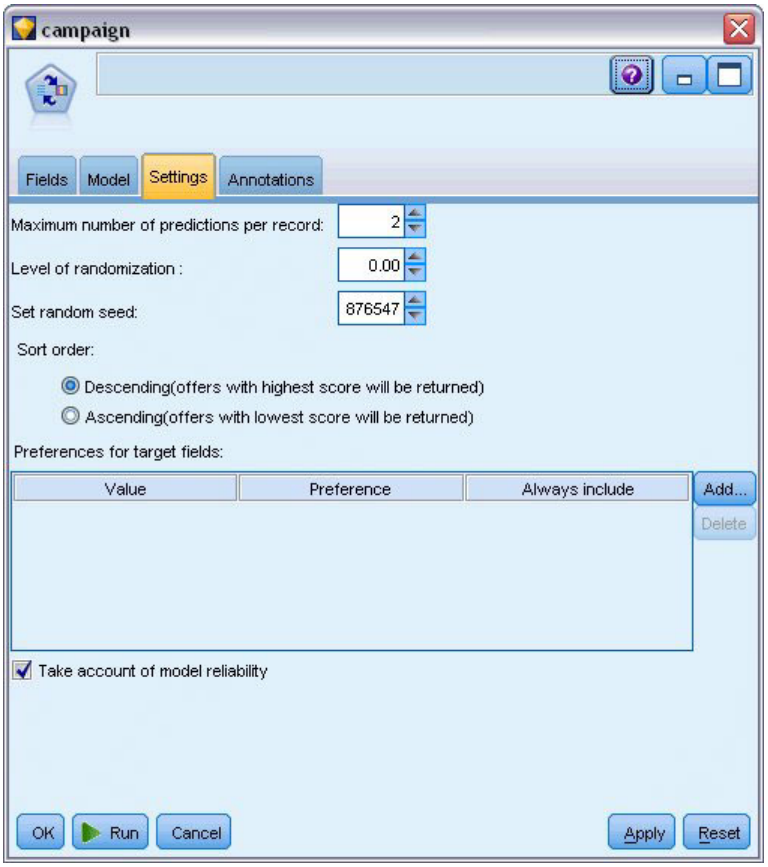


図 226. SLRM ノード設定

モデルの参照

1. モデル・ナゲットを開きます。「モデル」タブには最初、各提案の予測の推定精度およびモデル推定時の各予測の相対重要度が表示されます。

   各予測値と対象変数の相関を表示するには、右側のペインの「**表示**」リストから 「**応答との関連**」を選択します。
2. 予測がある 4 つの各提案の間で表示を切り替えるには、左側のペインにある「**表示**」リストから必要な提案を選択します。

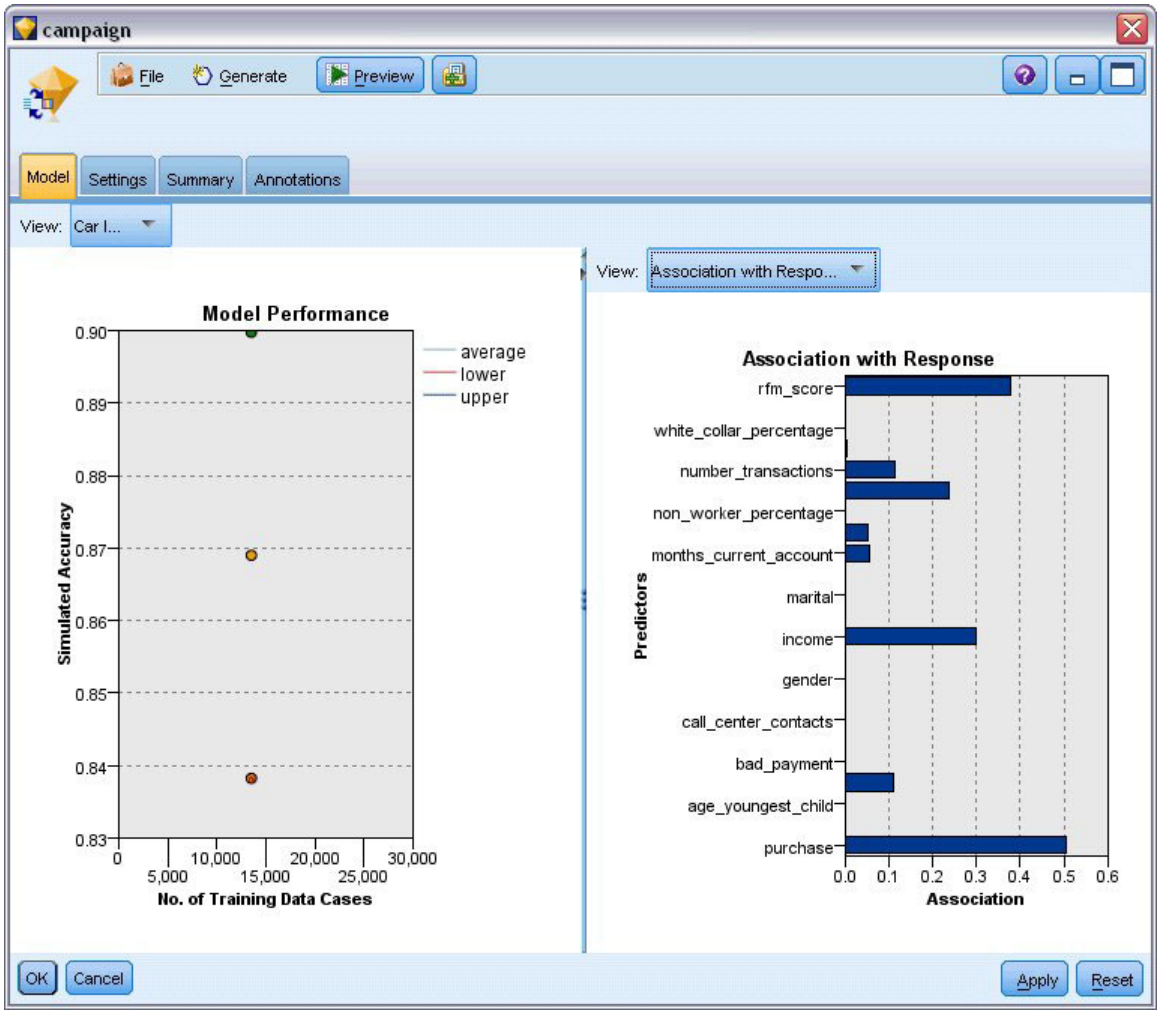


図 227. SLRM モデル・ナゲット

3. モデル・ナゲット・ウィンドウを閉じます。
4. ストリーム領域で、*pm_customer_train1.sav* を示す IBM SPSS Statistics ファイル入力ノードの接続を解除します。
5. IBM SPSS Modeler インストール環境の *Demos* フォルダーにある *pm_customer_train2.sav* を指し示す Statistics ファイル入力ノードを追加し、それを置換ノードに接続します。

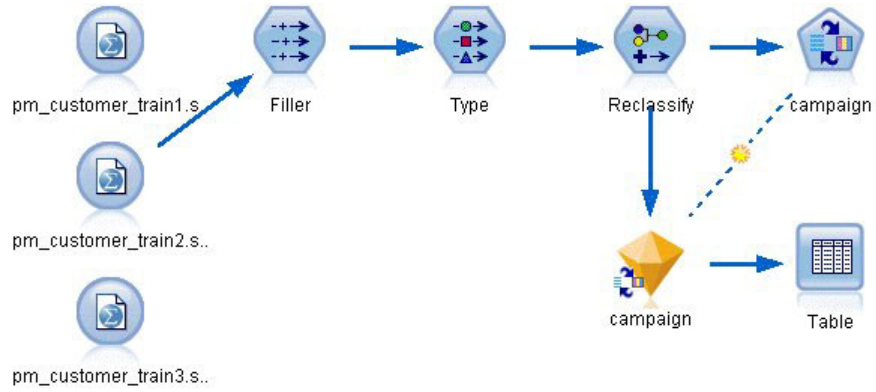


図 228. SLRM ストリームへの 2 番目のデータ・ソースの接続

6. SLRM ノードの「モデル」タブで、「**既存モデルの学習を継続**」を選択します。

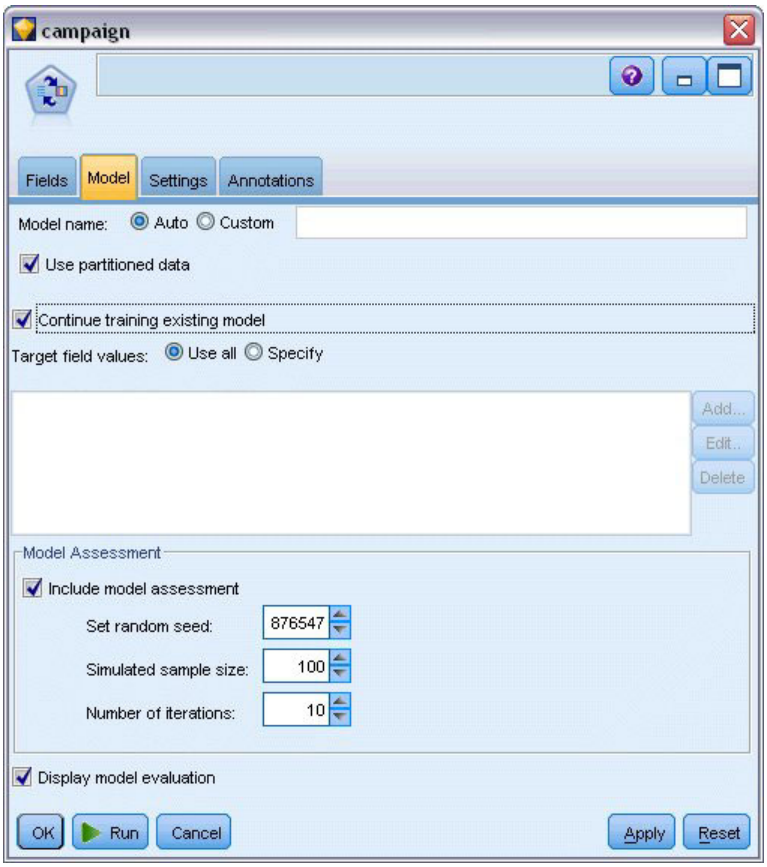


図 229. モデル学習の継続

7. 「**実行**」をクリックして、モデル・ナゲットを再作成します。詳細を表示するには、領域のナゲットをダブルクリックします。

   これで、「モデル」タブには、各提案の予測の精度の、改訂済み推定値が表示されます。

8. IBM SPSS Modeler インストール環境の *Demos* フォルダーにある *pm_customer_train3.sav* を指し示す Statistics ファイル入力ノードを追加し、それを置換ノードに接続します。

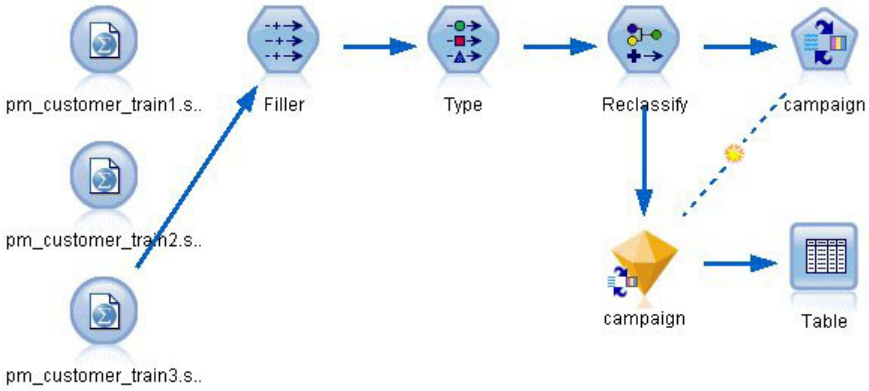


図 230. SLRM ストリームへの 3 番目のデータ・ソースの接続

9. 「**実行**」をクリックして、もう一度モデル・ナゲットを再作成します。詳細を表示するには、領域のナゲットをダブルクリックします。
10. これで、「モデル」タブには、各提案の予測の最終推定精度が表示されます。

    表示されているとおり、平均精度は、データ・ソースを追加すると若干減少します (86.9% から 85.4%)。ただし、この変動は最小限であり、利用可能なデータ内のわずかな異常値に起因している可能性があります。

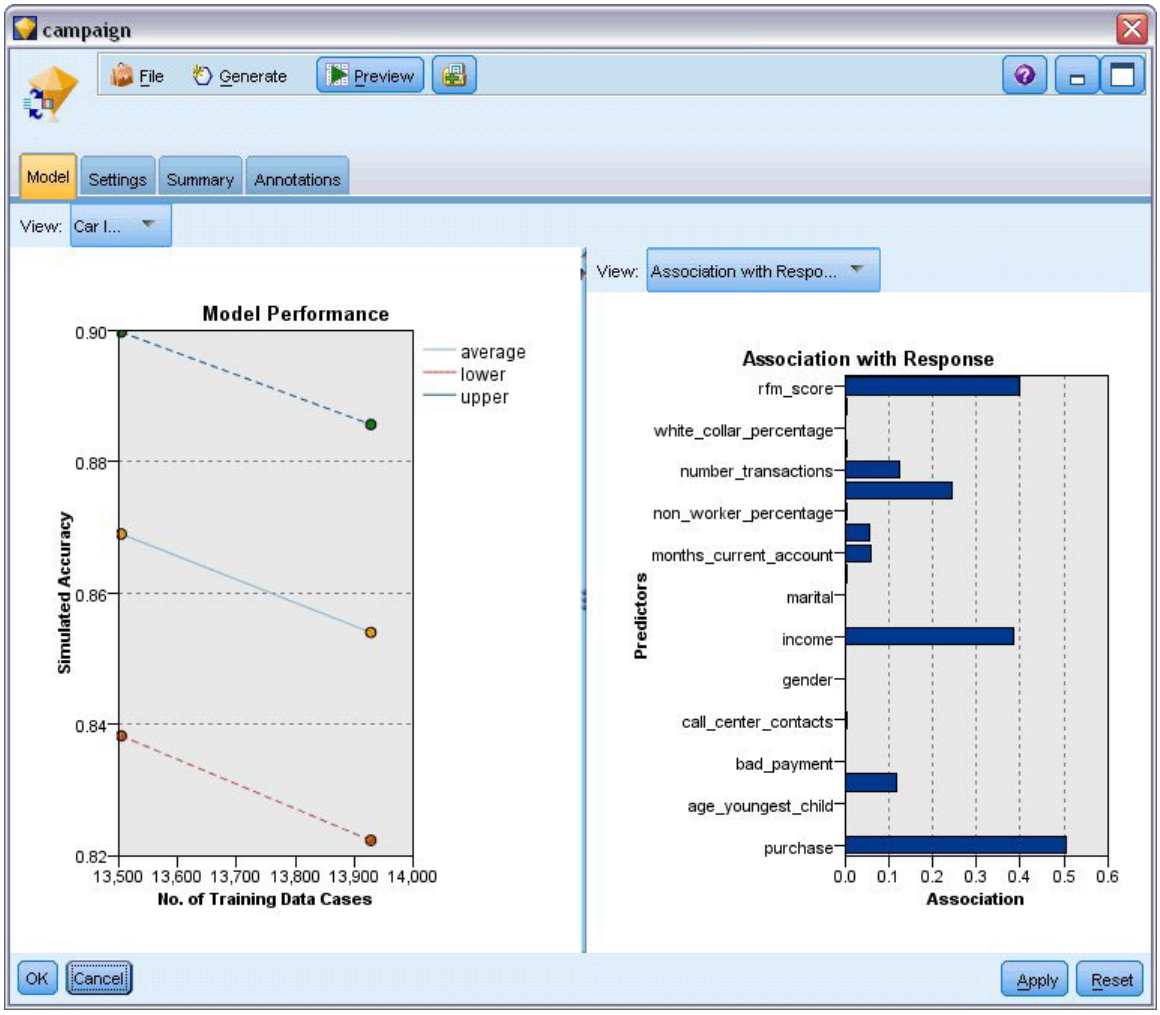
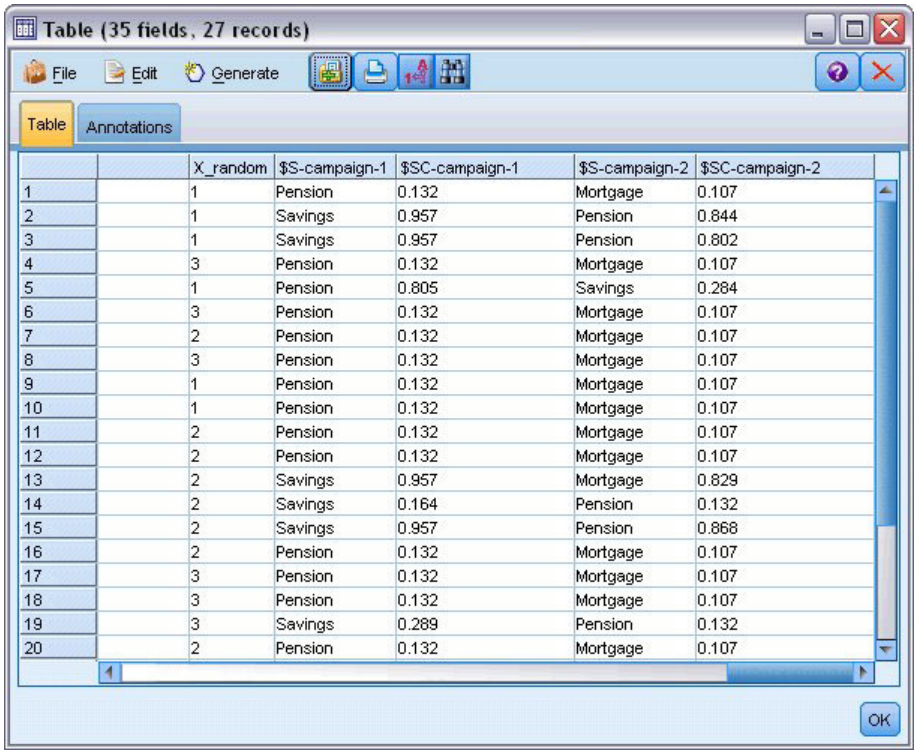


図 231. 更新された SLRM モデル・ナゲット

11. テーブル・ノードを最後 (3 番目) に生成されたモデルに接続し、テーブル・ノードを実行します。
12. テーブルを右へスクロールします。予測では、各顧客の詳細に応じて、顧客が受け入れる可能性が最も高い提案、およびその提案を受け入れる信頼度が示されています。

例えば、表示されたテーブルの 1 行目では、以前カー・ローンを利用した顧客が年金を提案した場合に受け入れる信頼度はわずか 13.2% (*$SC-campaign-1* 列の値 0.132 で表される) です。しかし、2 行目および 3 行目では、同様にカー・ローンを利用した顧客がさらに 2 人表示されています。これらの顧客のケースでは、これらの顧客および同じような履歴の他の顧客が預金口座の提案を受けた場合に預金口座を開設する信頼度は 95.7%です。また、年金を受け入れる信頼度は 80%を超えています。



		X_random	$S-campaign-1	$SC-campaign-1	$S-campaign-2	$SC-campaign-2
1		1	Pension	0.132	Mortgage	0.107
2		1	Savings	0.957	Pension	0.844
3		1	Savings	0.957	Pension	0.802
4		3	Pension	0.132	Mortgage	0.107
5		1	Pension	0.805	Savings	0.284
6		3	Pension	0.132	Mortgage	0.107
7		2	Pension	0.132	Mortgage	0.107
8		3	Pension	0.132	Mortgage	0.107
9		1	Pension	0.132	Mortgage	0.107
10		1	Pension	0.132	Mortgage	0.107
11		2	Pension	0.132	Mortgage	0.107
12		2	Pension	0.132	Mortgage	0.107
13		2	Savings	0.957	Mortgage	0.829
14		2	Savings	0.164	Pension	0.132
15		2	Savings	0.957	Pension	0.868
16		2	Pension	0.132	Mortgage	0.107
17		3	Pension	0.132	Mortgage	0.107
18		3	Pension	0.132	Mortgage	0.107
19		3	Savings	0.289	Pension	0.132
20		2	Pension	0.132	Mortgage	0.107

図 232. モデル出力 - 予測された提案と信頼度

IBM SPSS Modeler で使用されるモデル作成方法の数学的な基礎の説明については、製品のDVDの *¥Documentation* ディレクトリーにある「*IBM SPSS Modeler アルゴリズム・ガイド*」を参照してください。

これらの結果は学習データのみに基づくことに注意してください。モデルが実際の世界の他のデータにどれだけうまく一般化されるかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。

第 17 章 ローン返済不能の予測 (ベイズ・ネットワーク)

ベイズ・ネットワークを使用すると、観測された情報および記録された情報を「常識」という実際の知識と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。

この例では、*bankloan.sav* というデータ・ファイルを参照する *bayes_bankloan.str* というストリームを使用します。これらのファイルは IBM SPSS Modeler インストール環境の *Demos* ディレクトリーから使用でき、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することができます。*bayes_bankloan.str* ファイルは、*streams* ディレクトリー内にあります。

例えば、ローンが返済されない可能性について銀行が懸念しているとします。以前のローン返済不能データを使用してローン返済について問題が生じそうな潜在的顧客を予測することができる場合、こうした「リスクのある」顧客に対しては、融資を拒否するか、代わりの製品を提案することができます。

この例は、既存のローン返済不能データを使用して将来の返済不能者の可能性を予測することに焦点を当て、3 つの異なるベイズ・ネットワーク・モデル・タイプに注目してこの状況を予測するのにどのタイプが良いかを確定します。

ストリームの構築

1. *Demos* フォルダーの *bankloan.sav* を指し示す Statistics ファイル入力ノードを追加します。

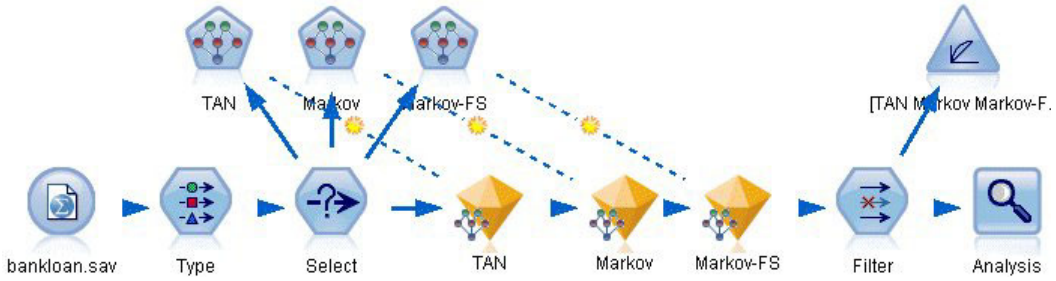


図 233. ベイズ・ネットワークのサンプル・ストリーム

2. データ型ノードを入力ノードに追加し、「**デフォルト**」フィールドの役割を「**対象**」に設定します。その他のフィールドの役割は、すべて「**入力**」に設定します。
3. 「**値の読み込み**」ボタンをクリックして、「*値*」列にデータを取り込みます。

図 234. 対象フィールドの選択

対象が Null 値を持つケースはモデルの構築に役に立ちません。こうしたケースを除外して、モデルの評価に使用されないようにすることができます。

4. 条件抽出ノードをデータ型ノードに追加します。
5. モードで「**破棄**」を選択します。
6. 「条件」ボックスに **default = '$null$'** と入力します。

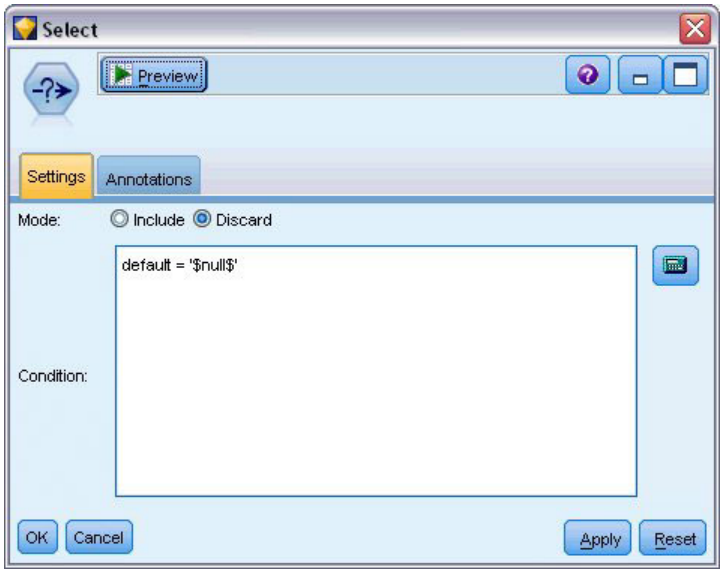


図 235. Null 値を持つ対象の破棄

さまざまなタイプのベイズ・ネットワークを構築できるため、複数のモデルを比較して最善の予測を提供するモデルを確認する価値があります。Tree Augmented Naive Bayes (TAN) モデルを最初に作成します。

7. ベイズ・ネットワーク・ノードを条件抽出ノードに接続します。
8. 「モデル」タブのモデル名で、「**カスタム**」を選択し、テキスト・ボックスに TAN と入力します。
9. 構造タイプに「**TAN**」を選択し、「**OK**」をクリックします。

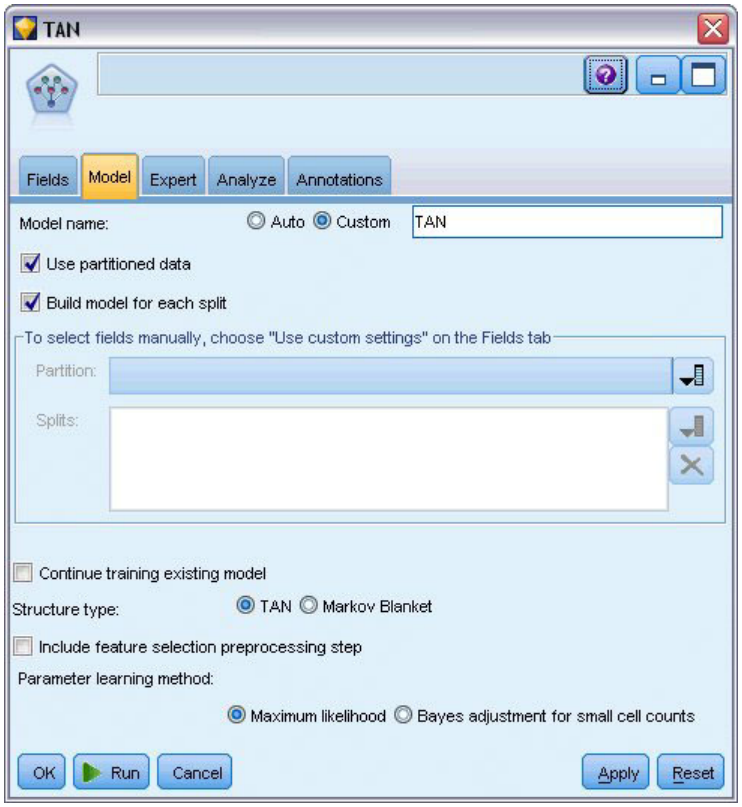


図 236. Tree Augmented Naive Bayes モデルの作成

2 番目に構築するモデル・タイプは Markov Blanket 構造になっています。

10. 2 番目のベイズ・ネットワーク・ノードを条件抽出ノードに接続します。
11. 「モデル」タブのモデル名で、「**カスタム**」を選択し、テキスト・ボックスに Markov と入力します。
12. 構造タイプに「**Markov Blanket**」を選択し、「**OK**」をクリックします。

図 237. Markov Blanket モデルの作成

3 番目に構築するモデル・タイプは Markov Blanket 構造になっており、またフィールド選択事前処理を使用して、対象変数に大きく関連している入力を選択します。

13. 3 番目のベイズ・ネットワーク・ノードを条件抽出ノードに接続します。
14. 「モデル」タブのモデル名で、「**カスタム**」を選択し、テキスト・ボックスに `Markov-FS` と入力します。
15. 構造タイプに「**Markov Blanket**」を選択します。
16. 「**フィールド選択の事前処理ステップを含む**」を選択し、「**OK**」をクリックします。

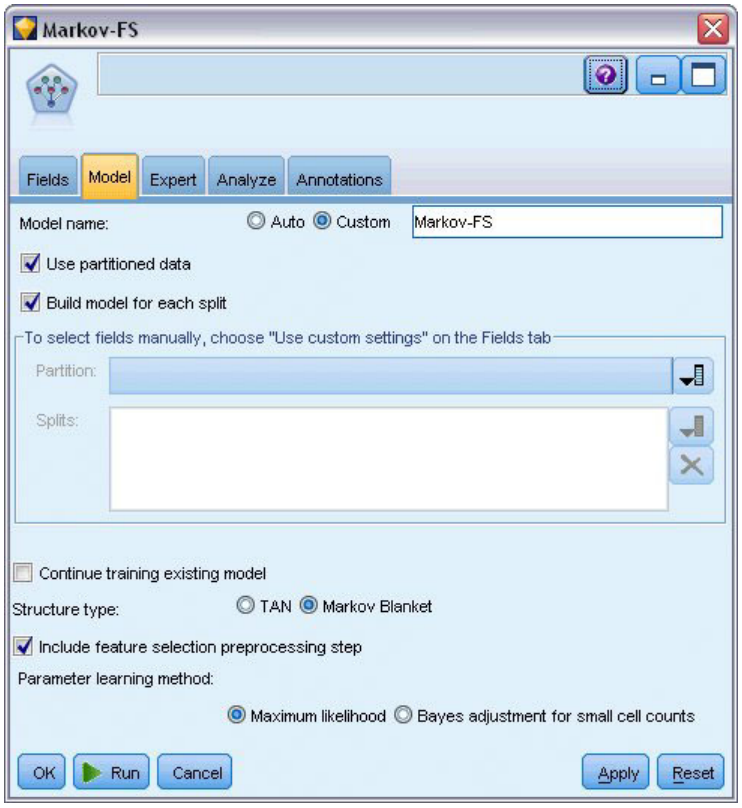


図 238. フィールド選択事前処理を使用した Markov Blanket モデルの作成

モデルの参照

1. ストリームを実行して、モデル・ナゲットを作成します。モデル・ナゲットは、ストリームおよび右上隅のモデル・パレットに追加されます。詳細を表示するには、ストリームの任意のモデル・ナゲットをダブルクリックします。

   モデル・ナゲットの「モデル」タブは、次の 2 つのペインに分けられています。左側のペインには、予測値間の関係に加え、対象と最も重要な予測値間の関係を表示するノードのネットワーク・グラフが含まれています。

   右側のペインには、モデル推定時に各予測値の相対重要度を示す*予測値の重要度*、また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確率値を含む*条件確率* が表示されます。

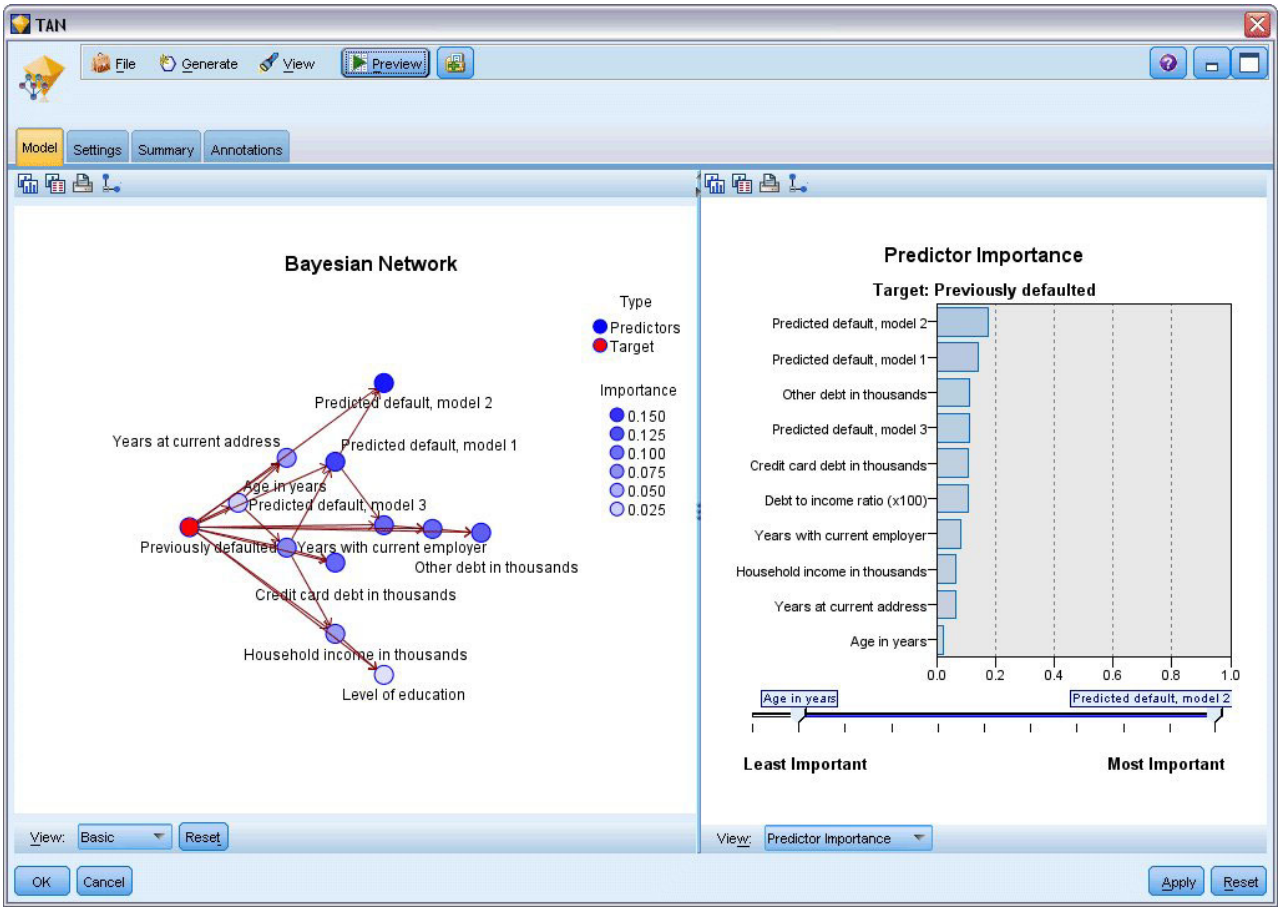


図 239. Tree Augmented Naive Bayes モデルの表示

2. TAN モデル・ナゲットを Markov ナゲットに接続します (警告ダイアログで、「**置換**」を選択します)。
3. Markov モデル・ナゲットを Markov-FS ナゲットに接続します (警告ダイアログで、「**置換**」を選択します)。
4. 3 つのナゲットを見やすいように条件抽出ノードに合わせて配置します。

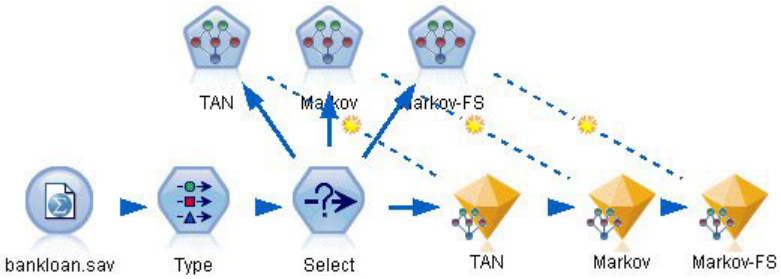


図 240. ストリーム内のナゲットの配置

5. 作成する評価グラフを明確にするためにモデル出力の名前を変更するには、フィルター・ノードを Markov-FS モデル・ナゲットに接続します。

6. 右側の「フィールド」列で、$B-default を TAN に、$B1-default を Markov に、$B2-default を Markov-FS に変更します。

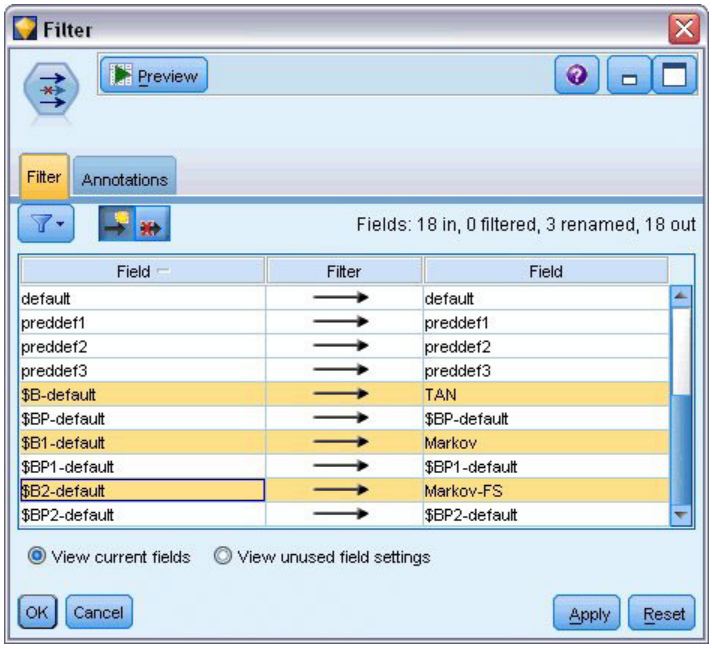


図 241. モデル・フィールド名の変更

モデルの予測精度を比較するために、ゲイン・グラフを作成することができます。

7. 評価グラフ・ノードをフィルター・ノードに接続し、デフォルトの設定を使用してグラフ・ノードを実行します。

グラフは、各モデル・タイプが類似した結果を作成していることを示していますが、Markov モデルが少し良い結果になっています。

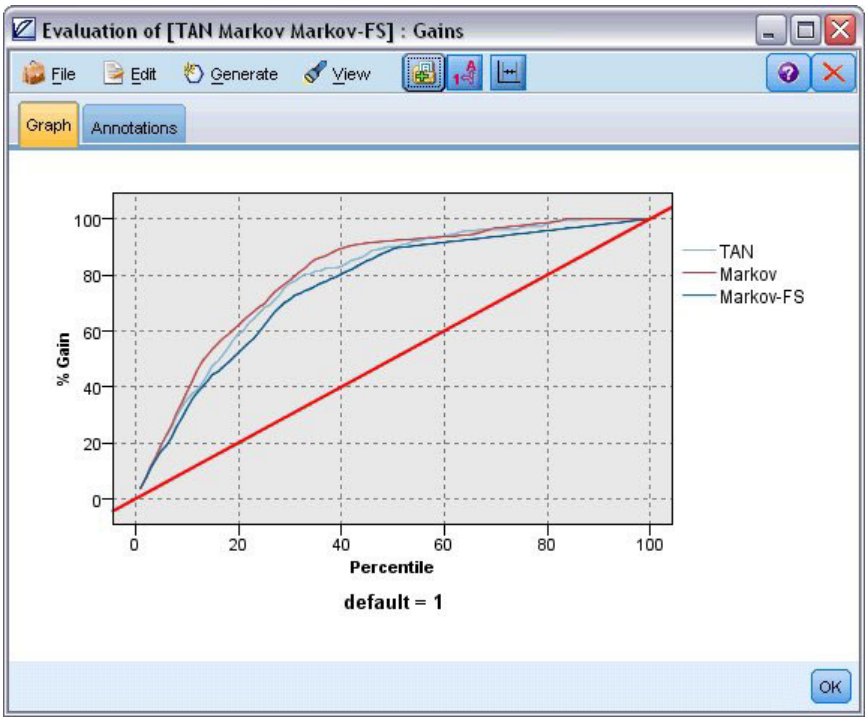


図 242. モデルの精度の評価

各モデルがどれほど良い予測を行っているかを確認するために、評価グラフではなく精度分析ノードを使用できます。これにより、正しい予測と正しくない予測の両方の割合によって精度を示します。

8. 精度分析ノードをフィルター・ノードに接続し、デフォルトの設定を使用して精度分析ノードを実行します。

評価グラフと同じように、これは、Markov モデルが正しい予測について若干優れていることを示しています。ただし、Markov-FS モデルは、Markov モデルより数ポイント下回っているだけです。これは、結果を計算するのに少ない入力を使用するため、データ収集や入力時間、処理時間が少なくなるという点で Markov-FS モデルを使用する方が良いということを示していると考えることもできます。

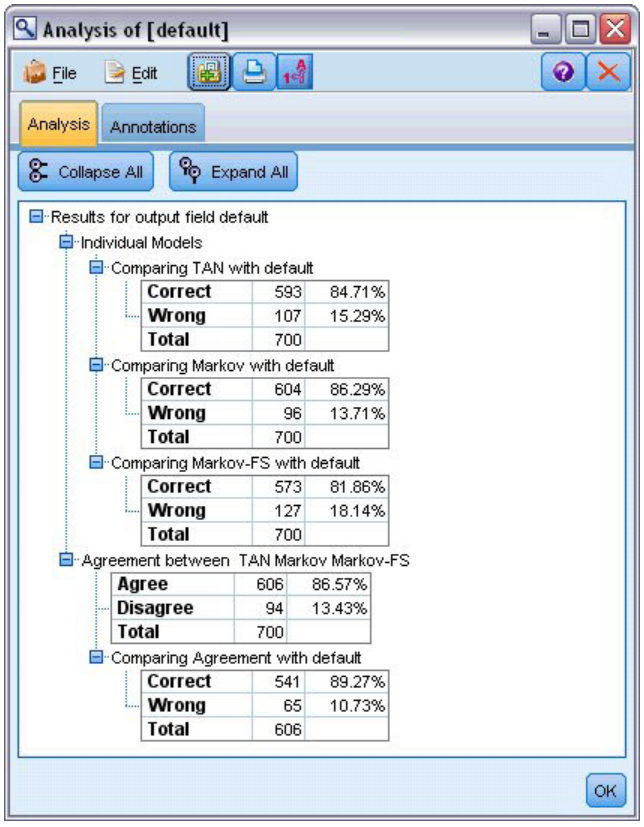


図 243. モデルの精度分析

IBM SPSS Modeler で使用されるモデル作成方法の数学的な基礎の説明については、インストール・ディスクの *¥Documentation* ディレクトリーにある「*IBM SPSS Modeler アルゴリズム・ガイド*」を参照してください。

これらの結果は学習データのみに基づくことに注意してください。モデルが実際の世界の他のデータにどれだけうまく一般化されるかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。

第 18 章 毎月ベースのモデルのリトレーニング (ベイズ・ネットワーク)

ベイズ・ネットワークを使用すると、観測された情報および記録された情報を「常識」という実際の知識と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。

この例では、*telco_Jan.sav* および *telco_Feb.sav* というデータ・ファイルを参照する *bayes_churn_retrain.str* というストリームを使用します。これらのファイルは IBM SPSS Modeler インストール環境の *Demos* ディレクトリーから使用でき、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することができます。*bayes_churn_retrain.str* ファイルは、*streams* ディレクトリー内にあります。

例えば、競合他社に奪われる顧客の数 (解約) に関して、電気通信プロバイダーが懸念しているとします。過去の顧客データを使用して将来解約する可能性が高い顧客を予測することができる場合、これらの顧客を誘導、または他社のサービス・プロバイダーへの移行を思いとどまらせるような提案の対象とすることができます。

この例では、将来解約する可能性が高い顧客を予測するために既存の月の解約データを使用すること、およびそれに続く月のデータを追加してモデルを再調整およびリトレーニングすることに焦点を当てています。

ストリームの構築

1. *Demos* フォルダーの *telco_Jan.sav* を指し示す Statistics ファイル入力ノードを追加します。

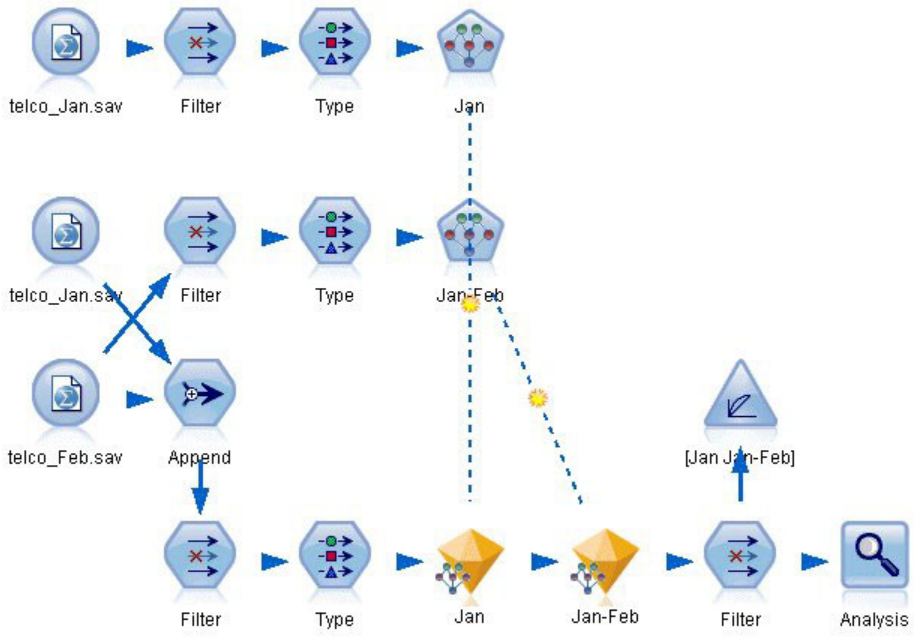


図 244. ベイズ・ネットワークのサンプル・ストリーム

以前の分析では、解約の予測時にいくつかのデータ・フィールドはほとんど重要ではないことが示されていました。これらのフィールドをデータ・セットからフィルターで除外して、モデルの構築およびスコアリング時の処理速度を向上させることができます。

2. フィルター・ノードを入力ノードに追加します。
3. *address*、*age*、*churn*、*custcat*、*ed*、*employ*、*gender*、*marital*、*reside*、*retire*、および *tenure* を除くすべてのフィールドを除外します。
4. 「**OK**」をクリックします。

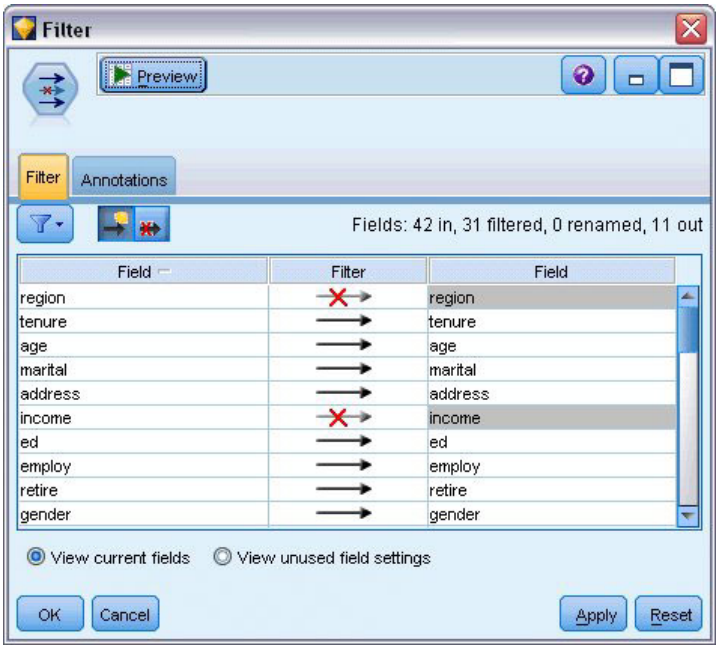


図 245. 不要なフィールドのフィルタリング

5. データ型ノードをフィルター・ノードに追加します。
6. データ型ノードを開き、「**値の読み込み**」ボタンをクリックして「*値*」列にデータを取り込みます。
7. 評価ノードでどの値が真でどの値が偽かを評価できるようにするために、*churn* フィールドの測定の尺度を「**フラグ**」に設定し、その役割を「**対象**」に設定します。「**OK**」をクリックします。

図 246. 対象フィールドの選択

さまざまな種類のベイズ・ネットワークを構築することができます。ただし、この例では、Tree Augmented Naive Bayes (TAN) モデルを構築します。これは、大規模ネットワークを作成し、データ変数間で可能なすべてのリンクを含むため、堅固な初期モデルが構築されます。

8. ベイズ・ネットワーク・ノードをデータ型ノードに接続します。
9. 「モデル」タブのモデル名で、「**カスタム**」を選択し、テキスト・ボックスに `Jan` と入力します。
10. パラメーター学習方法で、「**小さいセルの度数のベイズ調整**」を選択します。
11. 「**実行**」をクリックします。モデル・ナゲットがストリーム、および右上隅の「モデル」パレットに追加されます。

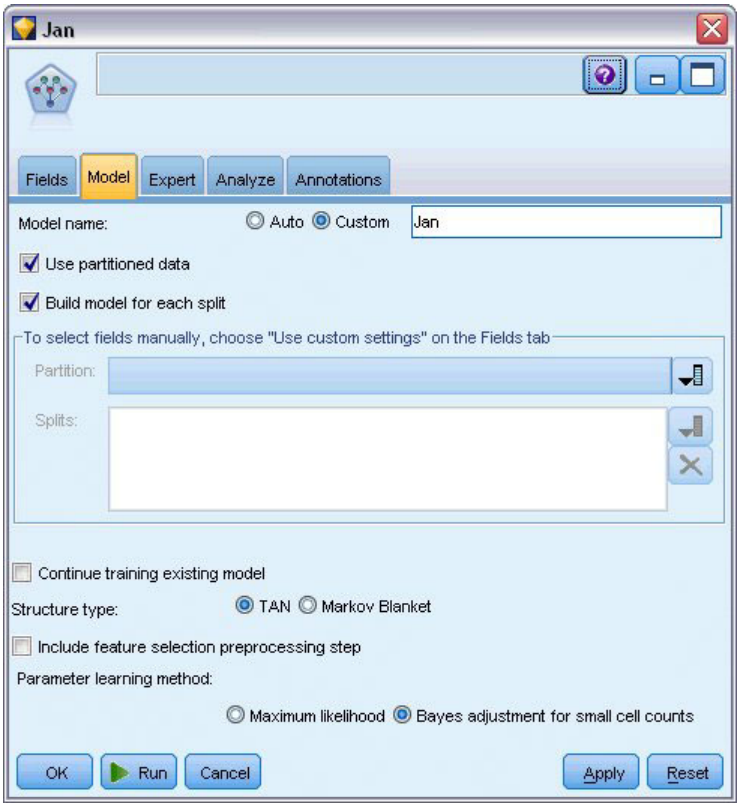


図 247. Tree Augmented Naive Bayes モデルの作成

12. *Demos* フォルダーの *telco_Feb.sav* を指し示す Statistics ファイル入力ノードを追加します。
13. この新しい入力ノードをフィルター・ノードに接続します (警告ダイアログで、「**置換**」を選択して以前の入力ノードへの接続を置き換えます)。

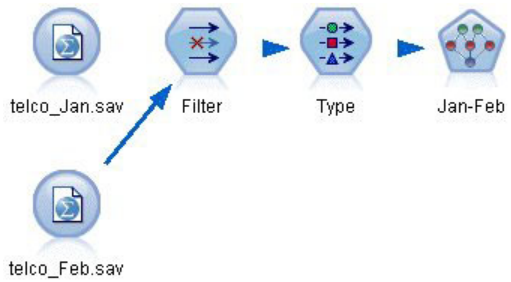


図 248. 2 番目の月のデータの追加

14. ベイズ・ネットワーク・ノードの「モデル」タブのモデル名で、「**カスタム**」を選択し、テキスト・ボックスに `Jan-Feb` と入力します。
15. 「**既存モデルの学習を継続**」を選択します。
16. 「**実行**」をクリックします。モデル・ナゲットはストリーム内の既存のモデル・ナゲットを上書きしますが、右上隅の「モデル」パレットにも追加されます。

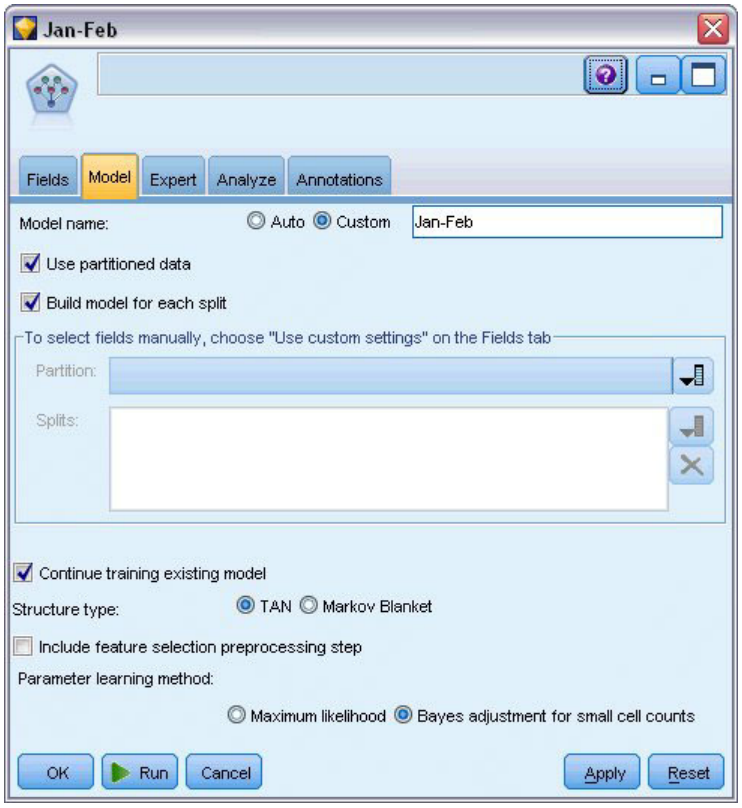


図 249. モデルのリトレーニング

モデルの評価

モデルを比較するには、2 つのデータ・セットを結合させる必要があります。

1. 結合ノードを追加して、*telco_Jan.sav* と *telco_Feb.sav* の両方の入力ノードをそのノードに接続します。

図 250. 2 つのデータ・ソースの結合

2. ストリームの初期のフィルター・ノードおよびデータ型ノードをコピーして、ストリーム領域に貼り付けます。
3. 結合ノードを新しくコピーされたフィルター・ノードに接続します。

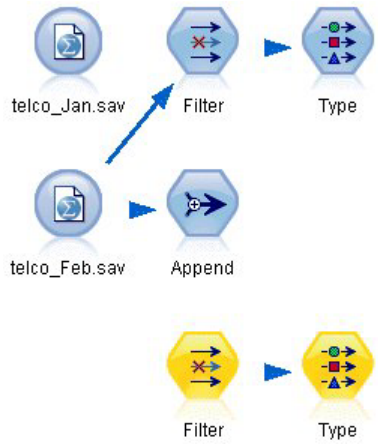


図 251. コピーされたノードのストリームへの貼り付け

2 つのベイズ・ネットワーク・モデルのナゲットは、右上隅の「モデル」パレットにあります。

4. Jan モデル・ナゲットをダブルクリックしてストリームに追加し、新しくコピーされたデータ型ノードにそれを接続します。
5. ストリーム内に既にある Jan-Feb モデル・ナゲットを Jan モデル・ナゲットに接続します。
6. Jan モデル・ナゲットを開きます。

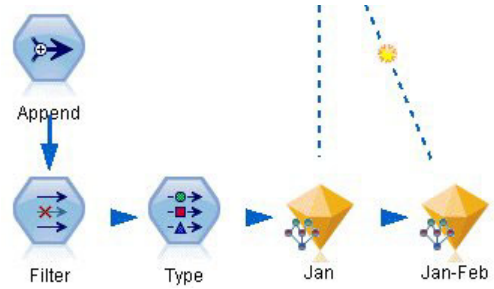


図 252. ナゲットのストリームへの追加

ベイズ・ネットワーク・モデル・ナゲットの「モデル」タブは、2 つの列に分けられています。左側の列には、予測値間の関係に加え、対象と最も重要な予測値間の関係を表示するノードのネットワーク・グラフが含まれています。

右側の列には、モデル推定時に各予測値の相対重要度を示す*予測値の重要度*、また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確率値を含む*条件確率* が表示されます。

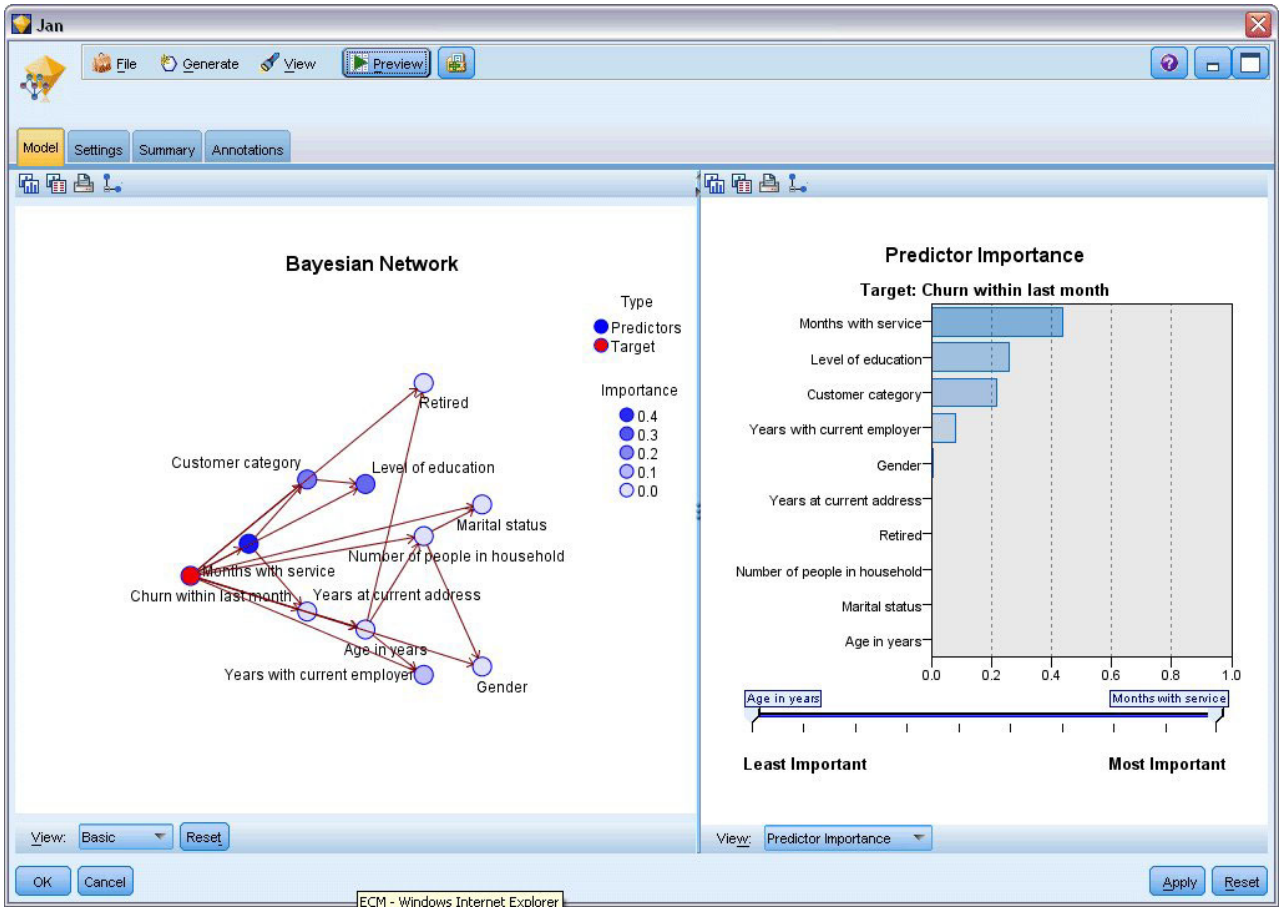


図 253. 予測値の重要度を示すベイズ・ネットワーク・モデル

ノードの条件確率を表示するには、左側の列のノードをクリックします。右側の列が更新され、必要な詳細が表示されます。

条件確率は、データ値がノードの親および兄弟ノードに関連して分割された各ビンごとに表示されます。

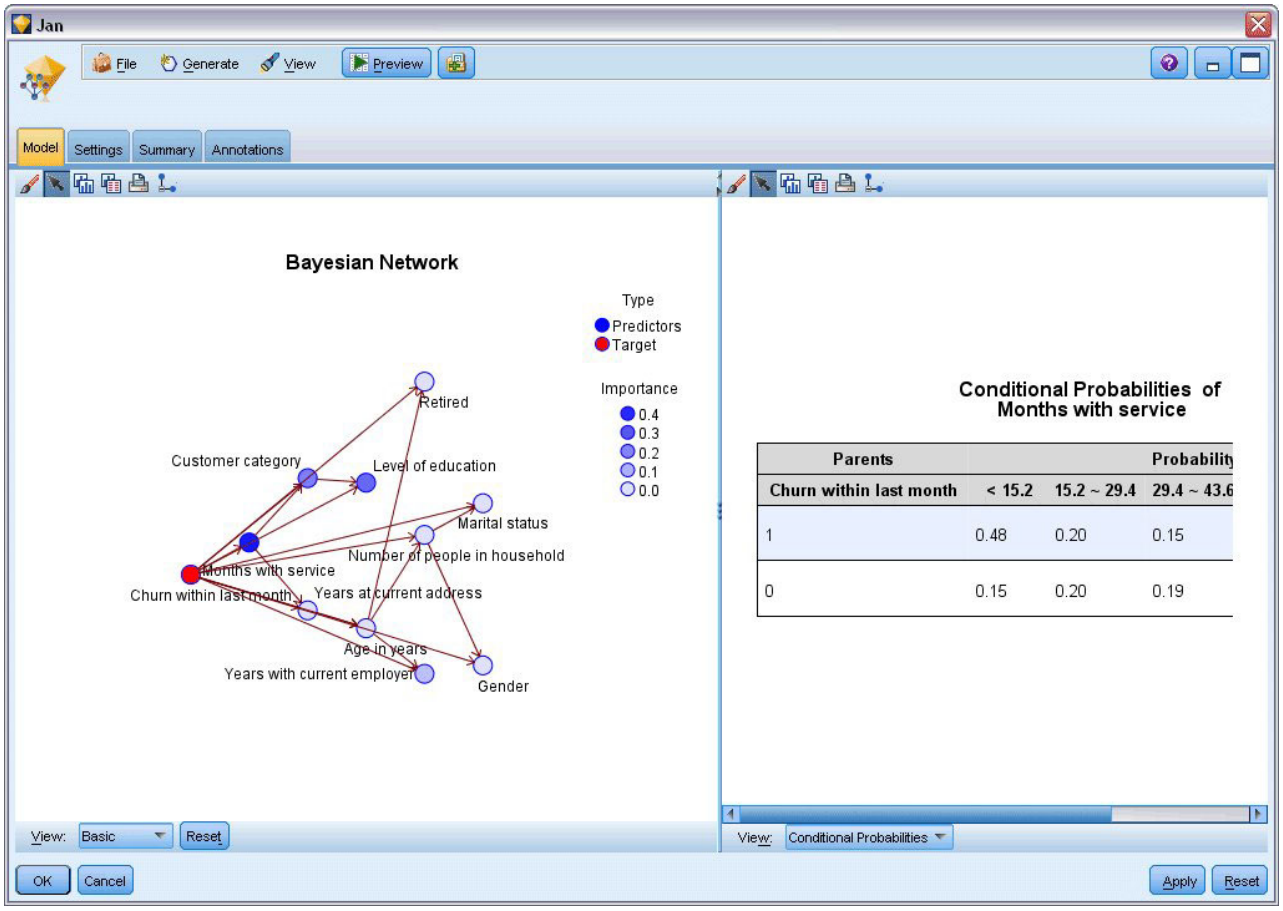


図 254. 条件確率を表示するベイズ・ネットワーク・モデル

7. 明確にするためにモデル出力の名前を変更するには、フィルター・ノードを Jan-Feb モデル・ナゲットに接続します。
8. 右側の「フィールド」列で、$B-churn を Jan に、$B1-churn を Jan-Feb に名前変更します。

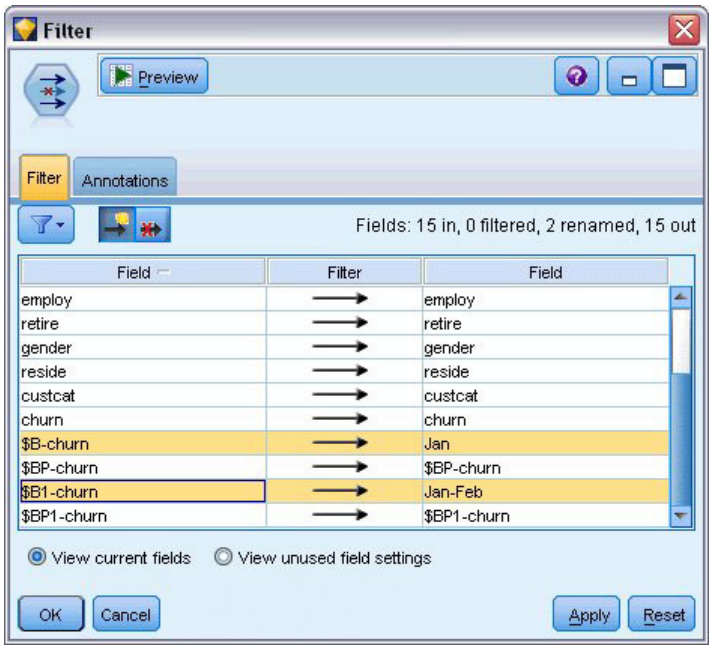


図 255. モデル・フィールド名の変更

各モデルが解約をいかに正確に予測しているかチェックするには、精度分析ノードを使用します。これにより、正しい予測および正しくない予測の両方の割合によって精度が示されます。

9. 精度分析ノードをフィルター・ノードに接続します。
10. 精度分析ノードを開いて、「**実行**」をクリックします。

これにより、両モデルが解約の予測において同じような精度であることが示されます。

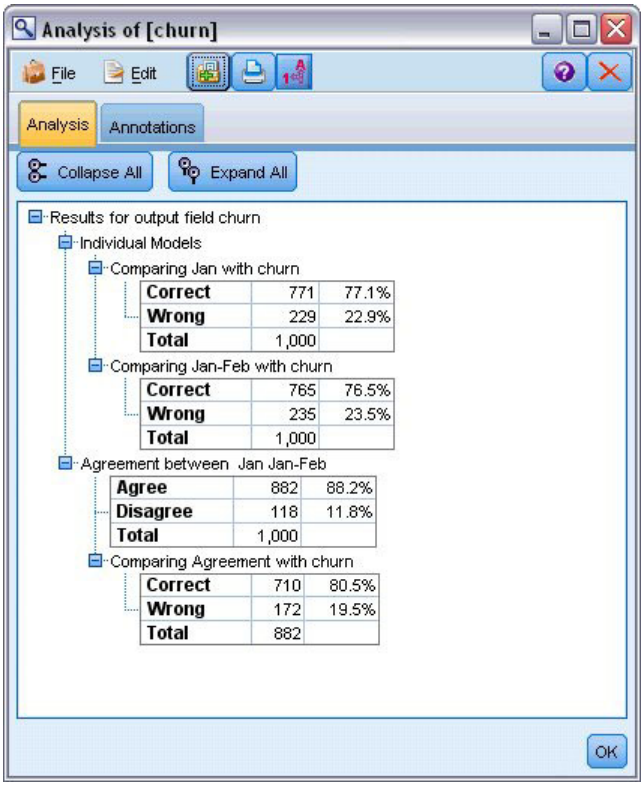


図 256. モデルの精度分析

精度分析ノードの代わりに評価グラフを使用し、ゲイン・グラフを作成すると、モデルの予測精度を比較することができます。

11. 評価グラフ・ノードをフィルター・ノードに接続します。

次に、デフォルト設定を使用してグラフ・ノードを実行します。

グラフは精度分析ノードと同様に、各モデル・タイプが同じような結果を作成していることを示しますが、両方の月のデータを使用するリトレーニング・モデルの方が予測の信頼度が高いため、少し良い結果になっています。

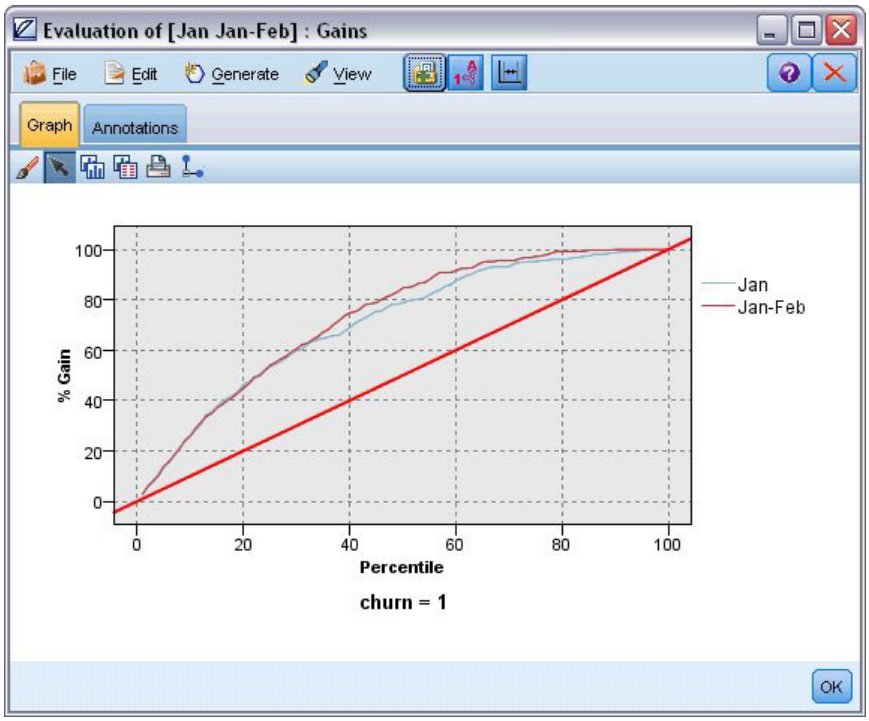


図 257. モデルの精度の評価

IBM SPSS Modeler で使用されるモデル作成方法の数学的な基礎の説明については、インストール・ディスクの *¥Documentation* ディレクトリーにある「*IBM SPSS Modeler アルゴリズム・ガイド*」を参照してください。

これらの結果は学習データのみに基づくことに注意してください。モデルが実際の世界の他のデータにどれだけうまく一般化されるかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。

第 19 章 小売業の販売促進活動 (ニューラル・ネットワーク/C&RT)

この例では、小売業の製品ラインについて、また販売促進活動が売上額にどのような効果を及ぼすかについて表すデータを扱います。 (これは架空のデータです)。この例では、将来の販売促進活動の効果を予測することを目標にしています。稼働状況監視の例と同じように、データ・マイニング・プロセスは、探索、データの準備、学習、およびテストの各フェーズで構成されます。

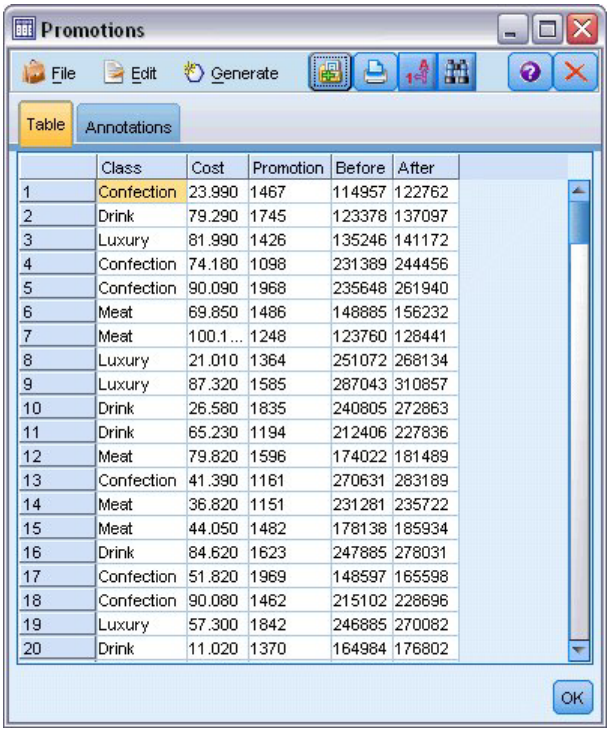
この例では、*goodsplot.str* および *goodslearn.str* というストリームを使用します。これらは、*GOODS1n* および *GOODS2n* というデータ・ファイルを参照します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。ストリーム *goodsplot.str* は、*streams* フォルダー内にあります。また、*goodslearn.str* ファイルは、*streams* ディレクトリー内にあります。

データの検証

各レコードには、次の内容が含まれています。

- *Class*。製品タイプ。
- *Cost*。単価。
- *Promotion*。特定の販売促進活動に費やした金額の指標。
- *Before*。販売促進活動前の収益。
- *After*。販売促進活動後の収益。

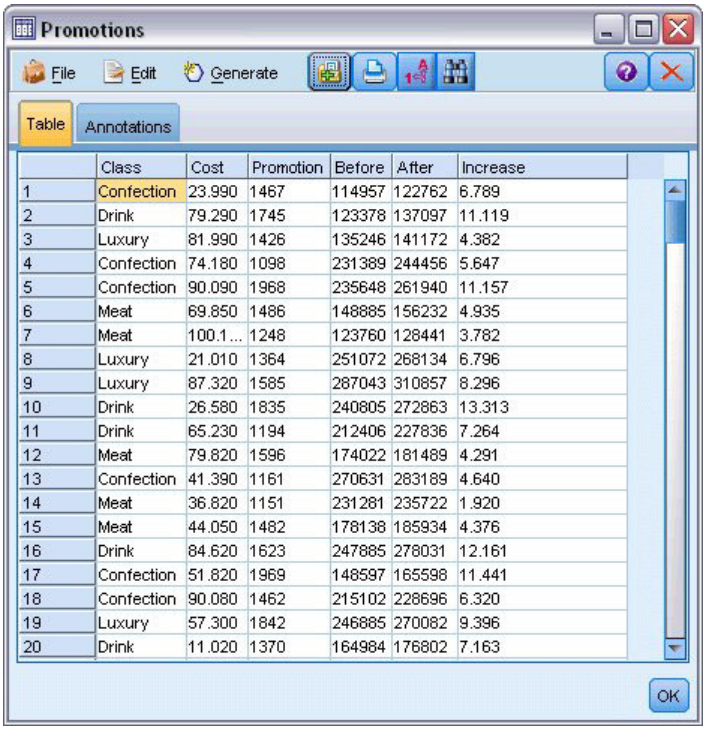
ストリーム *goodsplot.str* には、テーブルにデータを表示するための単純なストリームが含まれています。2 つの収益フィールド (*Before* および *After*) は、絶対項で表されていますが、販売促進活動後の (販売促進活動の効果と考えられる) 収益の増加の方が役に立つ数値と思われます。



	Class	Cost	Promotion	Before	After
1	Confection	23.990	1467	114957	122762
2	Drink	79.290	1745	123378	137097
3	Luxury	81.990	1426	135246	141172
4	Confection	74.180	1098	231389	244456
5	Confection	90.090	1968	235648	261940
6	Meat	69.850	1486	148885	156232
7	Meat	100.1...	1248	123760	128441
8	Luxury	21.010	1364	251072	268134
9	Luxury	87.320	1585	287043	310857
10	Drink	26.580	1835	240805	272863
11	Drink	65.230	1194	212406	227836
12	Meat	79.820	1596	174022	181489
13	Confection	41.390	1161	270631	283189
14	Meat	36.820	1151	231281	235722
15	Meat	44.050	1482	178138	185934
16	Drink	84.620	1623	247885	278031
17	Confection	51.820	1969	148597	165598
18	Confection	90.080	1462	215102	228696
19	Luxury	57.300	1842	246885	270082
20	Drink	11.020	1370	164984	176802

図 258. 製品売上額への販売促進活動の効果

goodsplot.str には、この値 (販売促進活動前の収益に対するパーセンテージ) を派生させるノードも *Increase* というフィールドに含まれており、このフィールドを示すテーブルが表示されます。



	Class	Cost	Promotion	Before	After	Increase
1	Confection	23.990	1467	114957	122762	6.789
2	Drink	79.290	1745	123378	137097	11.119
3	Luxury	81.990	1426	135246	141172	4.382
4	Confection	74.180	1098	231389	244456	5.647
5	Confection	90.090	1968	235648	261940	11.157
6	Meat	69.850	1486	148885	156232	4.935
7	Meat	100.1...	1248	123760	128441	3.782
8	Luxury	21.010	1364	251072	268134	6.796
9	Luxury	87.320	1585	287043	310857	8.296
10	Drink	26.580	1835	240805	272863	13.313
11	Drink	65.230	1194	212406	227836	7.264
12	Meat	79.820	1596	174022	181489	4.291
13	Confection	41.390	1161	270631	283189	4.640
14	Meat	36.820	1151	231281	235722	1.920
15	Meat	44.050	1482	178138	185934	4.376
16	Drink	84.620	1623	247885	278031	12.161
17	Confection	51.820	1969	148597	165598	11.441
18	Confection	90.080	1462	215102	228696	6.320
19	Luxury	57.300	1842	246885	270082	9.396
20	Drink	11.020	1370	164984	176802	7.163

図 259. 販売促進活動後の収益増加

さらにこのストリームは、増加のヒストグラムと、展開された販売促進活動にかかったコストと増加を比較した散布図を表示し、そこに関連する製品カテゴリーをオーバーレイします。

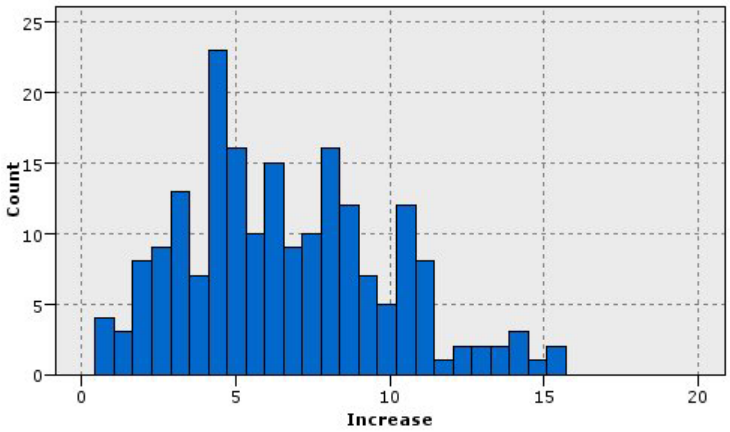


図 260. 収益増加のヒストグラム

散布図を見ると、製品の各クラスについて、収益増加と販売促進活動コスト増加の間には、ほとんど線型に近い関係が存在します。したがって、ディシジョン・ツリーまたはニューラル・ネットワークを使用することで、他の使用可能なフィールドから、十分な精度で収益増加を予測できると考えられます。

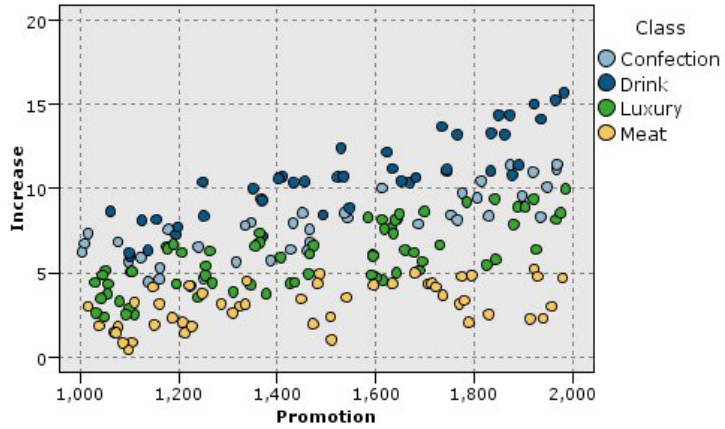


図 261. 収益増加と販売促進活動コストの比較

学習とテスト

ストリーム *goodslearn.str* は、ニューラル・ネットワークとディシジョン・ツリーを学習し、この収益増加の予測を行います。

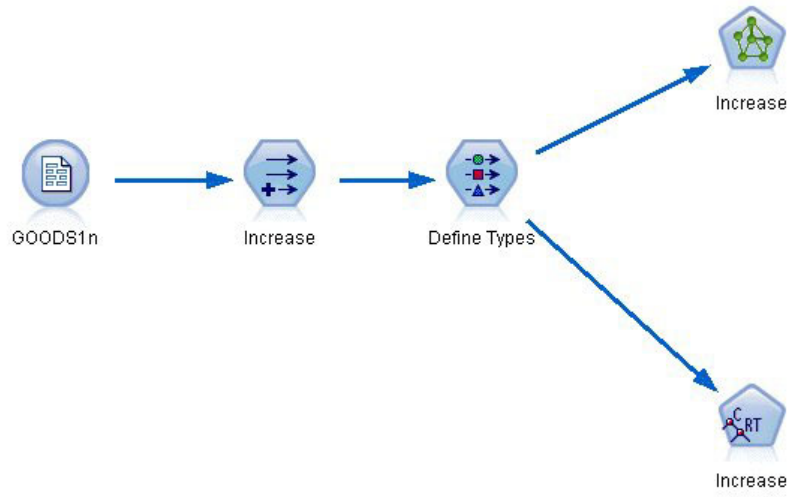


図 262. ストリーム goodslearn.str のモデル作成

モデル作成ノードを実行し、実際のモデルを生成したら、学習プロセスの結果をテストすることができます。テストを行うには、データ型ノードと新しい精度分析ノードの間に連続してディシジョン・ツリーとネットワークを接続し、入力 (データ) ファイルを *GOODS2n* に変更してから、精度分析ノードを実行します。このノードの出力から (特に予測増加量と正しい回答の間の線型相関から)、学習したシステムが収益増加をかなり正確に予測したことが分かります。

さらに詳細な検証では、学習したシステムが比較的大きな誤差を作成したケースに焦点を当てることができます。これは、収益増加の予測値と実際の増加を対比してプロットすれば特定できます。このグラフでの外れ値は、IBM SPSS Modeler の対話式グラフィックスを使用して選択できます。またそれらのプロパティーを使用して、データ詳細や学習プロセスを微調整して精度を向上させることができる場合もあります。

第 20 章 稼働状況の監視 (ニューラル・ネット/C5.0)

この例では、機械のステータス情報の監視と、障害の状態の認識やその予測に関する問題を取り上げます。データは架空のシミュレーションから作成され、時間の経過とともに測定されたさまざまな連結系列で構成されています。各レコードは、次の観点からの情報を基にした、機械のスナップショット・レポートになります。

- *時間*。整数。
- *電力*。整数。
- *温度*。整数。
- *圧力*。正常な場合は 0、瞬間の圧力についての警告がある場合は 1。
- *実行時間*。最後に修理、調整してからの経過時間。
- *ステータス*。正常な場合は 0、エラーが発生した場合、エラー・コード (101、202、303) に変化。
- *結果*。この時系列に表示されるエラー・コード。エラーが発生しなかった場合は 0。(これらのコードは、後々使用可能になります)。

この例では、*condplot.str* および *condlearn.str* というストリームを使用します。これらは、*COND1n* および *COND2n* というデータ・ファイルを参照します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*condplot.str* および *condlearn.str* ファイルは、*streams* ディレクトリー内にあります。

以下の表に示すように、各時系列データにおいて、正常に稼働している期間のレコードがあり、そのあとに、障害が起こるまでの期間のレコードがあります。

時間	電力	温度	圧力	実行時間	ステータス	結果
0	1059	259	0	404	0	0
1	1059	259	0	404	0	0
			...			
51	1059	259	0	404	0	0
52	1059	259	0	404	0	0
53	1007	259	0	404	0	303
54	998	259	0	404	0	303
			...			
89	839	259	0	404	0	303
90	834	259	0	404	303	303
0	965	251	0	209	0	0
1	965	251	0	209	0	0
			...			
51	965	251	0	209	0	0
52	965	251	0	209	0	0
53	938	251	0	209	0	101
54	936	251	0	209	0	101

時間	電力	温度	圧力	実行時間	ステータス	結果
			...			
208	644	251	0	209	0	101
209	640	251	0	209	101	101

次の過程は、ほとんどのデータ・マイニングのプロジェクトにおいて一般的に行われます。

- データを検証して、対象とする状態を予測したり認識したりするのに有効であると考えられる属性を判別します。
- これらの属性を保持するか (既に含まれている場合)、あるいは必要であればこれらの属性を派生させてデータに追加します。
- 結果として作成されたデータを、ルールとニューラル・ネットの学習に使用します。
- 学習を行ったシステムを、独立したテスト・データを使用してテストします。

データの検証

ファイル *condplot.str* は、プロセスの最初の部分を示しています。このファイルには、さまざまなグラフをプロットするストリームが含まれています。温度や電力量の時系列データに明らかなパターンがある場合、差し迫ったエラー状態を識別したり、あるいはそれらの発生を予測したりすることも可能になります。以下のストリームは温度と電力量の両方について、異なる 3 つのエラー・コードに関連付けられた時系列を個別のグラフをプロットします。つまり、合計 6 つのグラフが作成されることになります。条件抽出ノードによって、それぞれのエラー・コードに関連するデータが選別されます。

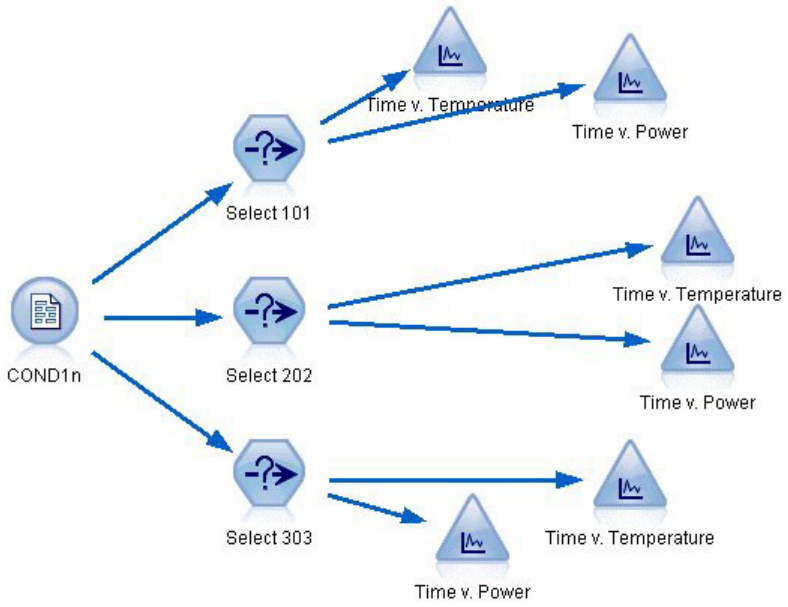


図 263. Condplot ストリーム

このストリームの結果を以下の図に示します。

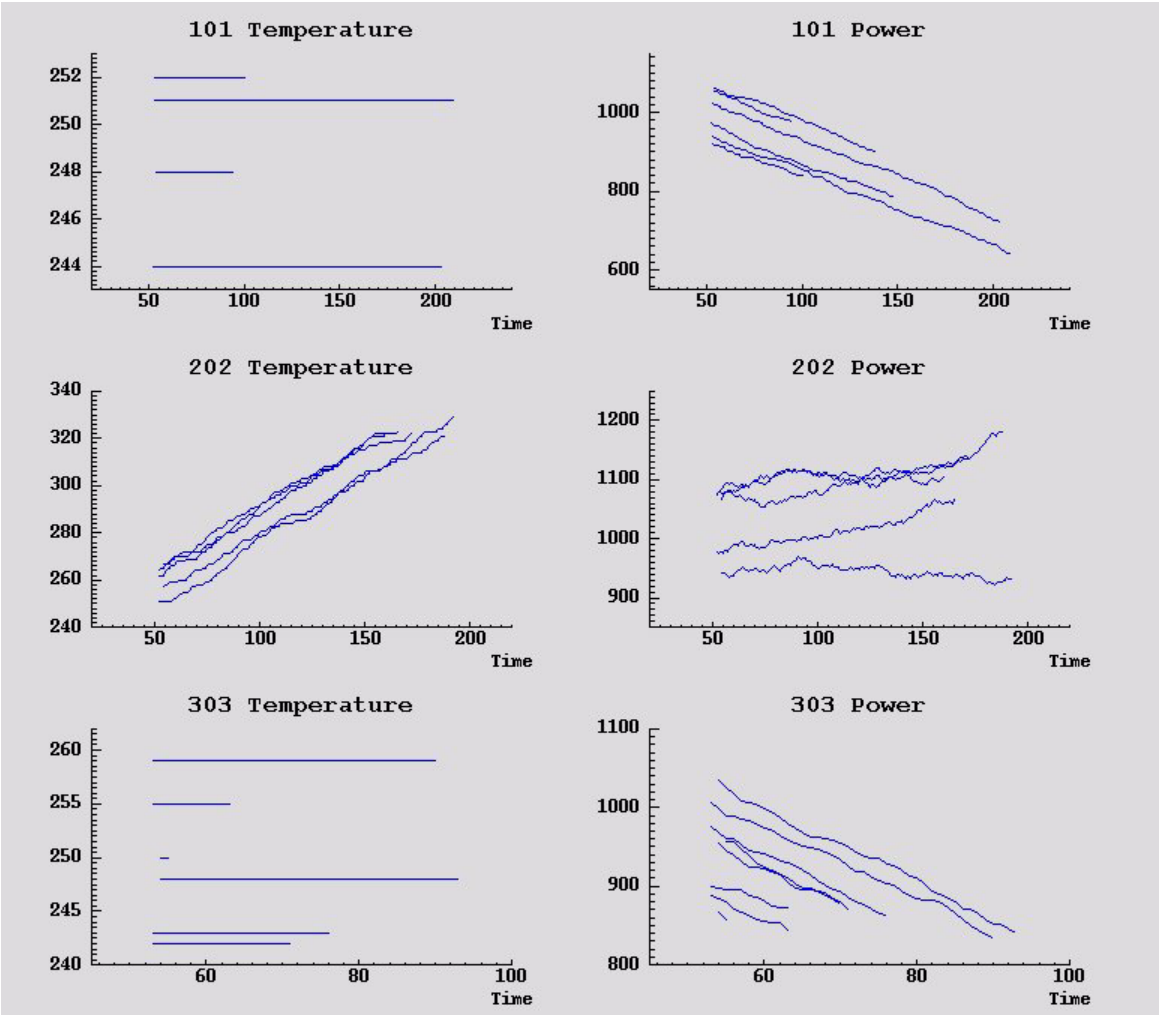


図 264. 時間の経過に伴う温度と電力量

グラフは、202 エラーと 101 や 303 のエラーが明確に区別されたパターンを示しています。202 エラーにおいては、時間の経過とともに温度が上昇し、また電力量が変動しています。しかし、他のエラーにはこのような動きは見られません。ただし、101 エラーと 303 エラーとを区別するパターンは、それほど明確ではありません。どちらのエラーにおいても温度は一定であり、また電力量は減少していますが、303 エラーでは、電力量の減少がより急激になっています。

これらのグラフによって、温度および電力量の変化の有無やその変化の割合、また変動の有無とその程度が、障害を予測したり区別したりすることに関連しているのがわかります。そのため、学習システムを適用する前に、これらの属性をデータに追加する必要があります。

データの準備

データの検証結果に基づいて、ストリーム *condlearn.str* は関連するデータのフィールドを作成し、障害を予測するための学習を行います。

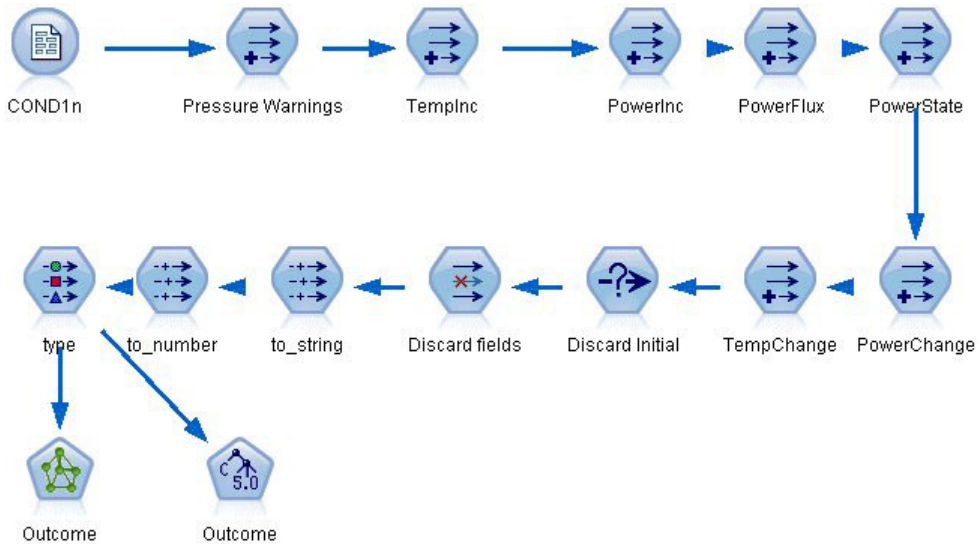


図 265. Condlearn ストリーム

このストリームは、さまざまなフィールド作成ノードを使用してモデル作成用にデータを準備します。

- 「**可変長ファイル・ノード**」。データ・ファイル *COND1n* を読み取ります。
- 「**圧力警告のフィールド作成**」。瞬間の圧力の警告数のカウント。時間が 0 に戻るとリセットされます。
- 「**温度上昇のフィールド作成**」。@DIFF1 を使用して、温度の瞬間の変化率を計算します。
- 「**電力量増加のフィールド作成**」。@DIFF1 を使用して、電力量の瞬間の変化率を計算します。
- 「**電力量変動のフィールド作成**」。これはフラグであり、1 つ前のレコードから該当するレコードになった際に電力量がその変化の方向を変えたときに真 (true) となります。つまり電力量の最高点あるいは最小点を示します。
- 「**電力量状況のフィールド作成**」。*Stable* (安定状態) から始まり、電力量の変動が 2 回連続して検出されると、*Fluctuating* (変動状態) に切り替わります。5 つの時間区分にわたって電力量の変動がないか、または*時間* がリセットされた場合にのみ、*Stable* (安定状態) に戻ります。
- 「**電力量変動**」。最後の 5 つの時間区分における*電力量上昇* の平均です。
- 「**温度変動**」。最後の 5 つの時間区分における*温度上昇* の平均です。
- 「**初期値の破棄 (条件抽出)**」。境界における*電力量* と*温度* の大きな (不正確な) 変動を避けるために、各時系列の最初のレコードを破棄します。
- 「**フィールドの破棄**」。レコードを*実行時間*、*状態*、*結果*、*圧力の警告*、*電力量状態*、*電力量変化*、および*温度変化* に分割します。
- 「**データ型**」。*結果* の役割を「**対象**」 (予測するフィールド) として定義します。さらに、測定の尺度の*結果* は「**名義型**」、*圧力の警告* は「**連続型**」、*電力量状態* は「**フラグ型**」として定義します。

学習

condlearn.str 内のストリームを実行すると、C5.0 ルールとニューラル・ネットワーク (ネット) が学習されます。ネットワークの学習にはある程度時間がかかりますが、早期に学習を中断して、使用に耐え得る結果を生成するネットワークを保存することもできます。学習が完了したら、マネージャー・ウィンドウの右上の「モデル」タブが点滅し、新しい 2 つのナゲットが作成されたことを知らせます。1 つはニューラル・

ネットを表し、もう 1 つはルールを表します。

図 266. モデル・ナゲットが表示されたモデル・マネージャー

既存のストリームにモデル・ナゲットを追加して、システムをテストしたり、モデルの結果をエクスポートしたりすることもできます。この例では、モデルの結果をテストします。

テスト

モデル・ナゲットをストリームに追加します。どちらもデータ型ノードに接続されています。

1. 示されたとおりにナゲットを再配置し、データ型ノードがニューラル・ネットワーク・ナゲットに接続され、それが C5.0 ナゲットに接続されるようにします。
2. 精度分析ノードを C5.0 ナゲットに接続します。
3. (*COND1n* ではなく) ファイル *COND2n* を読み取るように元のソース・ノードを編集します。*COND2n* には画面に表示されないテスト・データが含まれています。

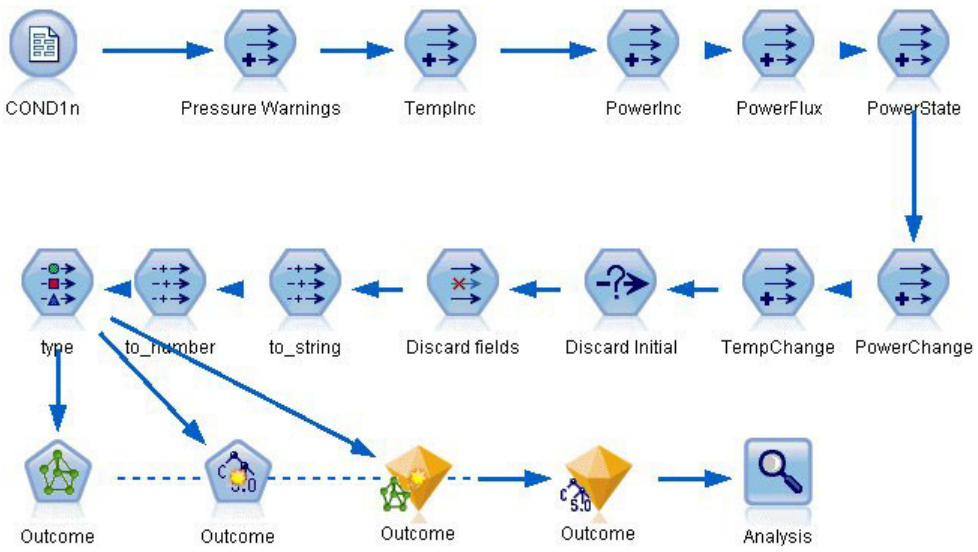


図 267. 学習済みのネットワークのテスト

4. 精度分析ノードを開いて、「実行」をクリックします。

そうすると、学習済みのネットワークとルールの精度を表す数値が生成されます。

第 21 章 電気通信会社の顧客の分類 (判別分析)

判別分析は、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。線型回帰と似ていますが、数値型対象フィールドではなくカテゴリー対象フィールドを取ります。

例えば、電気通信プロバイダーがその顧客ベースを、サービス使用パターンによって区分しており、顧客を 4 つのグループにカテゴリー化しているとします。人口統計データを使用して顧客がどのグループに所属するかを予測できれば、個々の見込み客にあわせてサービスをカスタマイズすることができます。

この例では、*telco.sav* というデータ・ファイルを参照する *telco_custcat_discriminant.str* というストリームを使用します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*telco_custcat_discriminant.str* ファイルは、*streams* ディレクトリー内にあります。

この例は、使用パターンを予測するための人口統計データの使用方法に注目します。以下のように、対象フィールド *custcat* には、4 つの顧客グループに対応する 4 つの可能な値があります。

値	ラベル
1	基本サービス
2	E-サービス
3	プラス・サービス
4	トータル・サービス

ストリームの作成

1. 最初に、変数および値ラベルを出力に表示するためにストリームのプロパティーを設定します。メニューから次の項目を選択します。

   「ファイル」 > 「ストリームのプロパティー...」 > 「オプション」 > 「全般」
2. **「出力中のフィールドと値ラベルを表示する」**が選択されていることを確認し、**「OK」**をクリックします。

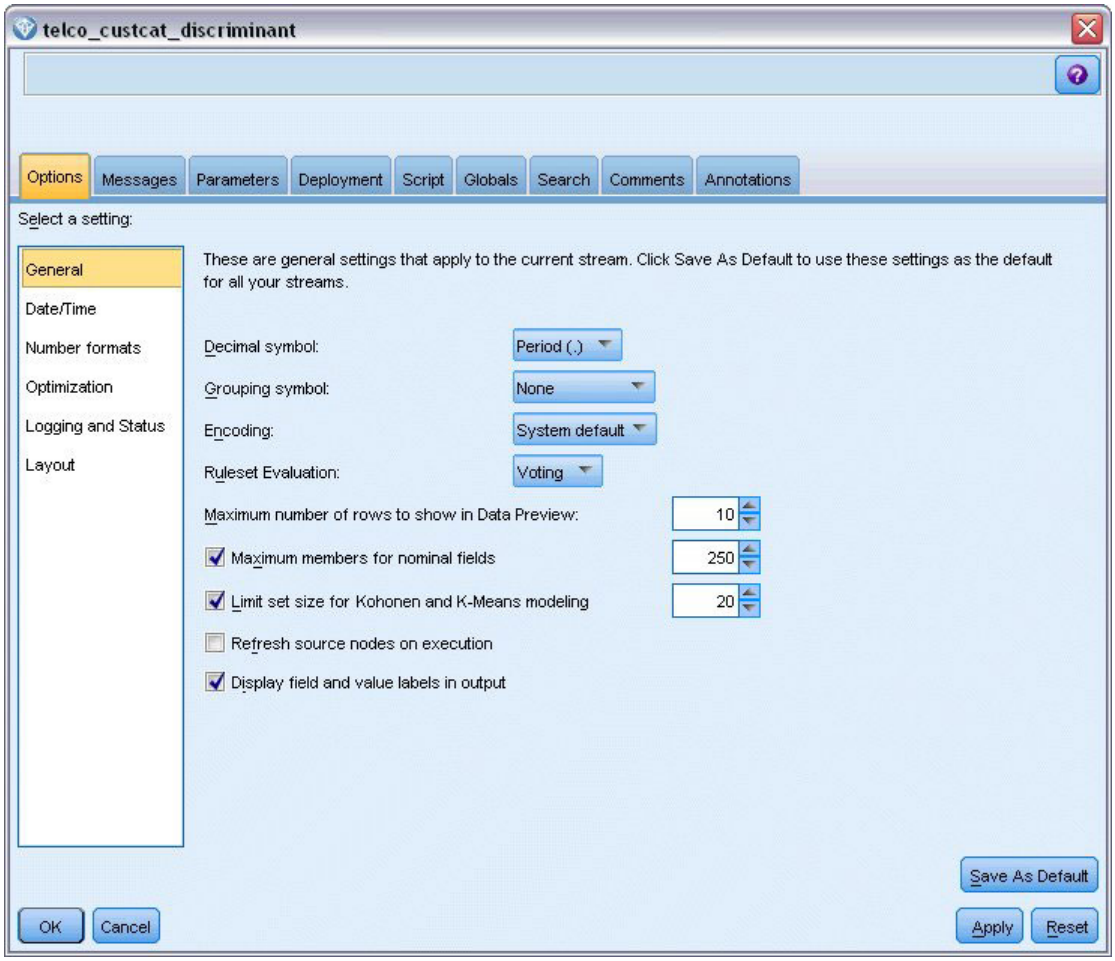


図 268. ストリームのプロパティー

3. *Demos* フォルダーの *telco.sav* を指し示す Statistics ファイル入力ノードを追加します。

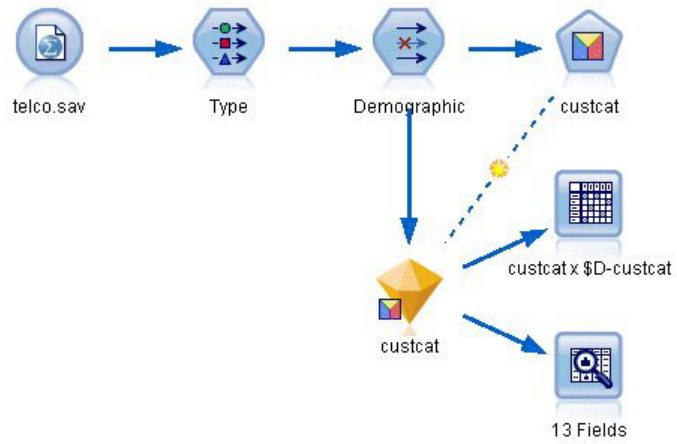


図 269. 判別分析を使用して顧客を分類するためのサンプル・ストリーム

a. データ型ノードを追加し、「**値の読み込み**」をクリックして、すべての測定の尺度が正しく設定されていることを確認します。例えば、値 0 および 1 を持つほとんどのフィールドはフラグと見なすこ

とができます。

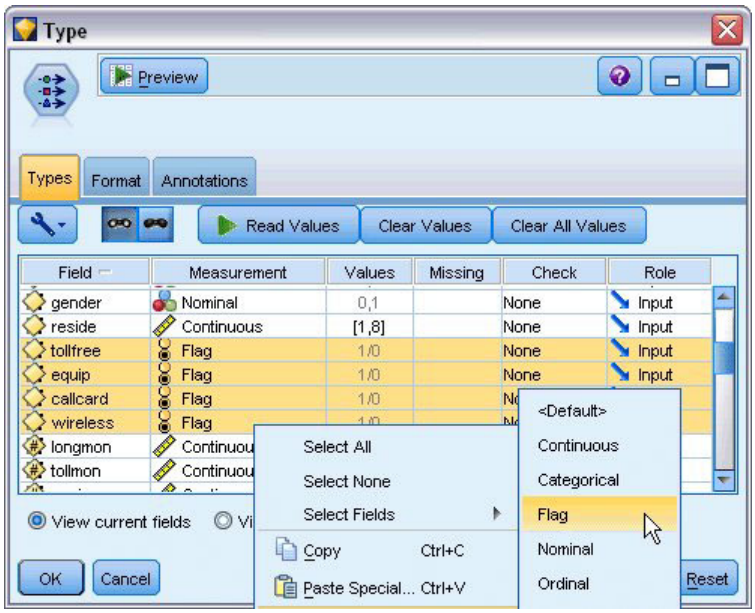


図 270. 複数のフィールドの測定の尺度の設定

ヒント: 類似の値 (0/1 など) を持つ複数のフィールドのプロパティーを変更するには、「*値*」列見出しをクリックしてフィールドを値の順序でソートし、Shift キーを押しながらマウスまたは矢印キーを使用して、変更するフィールドすべてを選択します。選択範囲を右クリックして、測定の尺度を変更するか、選択したフィールドの他の属性を変更します。

「*性別*」は、フラグではなく、2 つの値セットを持つフィールドと見なす方がより適切であるため、「尺度」の値は 「**名義型**」 のままにします。

b. 「*custcat*」フィールドの役割を「**対象**」に設定します。その他のフィールドの役割は、すべて「**入力**」に設定します。

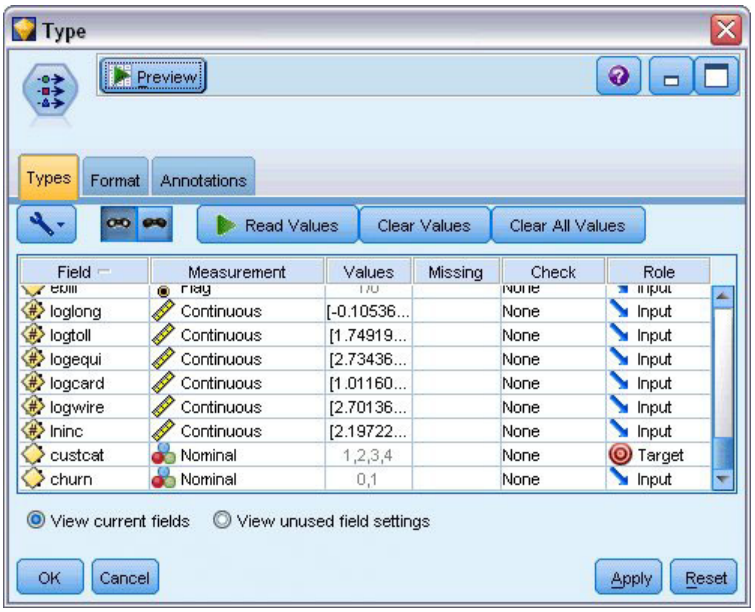


図 271. フィールドの役割の設定

この例はデモグラフィックに注目するため、フィルター・ノードを使用して関連するフィールド (*地域*、*年齢*、*結婚*、*住所*、*収入*、*学歴*、*雇用*、*退職*、*性別*、*居住*、および *custcat*) のみを含めます。この分析の目的上、その他のフィールドは除外してもかまいません。

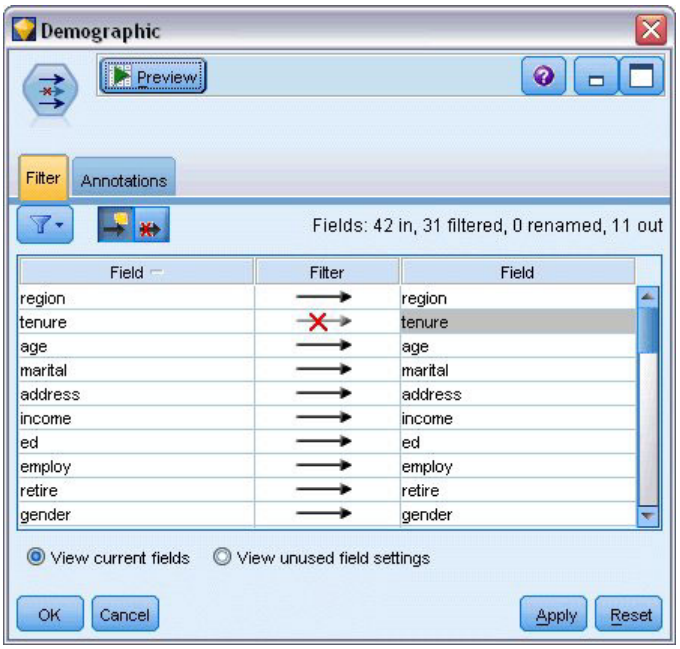


図 272. デモグラフィック・フィールドのフィルタリング

(または、これらのフィールドを除外するのではなく役割を「**なし**」に変更するか、モデル作成ノードで使用するフィールドを選択することもできます。)

4. 判別分析ノードで、「モデル」タブをクリックし、「**ステップワイズ**」法を選択します。

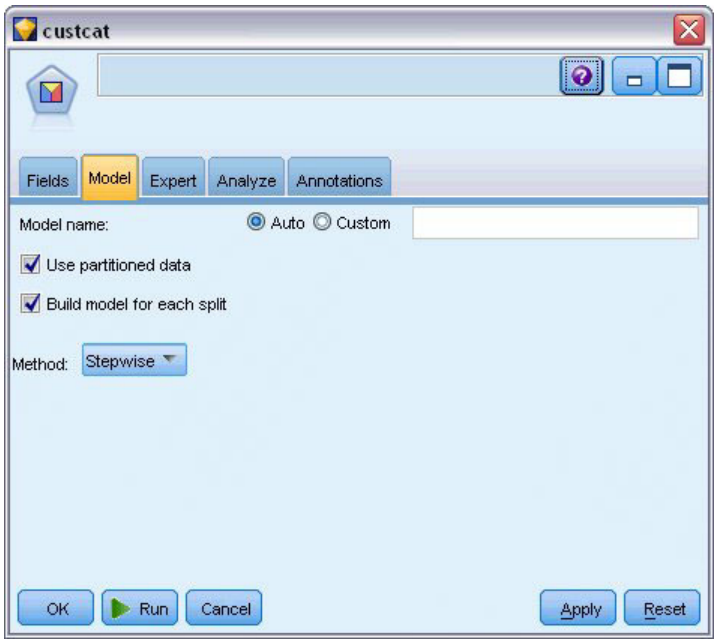


図 273. モデル・オプションの選択

5. 「エキスパート」タブで、モードを「**エキスパート**」に設定し、「**出力**」をクリックします。
6. 「詳細出力」ダイアログ・ボックスで、「**集計表**」、「**地域マップ**」、および「**ステップの要約**」を選択し、「**OK**」をクリックします。

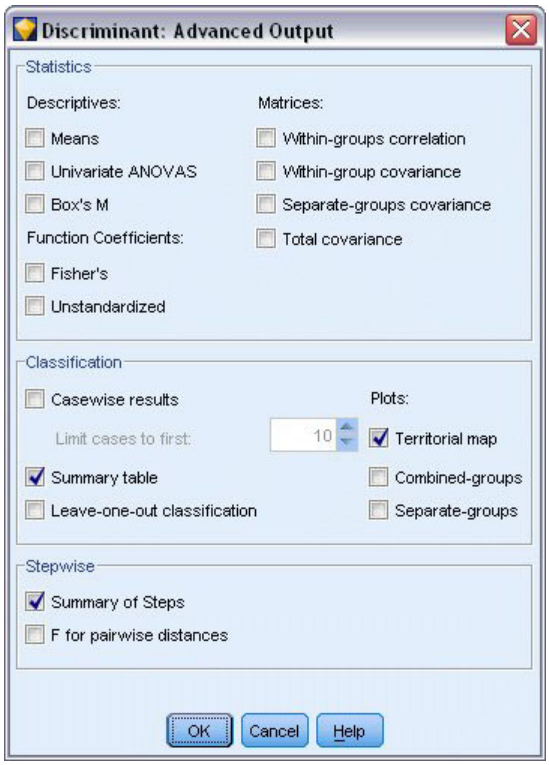


図 274. 出力オプションの選択

モデルの検証

1. 「**実行**」をクリックしてモデルを作成します。これはストリームと、右上の「モデル」パレットに追加されます。詳細を表示するには、ストリームのモデル・ナゲットをダブルクリックします。

   「要約」タブに、検証用に提示された対象およびすべての入力 (予測値フィールド) のリストが (他の項目とともに) 表示されます。

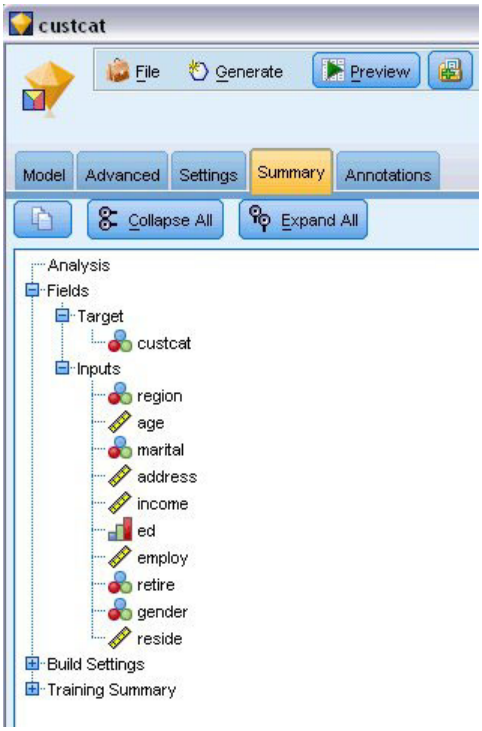


図 275. 対象および入力フィールドが表示されたモデルの要約

この判別分析結果の詳細を表示するには、次のようにします。

2. 「詳細」タブをクリックします。
3. 「外部ブラウザーで起動」ボタン (「モデル」タブのすぐ下) をクリックして、Web ブラウザーで結果を表示します。

通信業界の顧客を分類するために使用する判別分析の出力の分析

ステップワイズ法の判別分析

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	Age in years	1.000	1.000	7.521	.978
	Marital status	1.000	1.000	3.500	.990
	Years at current address	1.000	1.000	8.433	.975
	Household income in thousands	1.000	1.000	6.689	.980
	Level of education	1.000	1.000	61.454	.844
	Retired	1.000	1.000	3.005	.991
	Years with current employer	1.000	1.000	16.976	.951
	Gender	1.000	1.000	.373	.999
	Number of people in household	1.000	1.000	3.976	.988

図 276. 分析に含まれない変数、ステップ 0

多数の予測値がある場合、ステップワイズ法はモデルで使用する「最良」の変数を自動的に選択する際に役立ちます。ステップワイズ法は予測値を一切含まないモデルで開始されます。各ステップで、投入基準 (デフォルトは 3.84) を超える投入する *F* の最大値を有する予測値をモデルに追加します。

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
3	Age in years	.535	.535	.252	.795
	Marital status	.605	.593	1.507	.792
	Years at current address	.776	.771	3.514	.787
	Household income in thousands	.688	.657	.687	.794
	Retired	.917	.880	.353	.795
	Gender	.997	.931	.395	.795

図 277. 分析に含まれない変数、ステップ 3

最終ステップで分析から外されたすべての変数の投入する *F* 値は 3.84 より小さいため、これ以上追加されません。

Step		Tolerance	F to Remove	Wilks' Lambda
1	Level of education	1.000	61.454	
2	Level of education	.953	59.108	.951
	Years with current employer	.953	14.933	.844
3	Level of education	.951	60.046	.940
	Years with current employer	.934	15.824	.834
	Number of people in household	.979	4.841	.807

図 278. 分析に含まれる変数

この表は、各ステップの分析に含まれる変数の統計を表示しています。*許容度* は方程式においてその他の独立変数によって構成されない変数の分散の比率です。非常に低い許容度の変数はモデルに対する情報をほとんど提供せず、計算上の問題を引き起こすことがあります。

除去する *F* 値は、変数が現行のモデルから除去された場合に発生することを説明するのに役立ちます (その他の変数が残っているとする)。 変数に投入する除去する *F* は、前のステップの 投入する *F* と同様です (「分析に含まれない変数」表に表示)。

ステップワイズ法に関する注意事項

ステップワイズ法は便利ですが、制約があります。ステップワイズ法は統計の利点のみに基づいてモデルを選択するため、*実際的有意確率*のない予測値を選択する可能性があります。データを扱った経験があり、重要な予測値を予想しているのであれば、その知識を使用すべきであり、ステップワイズ法は避ける必要があります。ただし、多数の予測値があり、どこから始めればいいのかわからない場合は、ステップワイズ分析を実行し、選択されたモデルを修正する方が、何もないよりはよいでしょう。

モデルの適合度をチェック

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.198	80.2	80.2	.407
2	.048	19.4	99.6	.214
3	.001	.4	100.0	.031

図 279. 固有値

モデルで明示されるほぼすべての分散は、最初の 2 種類の判別関数に起因します。3 種類の関数が自動的に適合されますが、固有値が非常に小さいため、3 つ目はほぼ支障なく無視することができます。

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.796	227.345	9	.000
2 through 3	.953	47.486	4	.000
3	.999	.929	1	.335

図 280. Wilks のラムダ

Wilks のラムダは、最初の 2 種類の関数のみが有用であるとしています。各関数のセットにおいて、これはリストされた関数の手段がグループにわたって相当するという仮説をテストします。関数 3 のテストは 0.10 を上回る有意確率の値を備えているため、この関数はモデルにほとんど寄与しません。

構造行列

	Function		
	1	2	3
Level of education	.966*	-.090	-.244
Years with current employer	-.182	.964*	-.193
Age in years[a]	-.162	.598*	-.285
Household income in thousands[a]	.109	.514*	-.190
Years at current address[a]	-.151	.394*	-.214
Retired[a]	-.108	.230*	-.137
Gender[a]	.008	.054*	.009
Number of people in household	.232	.097	.968*
Marital status[a]	.132	.134	.600*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

図 281. 構造行列

複数の判別関数がある場合、標準関数の 1 つを用いて各変数で絶対値が最大の相関にアスタリスク (*) マークを付けます。 各関数において、マークを付けられた変数は相関のサイズによって並べ替えられます。

- *教育水準* は第 1 の関数と最も強い相関性があり、これはこの関数と最も強い相関性のある唯一の変数です。
- *現在の会社の勤務年数*、*年齢*、*世帯収入 (千単位)*、*現住所の居住年数*、*退職者*、および*性別* は第 2 の関数と最も強い相関性があるが、*性別* および*退職者* は他に比べ相関性は弱いです。 その他の変数はこの関数を「安定度」関数としてマークを付けます。
- *世帯人数* および*配偶者の有無* は第 3 の判別関数と最も強い相関性がありますが、これは無価値な関数であるため、ほとんど無効な予測値です。

地域マップ

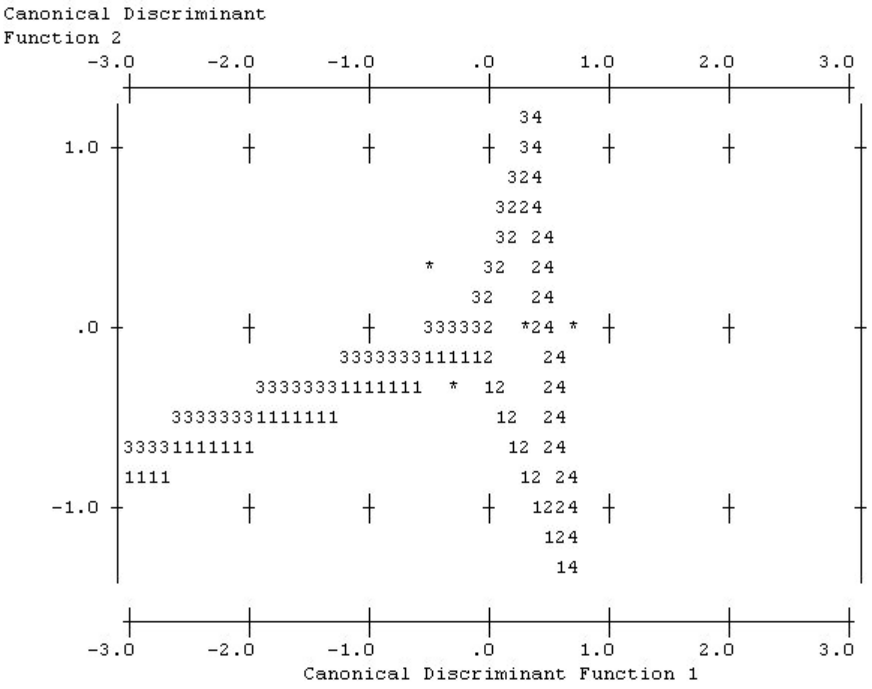


図 282. 地域マップ

地域マップは、グループと判別関数の関係を研究するのに役立ちます。構造行列の結果と組み合わせると、予測値とグループの関係が図で説明されます。横軸に示される第 1 関数は、グループ 4 (総合サービス の顧客) をその他から分離させます。 *教育水準* は第 1 の関数と強く正の相関性があるため、これは *総合サービス* の顧客が一般的に最も教育水準の高いことを示唆しています。 第 2 の関数はグループ 1 および 3 (*基本サービス* および *付加サービス* の顧客) を分離させます。*付加サービス* の顧客は、*基本サービス* の顧客よりも勤務期間が長く、年齢が高い傾向にあります。 *E-サービス* の顧客はその他の顧客からはあまり分離されず、マップは教育水準が高く、実務経験が中程度であることを示唆しています。

一般に、アスタリスク (*) マークの付いたグループの重心の地域線に対する近さは、すべてのグループ間の分離はあまり強くないことを示唆します。

最初の 2 種類の判別関数のみが示されていますが、第 3 の関数はそれほど重要ではないということがわかったため、地域マップは判別モデルの包括的な図を示します。

分類結果

			Predicted Group Membership				
		Customer category	Basic service	E-service	Plus service	Total service	Total
Original	Count	Basic service	125	11	61	69	266
		E-service	49	15	58	95	217
		Plus service	102	14	112	53	281
		Total service	40	16	37	143	236
	%	Basic service	47.0	4.1	22.9	25.9	100.0
		E-service	22.6	6.9	26.7	43.8	100.0
		Plus service	36.3	5.0	39.9	18.9	100.0
		Total service	16.9	6.8	15.7	60.6	100.0

a. 39.5% of original grouped cases correctly classified.

図 283. 分類結果

Wilks のラムダから、このモデルは予想していたよりも優れていることがわかりますが、どの程度優れているかを判断するには分類結果を調べる必要があります。監視データを考慮して、「Null」モデル (つまり、予測のないモデル) はすべての顧客をモーダル・グループである *付加サービス* に分類します。したがって、Null モデルは時間の 281/1000 = 28.1% が適切です。モデルは顧客の 11.4% から 39.5% を取得します。特に、 モデルは *トータル・サービス* の顧客を識別することに優れています。 ただし、*E*-サービスの顧客を分類するジョブは非常に苦手としています。これらの顧客を分類するには、別の予測値を見つける必要がある場合があります。

要約

それぞれの顧客からの人口統計情報に基づき、定義済みの 4 種類の「サービス使用」グループの 1 つに顧客を分類する判別モデルを作成しました。構造行列および地域マップを使用して、顧客ベースをセグメント化するのにどの変数が最も役立つかを特定しました。最後に、分類結果はモデルが *E*-サービス の顧客を分類するには向いていないことを示しています。 これらの顧客を適切に分類する別の予測値変数を判断するにはさらなる調査が必要ですが、予測しようとしている内容によっては、モデルはニーズを完全に満たすことができる場合があります。 例えば、*E*-サービス の顧客を識別することに関心がない場合、モデルは十分に正確である可能性があります。これは、E-サービスがほとんど利益を生み出さない特売商品である場合のことです。例えば、投資に対する最高の投資収益率が*付加サービス* または*総合サービス* の顧客からもたらされる場合、モデルは必要な情報を提供することがあります。

これらの結果は学習データのみに基づいていることに注意してください。モデルが他のデータに対してどの程度一般化されているかを評価するには、データ区分ノードを使用して、テストおよび検証用にレコードのサブセットを提供します。

IBM SPSS Modeler で使用されるモデル作成手法の数学的な基礎の説明は、「IBM SPSS Modeler アルゴリズム・ガイド」を参照してください。このガイドは、インストール・ディスクの「*¥Documentation*」ディレクトリーにあります。

第 22 章 区間打ち切り生存データの分析 (一般化線型モデル)

区間打ち切り生存データを分析して (つまり、対象となるイベントの正確な時刻は不明で、指定の区間内に発生したものしか分からない場合)、区間におけるイベントのハザードに Cox モデルを適用している場合、補数対数-対数回帰モデルが生じます。

潰瘍の再発を防ぐための 2 種類の治療の有効性を比較するために設計された研究の部分情報は、*ulcer_recurrence.sav* に収集されます。 このデータ・セットは、他の場所[1]で表示および分析されています。一般化線型モデルを使用すると、補数対数-対数回帰モデルの結果を複製できます。

この例では *ulcer_genlin.str* というストリームを使用し、*ulcer_recurrence.sav* というデータ・ファイルを参照します。 データ・ファイルは *Demos* フォルダーにあり、ストリーム・ファイルは *streams* サブフォルダーにあります。

ストリームの作成

1. *Demos* フォルダーの *ulcer_recurrence.sav* を指し示す Statistics ファイル入力ノードを追加します。

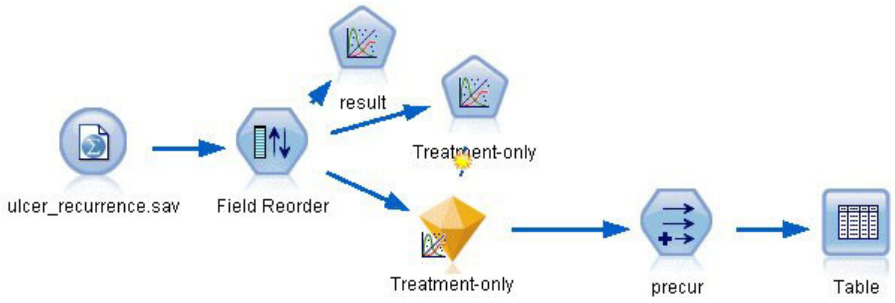


図 284. 潰瘍の再発を予測するサンプル・ストリーム

2. ソース・ノードの「フィルター」タブで、「*ID*」および「*時刻*」を除外します。

1. Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

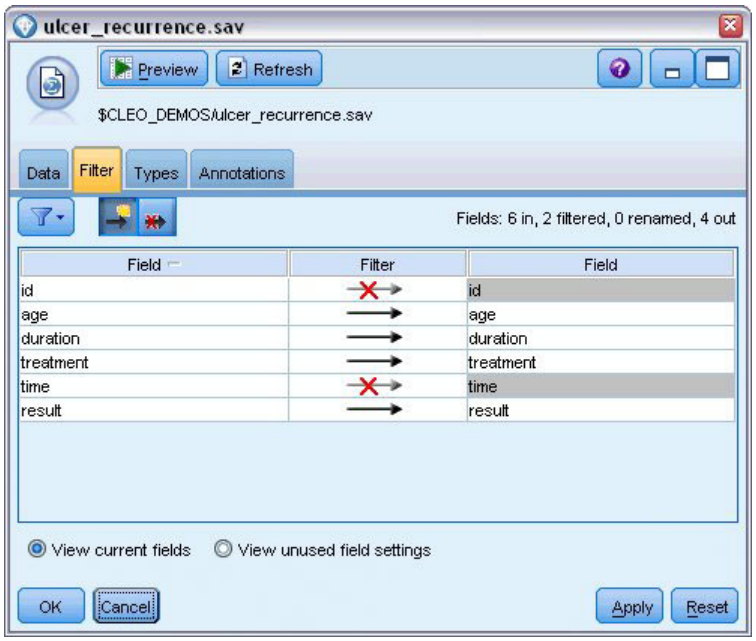


図 285. 不要なフィールドのフィルター

3. ソース・ノードの「タイプ」タブで、「*result*」フィールドの役割に「**対象**」を設定し、測定の尺度に「**フラグ**」を設定します。 1 の結果は、潰瘍が再発していることを示します。 その他のフィールドの役割は、すべて「**入力**」に設定します。
4. 「**値の読み取り**」をクリックし、データをインスタンス化します。

図 286. フィールドの役割の設定

5. フィールドの並べ替えノードを追加し、「*期間*」、「*治療*」、および「*年齢*」を入力順に指定します。これによりフィールドをモデルに入力する順番が決まり、Collett の結果を複製できるようにします。

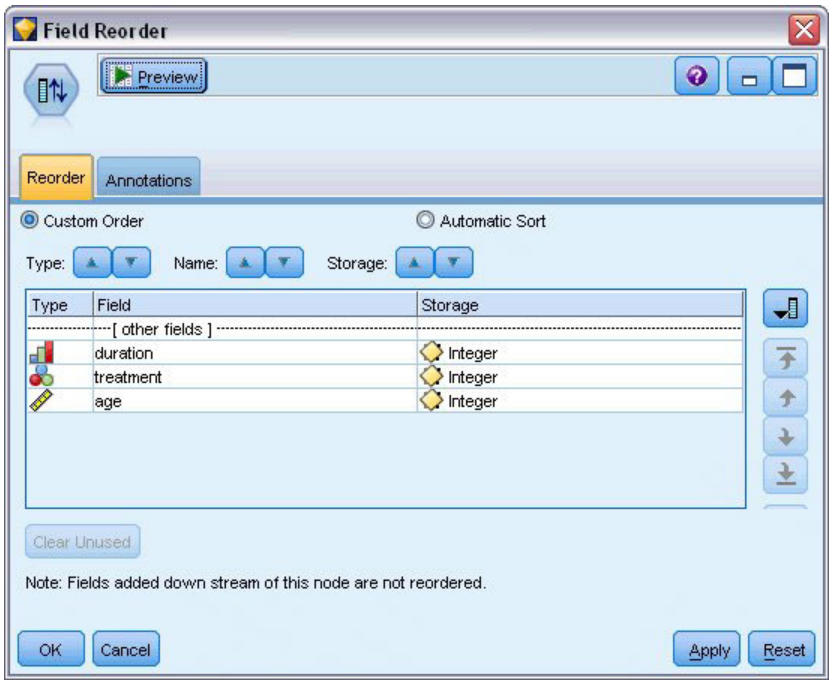


図 287. 希望どおりにフィールドをモデルに入力するためのフィールドの並べ替え

6. 一般化線型ノードをソース・ノードに接続します。一般化線型ノードで「**モデル**」タブをクリックします。
7. 「**最初 (最小)**」を対象の参照カテゴリーとして選択します。 これは、第 2 のカテゴリーが対象となるイベントであり、モデルにおけるこの効果はパラメーター推定値の解釈に表れることを示します。正の係数を持つ連続型予測値は、予測値が増加する再発の確率の増加を示します。大きな係数を持つ名義型予測のカテゴリーは、セットのその他のカテゴリーに対して再発の確率の増加を示します。

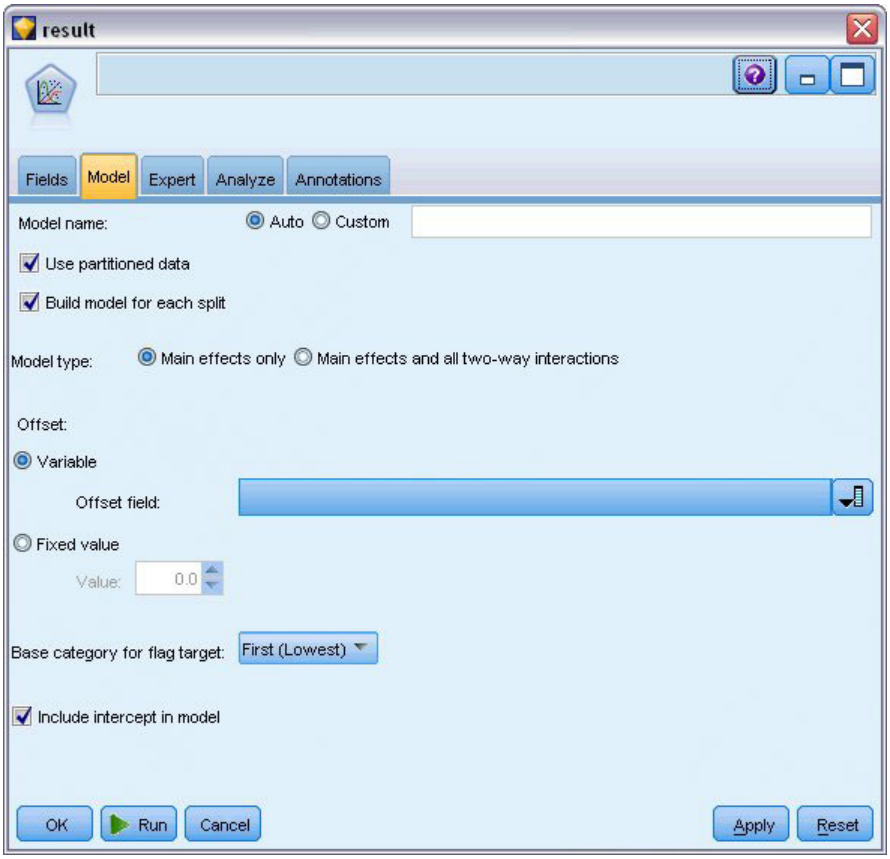


図 288. モデル・オプションの選択

8. 「**エキスパート**」タブをクリックし、「**エキスパート**」を選択して、エキスパート・モデル作成オプションを有効にします。
9. 「**2 項**」を分布として、「**補数対数-対数**」をリンク関数として選択します。
10. 「**固定値**」をスケール・パラメーターを推定する方法として選択し、デフォルト値 1.0 のままにします。
11. 「**降順**」を因子のカテゴリー順として選択します。 これは、それぞれの因子の第 1 のカテゴリーが参照カテゴリーになるということを示します。モデルでのこの選択の効果は、パラメーター推定値の解釈に表れます。

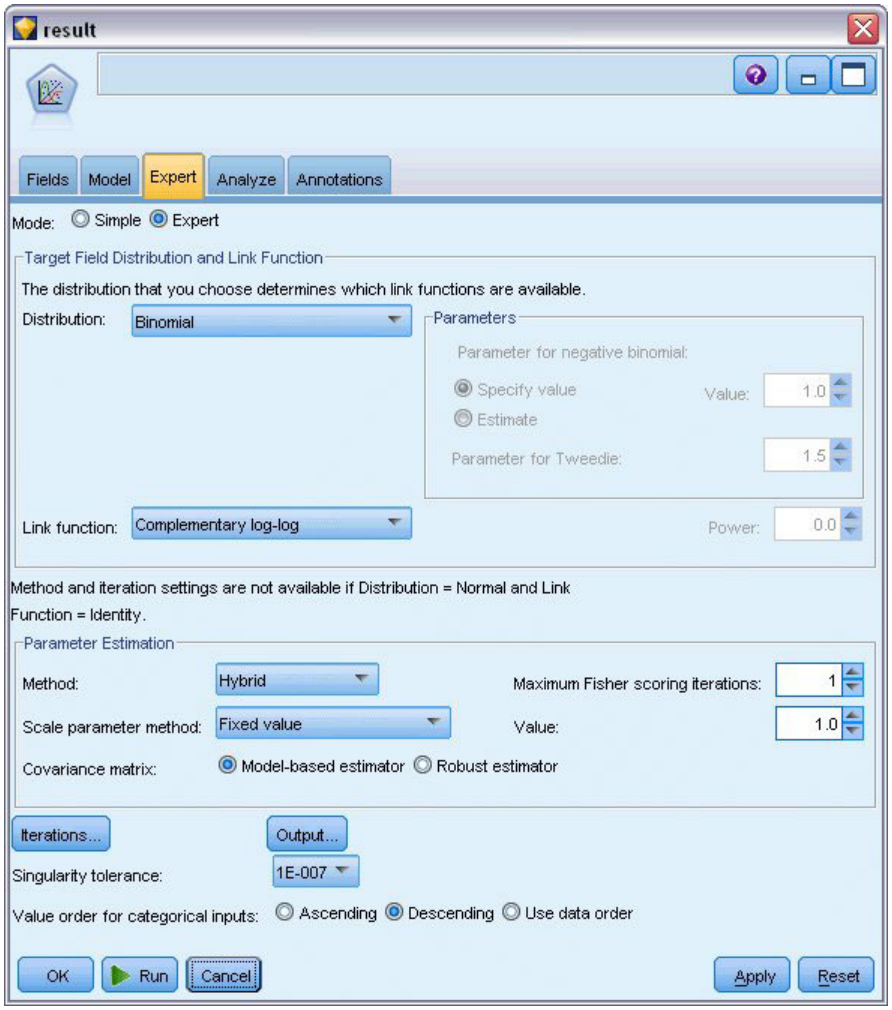


図 289. エキスパート・オプションの選択

12. ストリームを実行してモデル・ナゲットを作成します。これはストリーム領域および右上隅の「モデル」パレットに追加されます。モデルの詳細を表示するには、ナゲットを右クリックし、「**編集**」または「**参照**」を選択します。

モデル効果の検定

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	.536	1	.464
duration	.003	1	.958
treatment	.382	1	.537
age	.358	1	.550

Dependent Variable: Result
Model: (Intercept), duration, treatment, age

図 290. 主効果モデルのモデル効果検定

統計的に有意なモデル効果はありません。ただし、治療効果における目立った相違は臨床的に関心があるため、治療のみの縮小モデルをモデル項として適合させます。

治療のみのモデルの適合

1. 一般化線型ノードの「フィールド」タブで、「**カスタム設定を使用**」をクリックします。
2. 「*結果*」を対象として選択します。
3. 「*治療*」を単一の入力として選択します。

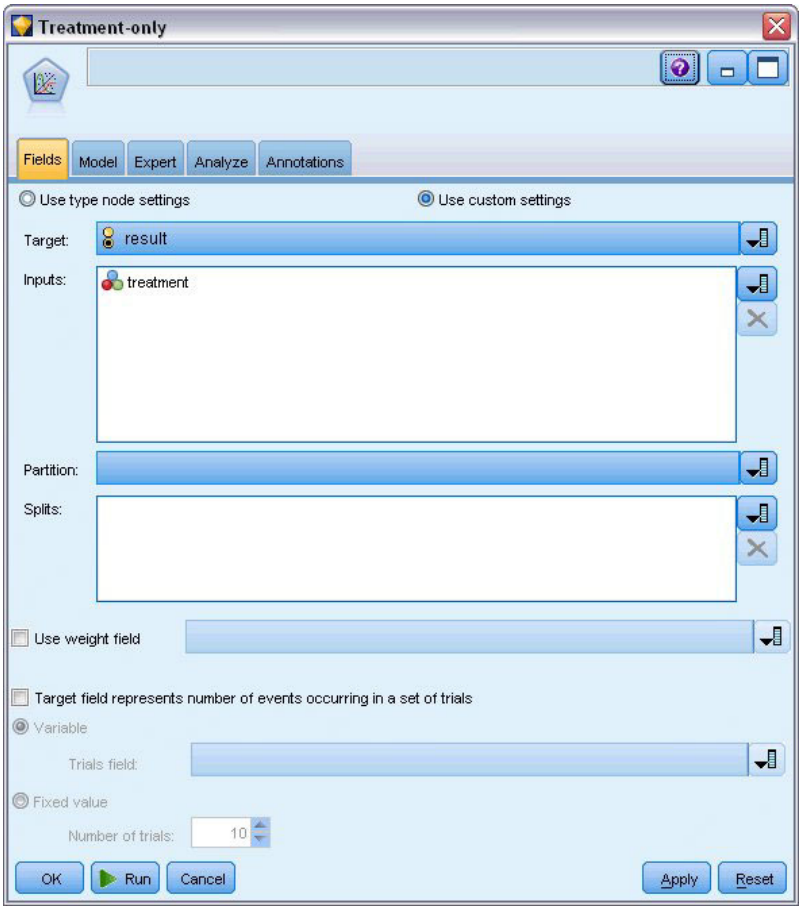


図 291. フィールド・オプションの選択

4. ストリームを実行して、作成されたモデル・ナゲットを開きます。

モデル・ナゲットで、「**詳細**」タブを選択して下部までスクロールします。

パラメーター推定値

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1.442	.5012	-2.425	-.460	8.282	1	.004
[treatment=1]	.378	.6288	-.855	1.610	.361	1	.548
[treatment=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result
Model: (Intercept), treatment

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

図 292. 治療のみのモデルに関するパラメーター推定値

治療効果 (2 種類の治療レベル間での線型予測値の差。つまり、*[treatment=1]* の係数) は統計的にまだ有意ではありませんが、治療 *B* のパラメーター推定値が *A* の値よりも大きく、そのため最初の 12 カ月間の再発確率が増加しているため、治療 *A[treatment=0]* が *B [treatment=1]* よりも優れていると示唆されます。 線型予測値 (切片 + 治療効果) は、$P(recur_{12,\ t})$ が 12 カ月間の治療 t(=*A* or *B*) の後の再発確率である場合の $\log(-\log(1-P(recur_{12,t}))$ の推定値です。 これらの予測確率は、データ・セットのそれぞれの観測に対して生成されます。

再発および生存の予測確率

図 293. フィールド作成ノードの設定オプション

1. それぞれの患者について、モデルは予測結果とその予測結果の確率をスコアリングします。予測した再発確率を確認するには、生成したモデルをパレットにコピーしてフィールド作成ノードを接続します。
2. 「設定」タブで、派生フィールドとして「precur」を入力します。
3. 「**条件付き**」として派生させるように選択します。
4. 「計算器」ボタンをクリックし、「**If**」条件の式ビルダーを開きます。

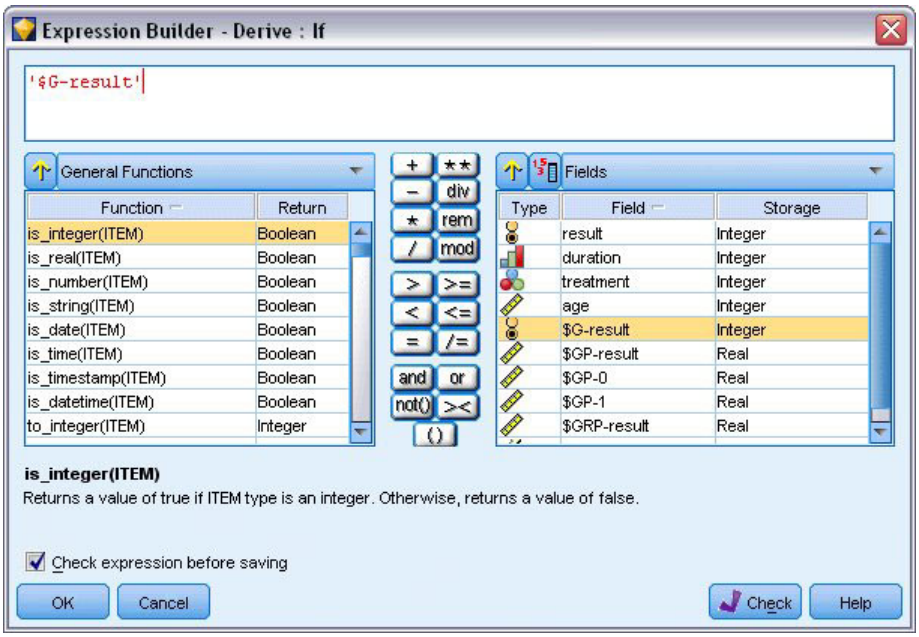


図 294. フィールド作成ノード: If 条件の式ビルダー

5. 「*$G-result*」フィールドを式に挿入します。
6. 「**OK**」をクリックします。

フィールド作成ノード「*precur*」は、「*$G-result*」が 1 の場合は「**Then**」式の値を、0 の場合は「**Else**」式の値を取得します。

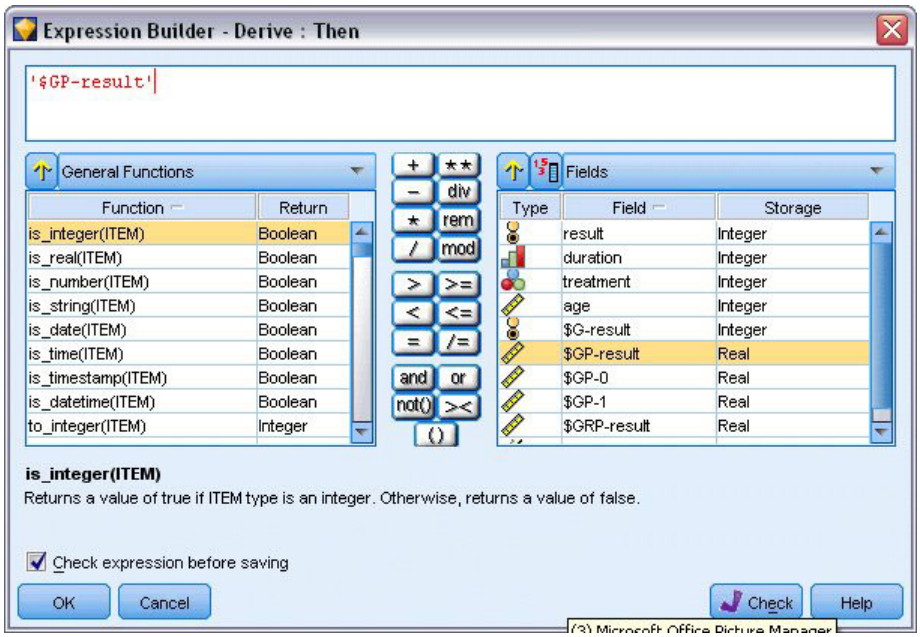


図 295. フィールド作成ノード: Then 式の式ビルダー

7. 「計算器」ボタンをクリックし、「**Then**」式の式ビルダーを開きます。
8. 「*$GP-result*」フィールドを式に挿入します。
9. 「**OK**」をクリックします。

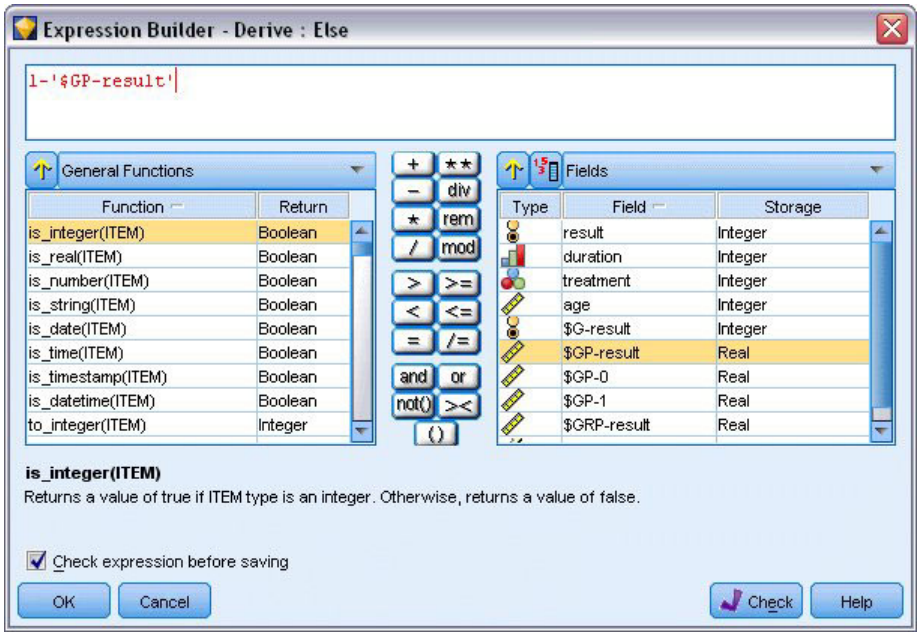
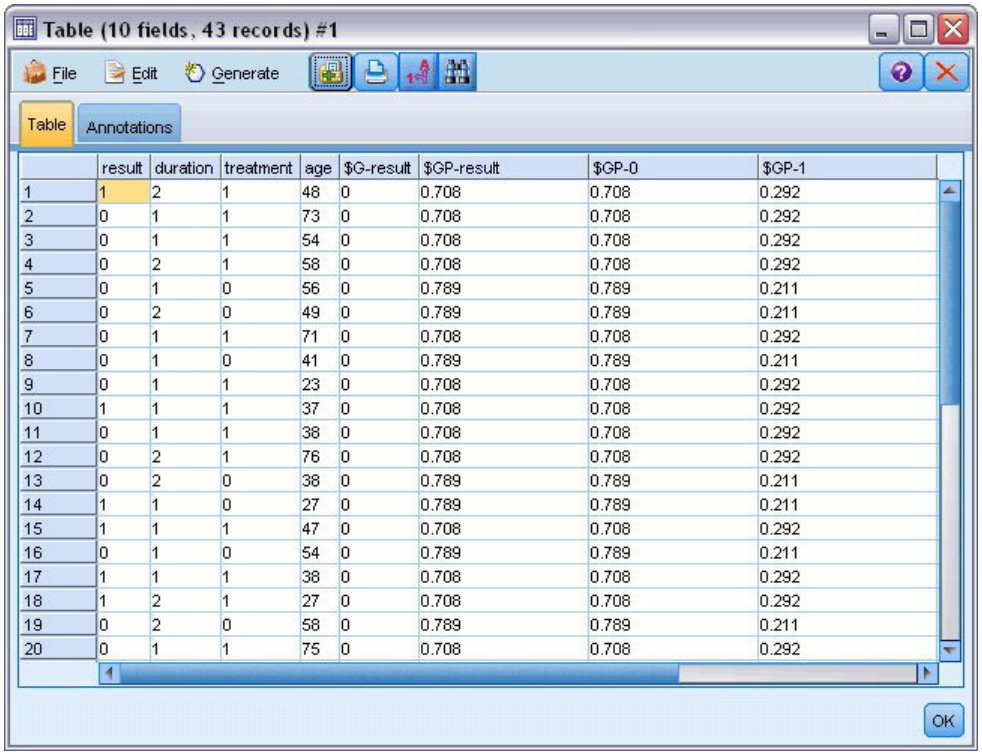


図 296. フィールド作成ノード: Else 式の式ビルダー

10. 「計算器」ボタンをクリックし、「**Else**」式の式ビルダーを開きます。
11. 「1-」を式に入力して、「*$GP-result*」フィールドを式に挿入します。
12. 「**OK**」をクリックします。

図 297. フィールド作成ノードの設定オプション

13. テーブル・ノードをフィールド作成ノードに接続して実行します。



	result	duration	treatment	age	$G-result	$GP-result	$GP-0	$GP-1
1	1	2	1	48	0	0.708	0.708	0.292
2	0	1	1	73	0	0.708	0.708	0.292
3	0	1	1	54	0	0.708	0.708	0.292
4	0	2	1	58	0	0.708	0.708	0.292
5	0	1	0	56	0	0.789	0.789	0.211
6	0	2	0	49	0	0.789	0.789	0.211
7	0	1	1	71	0	0.708	0.708	0.292
8	0	1	0	41	0	0.789	0.789	0.211
9	0	1	1	23	0	0.708	0.708	0.292
10	1	1	1	37	0	0.708	0.708	0.292
11	0	1	1	38	0	0.708	0.708	0.292
12	0	2	1	76	0	0.708	0.708	0.292
13	0	2	0	38	0	0.789	0.789	0.211
14	1	1	0	27	0	0.789	0.789	0.211
15	1	1	1	47	0	0.708	0.708	0.292
16	0	1	0	54	0	0.789	0.789	0.211
17	1	1	1	38	0	0.708	0.708	0.292
18	1	2	1	27	0	0.708	0.708	0.292
19	0	2	0	58	0	0.789	0.789	0.211
20	0	1	1	75	0	0.708	0.708	0.292

図 298. 予測確率

治療 *A* を割り当てた患者は最初の 12 カ月に再発する予測確率は 0.211、治療 *B* では 0.292 です。$1–P(recur_{12,\ t})$ は 12 カ月後の生存確率であり、生存アナリストにおいてはより関心が高い場合があります。

期間による再発確率のモデル作成

現状でのモデルでの問題は、最初の調査時に収集した情報が無視されることです。すなわち、多くの患者は最初の 6 カ月間に再発しませんでした。「より優れた」モデルは、各区間中にイベントが発生したかどうかを記録する 2 値の回答をモデル化します。このモデルに適合するには、元のデータ・セットの復元が必要です。これは *ulcer_recurrence_recoded.sav* で見つけることができます。 このファイルには、次の 2 種類の追加変数が含まれます。

- *期間*、ケースが最初の調査期間または 2 番目に対応するかどうかを記録します。
- *期間の結果*、指定の期間中に指定の患者に再発があるかどうかを記録します。

それぞれの元のケース (患者) は、リスク集合に含まれる区間ごとに 1 つのケースに寄与します。したがって、例えば、患者 1 は 2 つのケースに寄与します。この場合、1 つのケースは再発が見られなかった最初の調査期間に関するものであり、もう 1 つのケースは再発が記録された 2 番目の調査期間に関するものです。一方、最初の期間に再発が記録されたため、患者 10 は単一のケースに寄与します。患者 16、28、および 34 は、6 カ月後に研究から外れたため、新しいデータ・セットの単一のケースにのみに寄与します。

1. *Demos* フォルダーの *ulcer_recurrence_recoded.sav* を指し示す Statistics ファイル入力ノードを追加します。

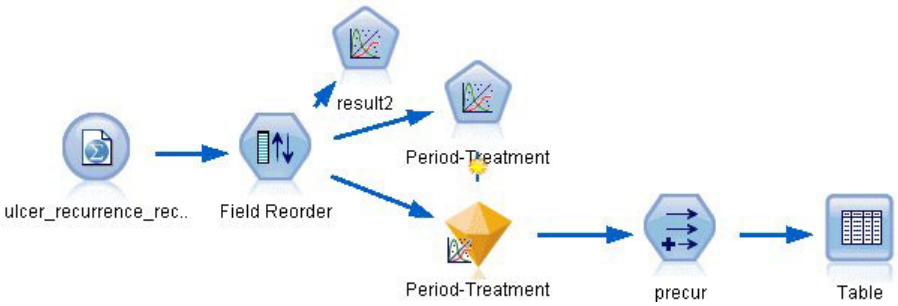


図 299. 潰瘍の再発を予測するサンプル・ストリーム

2. ソース・ノードの「フィルター」タブで、「*ID*」、「*時刻*」、および「*結果*」を除外します。

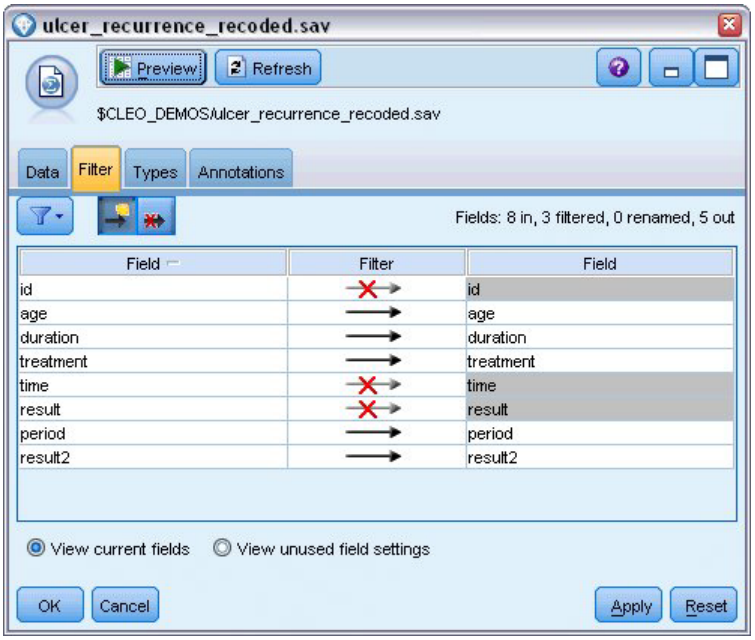


図 300. 不要なフィールドのフィルター

3. ソース・ノードの「タイプ」タブで、「*result2*」フィールドの役割に「**対象**」を設定し、測定の尺度に「**フラグ**」を設定します。 その他のフィールドの役割は、すべて「**入力**」に設定します。

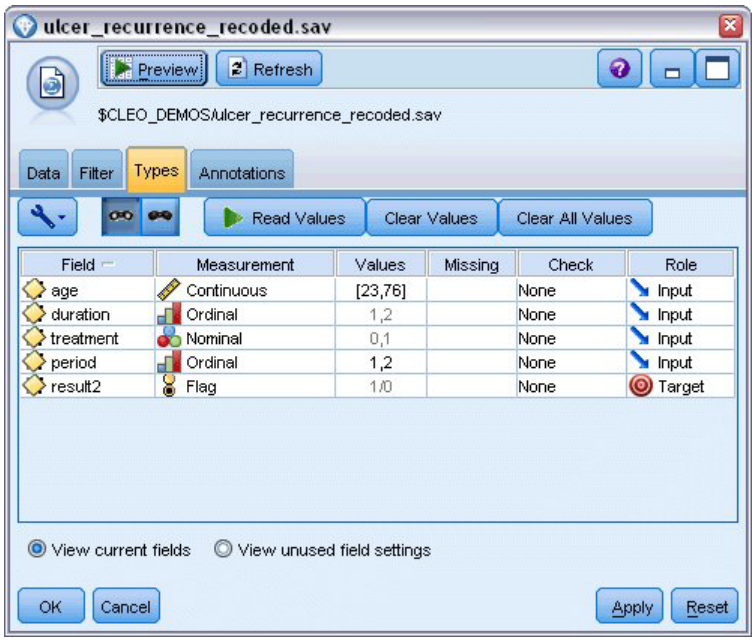


図 301. フィールドの役割の設定

4. フィールドの並べ替えノードを追加し、「*期間*」、「*所要時間*」、「*治療*」、および「*年齢*」を入力順に指定します。「*期間*」を第 1 の入力として作成する (モデルの切片項は含めない) と、ダミー変数のフルセットを適合して期間効果を取り込めるようになります。

図 302. 希望どおりにフィールドをモデルに入力するためのフィールドの並べ替え

5. 一般化線型ノードで、「**モデル**」タブをクリックします。

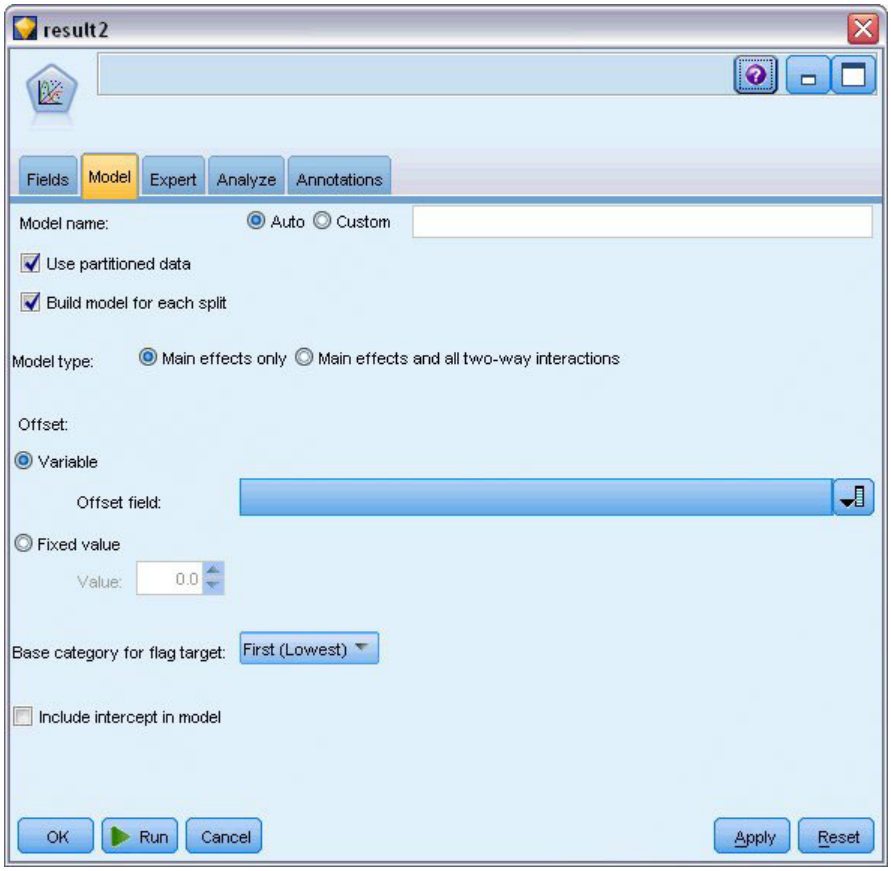


図 303. モデル・オプションの選択

6. 「**最初 (最小)**」を対象の参照カテゴリーとして選択します。 これは、第 2 のカテゴリーが対象となるイベントであり、モデルにおけるこの効果はパラメーター推定値の解釈に表れることを示します。
7. 「**モデル内に切片を含む**」を選択解除します。
8. 「**エキスパート**」タブをクリックし、「**エキスパート**」を選択して、エキスパート・モデル作成オプションを有効にします。

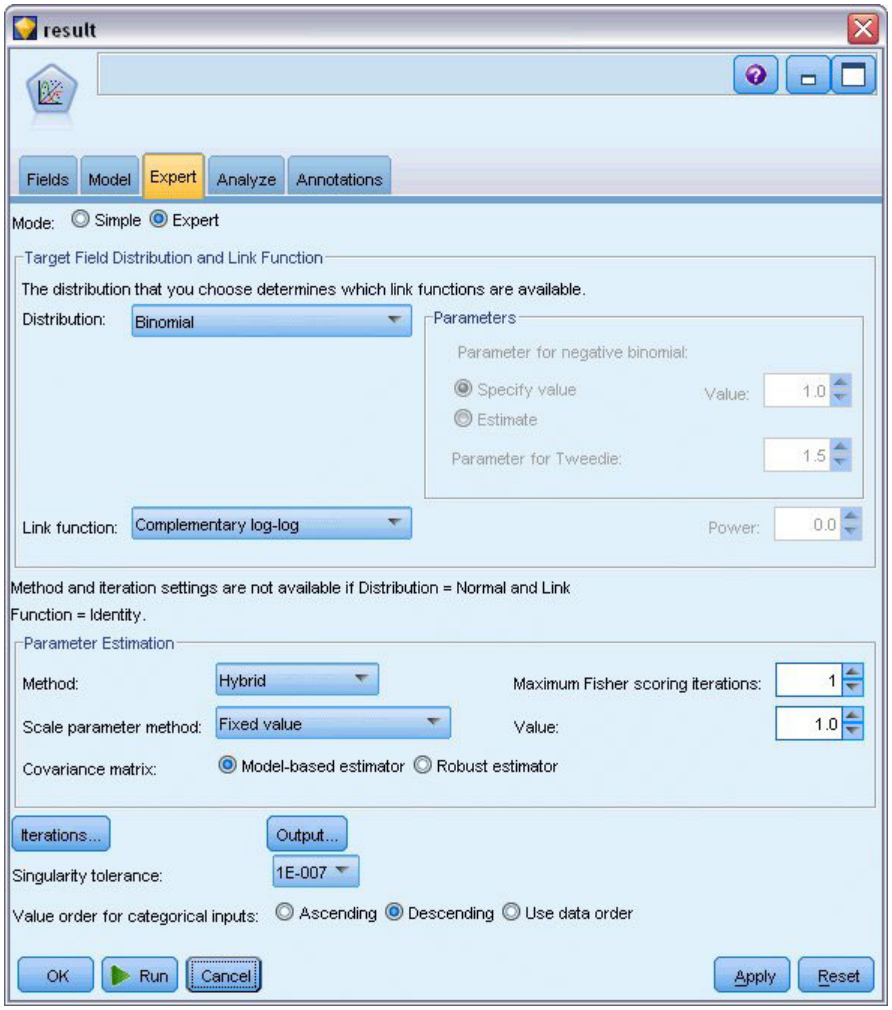


図 304. エキスパート・オプションの選択

9. 「**2 項**」を分布として、「**補数対数-対数**」をリンク関数として選択します。
10. 「**固定値**」をスケール・パラメーターを推定する方法として選択し、デフォルト値 1.0 のままにします。
11. 「**降順**」を因子のカテゴリー順として選択します。 これは、それぞれの因子の第 1 のカテゴリーが参照カテゴリーになるということを示します。モデルでのこの選択の効果は、パラメーター推定値の解釈に表れます。
12. ストリームを実行してモデル・ナゲットを作成します。これはストリーム領域および右上隅の「モデル」パレットに追加されます。モデルの詳細を表示するには、ナゲットを右クリックし、「**編集**」または「**参照**」を選択します。

モデル効果の検定

Source	Type III		
	Wald Chi-Square	df	Sig.
period	.464	1	.496
duration	.000	1	.988
treatment	.117	1	.732
age	.314	1	.575

Dependent Variable: Result by period
Model: period, duration, treatment, age

図 305. 主効果モデルのモデル効果検定

統計的に有意なモデル効果はありません。ただし、治療効果における目立った相違は臨床的に関心があるため、これらのモデル項のみで縮小モデルを適合させます。

縮小モデルの適合

1. 一般化線型ノードの「フィールド」タブで、「**カスタム設定を使用**」をクリックします。
2. 「*result2*」を対象として選択します。
3. 「*期間*」および「*治療*」を入力として選択します。

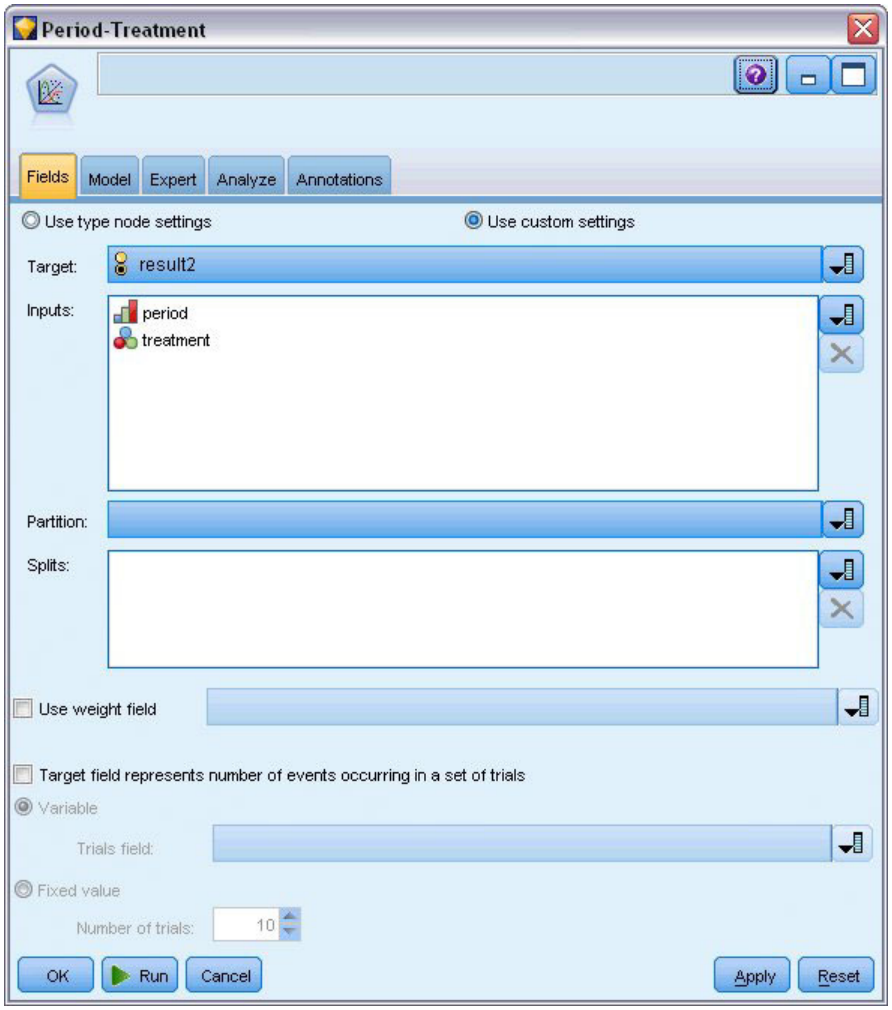


図 306. フィールド・オプションの選択

4. ノードを実行して生成されたモデルを参照し、生成されたモデルをパレットにコピーしてテーブル・ノードを接続し、それを実行します。

パラメーター推定値

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
[period=2]	-1.794	.5792	-2.929	-.659	9.597	1	.002
[period=1]	-2.206	.5912	-3.365	-1.047	13.926	1	.000
[treatment=1]	.195	.6279	-1.035	1.426	.097	1	.756
[treatment=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result by period
Model: period, treatment

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

図 307. 治療のみのモデルに関するパラメーター推定値

治療効果は統計的にまだ有意ではありませんが、治療 *B* のパラメーター予測値が最初の 12 カ月間の再発確率の増加に関連しているため、*A* は *B* より優れていると示唆されます。 期間の値は 0 とは異なり統計的に有意ですが、これは切片項が適合していないという事実によるものです。期間の効果 (*[period=1]* と *[period=2]* の線型予測値の差) は、モデル効果の検定から分かるように、統計的に有意ではありません。線型予測値 (期間の効果 + 治療効果) は、$P(recur_{p,\ t})$ が p(=1 or 2, 6 カ月または 12 カ月を表す)の期間、指定の治療 t(=*A* or *B*) を行った後の再発確率である場合の $\log(-\log(1-P(recur_{p,\ t}))$ の推定値です。 これらの予測確率は、データ・セットのそれぞれの観測に対して生成されます。

再発および生存の予測確率

図 308. フィールド作成ノードの設定オプション

1. それぞれの患者について、モデルは予測結果とその予測結果の確率をスコアリングします。予測した再発確率を確認するには、生成したモデルをパレットにコピーしてフィールド作成ノードを接続します。
2. 「設定」タブで、派生フィールドとして「precur」を入力します。
3. 「**条件付き**」として派生させるように選択します。
4. 「計算器」ボタンをクリックし、「**If**」条件の式ビルダーを開きます。

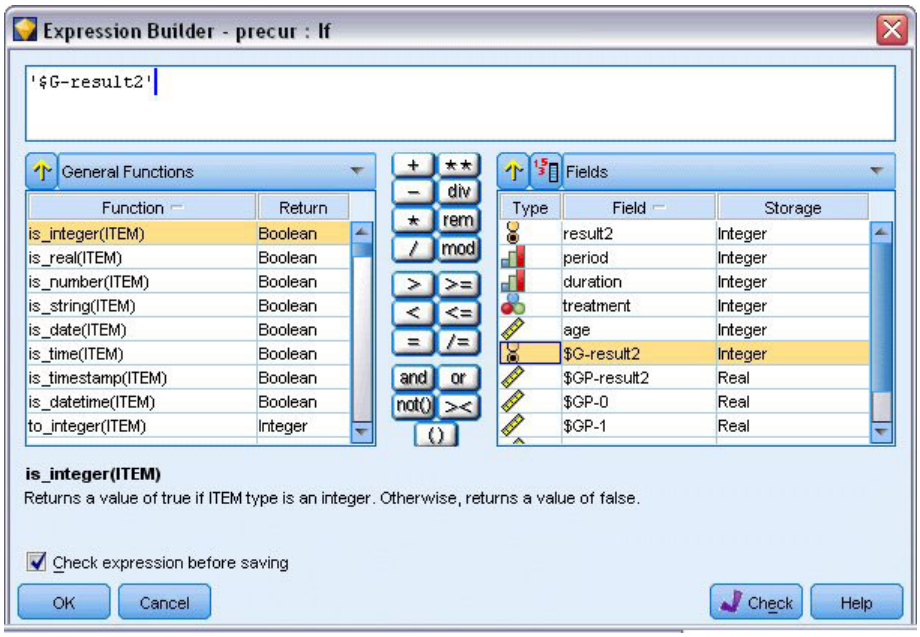


図 309. フィールド作成ノード: If 条件の式ビルダー

5. 「*$G-result2*」フィールドを式に挿入します。
6. 「**OK**」をクリックします。

フィールド作成ノード「*precur*」は、「*$G-result2*」が 1 の場合は「**Then**」式の値を、0 の場合は「**Else**」式の値を取得します。

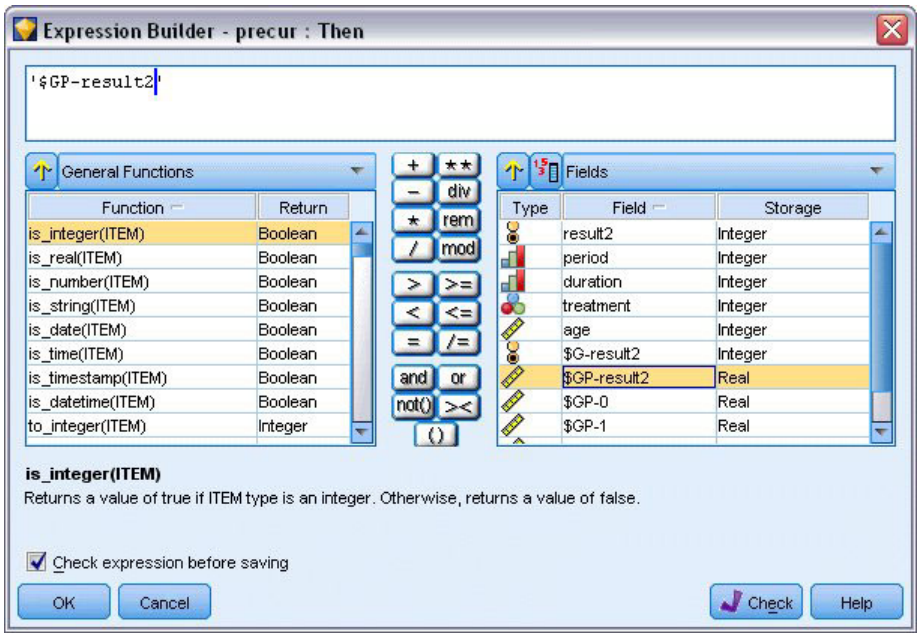


図 310. フィールド作成ノード: Then 式の式ビルダー

7. 「計算器」ボタンをクリックし、「**Then**」式の式ビルダーを開きます。
8. 「*$GP-result2*」フィールドを式に挿入します。

9. 「**OK**」をクリックします。

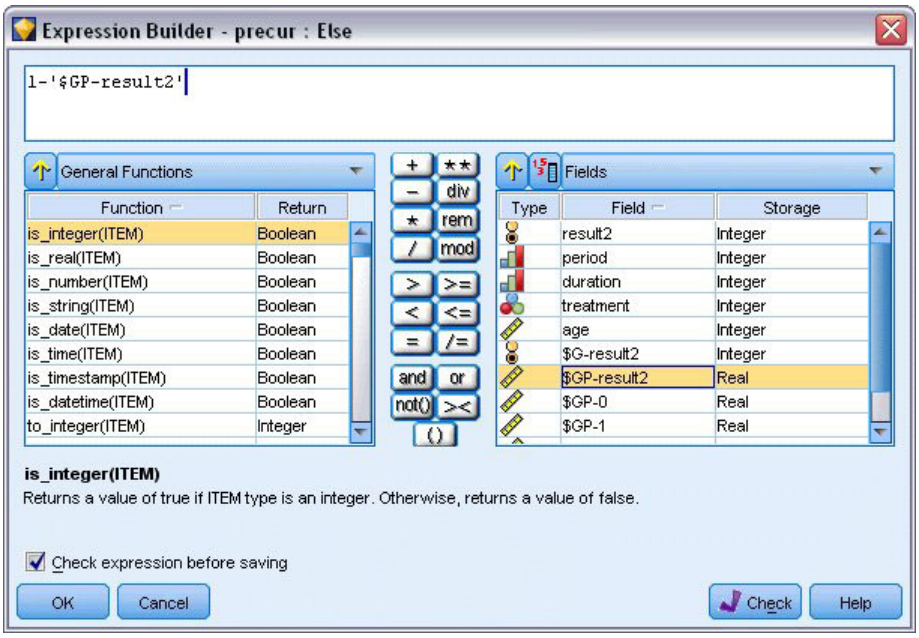
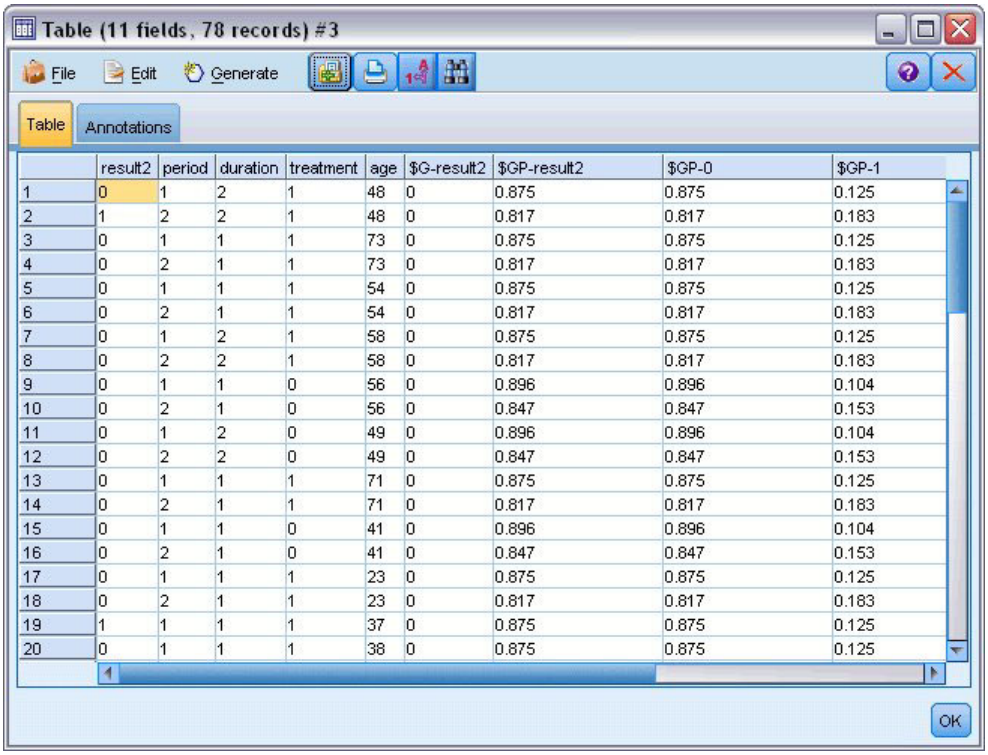


図 311. フィールド作成ノード: Else 式の式ビルダー

10. 「計算器」ボタンをクリックし、「**Else**」式の式ビルダーを開きます。
11. 「1-」を式に入力して、「*$GP-result2*」フィールドを式に挿入します。
12. 「**OK**」をクリックします。

図 312. フィールド作成ノードの設定オプション

13. テーブル・ノードをフィールド作成ノードに接続して実行します。



	result2	period	duration	treatment	age	$G-result2	$GP-result2	$GP-0	$GP-1
1	0	1	2	1	48	0	0.875	0.875	0.125
2	1	2	2	1	48	0	0.817	0.817	0.183
3	0	1	1	1	73	0	0.875	0.875	0.125
4	0	2	1	1	73	0	0.817	0.817	0.183
5	0	1	1	1	54	0	0.875	0.875	0.125
6	0	2	1	1	54	0	0.817	0.817	0.183
7	0	1	2	1	58	0	0.875	0.875	0.125
8	0	2	2	1	58	0	0.817	0.817	0.183
9	0	1	1	0	56	0	0.896	0.896	0.104
10	0	2	1	0	56	0	0.847	0.847	0.153
11	0	1	2	0	49	0	0.896	0.896	0.104
12	0	2	2	0	49	0	0.847	0.847	0.153
13	0	1	1	1	71	0	0.875	0.875	0.125
14	0	2	1	1	71	0	0.817	0.817	0.183
15	0	1	1	0	41	0	0.896	0.896	0.104
16	0	2	1	0	41	0	0.847	0.847	0.153
17	0	1	1	1	23	0	0.875	0.875	0.125
18	0	2	1	1	23	0	0.817	0.817	0.183
19	1	1	1	1	37	0	0.875	0.875	0.125
20	0	1	1	1	38	0	0.875	0.875	0.125

図 313. 予測確率

表 3. 再発の予測確率

治療	6 カ月	12 カ月
A	0.104	0.153
B	0.125	0.183

再発の予測確率から、12 カ月間の生存確率は、$1-(P(recur_{1,\ t}) + P(recur_{2,\ t})\times(1-P(recur_{1,\ t})))$ と推定できます。したがって、それぞれの治療は次のとおりです。

A: 1 − (0.104 + 0.153*0.896) = 0.759

B: 1 − (0.125 + 0.183*0.875) = 0.715

これは、*A* がより優れた治療として非統計的に有意なサポートであることを再度示しています。

要約

一般化線型モデルを使用して、区間打ち切り生存データに一連の補数対数-対数回帰モデルを適合させました。治療 *A* を選択するためのサポートがある一方、統計的に有意な結果を得るためにより大規模な研究が必要となる場合があります。 ただし、既存のデータを検討するさらなる手段がいくつかあります。

- 交互作用効果のあるモデルを最適合することが、特に「*期間*」と「*治療グループ*」の間では有益な場合があります。

IBM SPSS Modeler で使用されるモデル作成手法の数学的な基礎の説明は、「*IBM SPSS Modeler アルゴリズム・ガイド*」を参照してください。

関連手続き

一般化線型モデル手続きは、さまざまなモデルに適合する強力なツールです。

- 一般化推定方程式手続きは、一般化線型モデルを拡張して反復測定を可能にします。
- 線型混合モデル手続きを使用して、ランダム・コンポーネントまたは反復測定、またはその両方を持つスケール従属変数にモデルを適合させることができます。

推奨図書

一般化線型モデルの詳細については、次のテキストを参照してください。

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

第 23 章 ポワソン回帰を使用した船舶損傷率の分析 (一般化線型モデル)

一般化線型モデルは度数データを分析する場合にポワソン回帰を適合させるのに使用できます。 例えば、他の場所[2]を表示および分析するデータ・セットは、波による貨物船の損傷に関係します。 予測値が与えられると、ポワソン比率で発生するように事故カウントをモデル化することが可能であり、その結果として得られるモデルは、どの船舶タイプが最も損害を受けやすいかを判断する手助けになります。

この例では *ships_genlin.str* というストリームを使用し、*ships.sav* というデータ・ファイルを参照します。データ・ファイルは *Demos* フォルダーにあり、ストリーム・ファイルは *streams* サブフォルダーにあります。

未加工のセル度数のモデル化は、船舶タイプによってサービス集計月 が異なるため、この状態では誤った結果が導かれる可能性があります。 リスクの「露出」量を測定するこのような変数は、一般化線型モデル内でオフセット変数として処理されます。さらに、ポワソン回帰は従属変数の対数が予測では線型であると推定します。そのため、一般化線型モデルを使用してポワソン回帰を事故率に適合させるには、サービス集計月の対数 を使用する必要があります。

「過分散」ポワソン回帰の適合

1. *Demos* フォルダーの *ships.sav* を指し示す Statistics ファイル入力ノードを追加します。

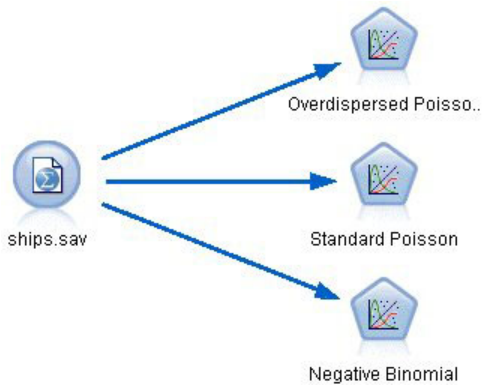


図 314. 損傷率を分析するサンプル・ストリーム

2. ソース・ノードの「フィルター」タブで、「*months_service*」フィールドを除外します。この変数の対数変換は、*log_months_service* に含まれ、分析に使用されます。

2. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

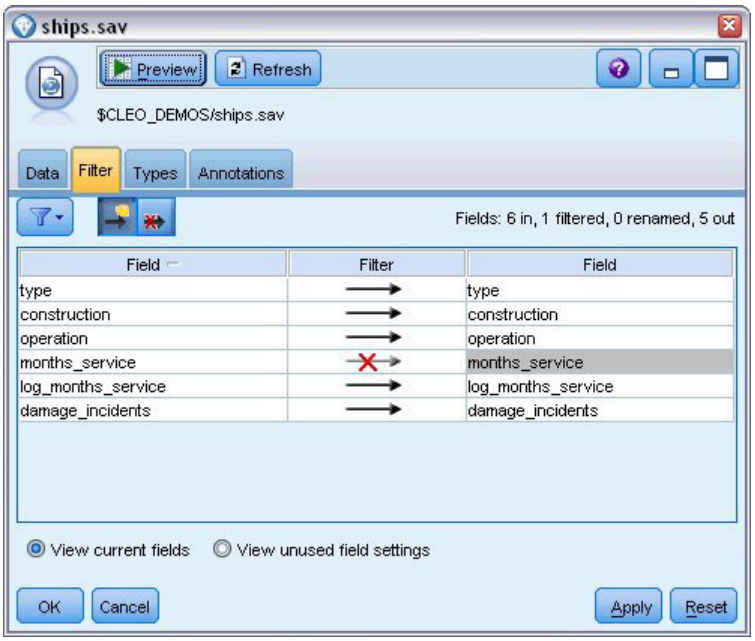


図 315. 不要なフィールドのフィルタリング

(あるいは、除外する代わりに、「タイプ」タブでこのフィールドの役割を「**なし**」に変更するか、モデル作成ノードで使用するフィールドを選択することもできます。)

3. ソース・ノードの「タイプ」タブで、*damage_incidents* フィールドの役割に「**対象**」を設定します。その他のフィールドの役割は、すべて「**入力**」に設定します。
4. 「**値の読み取り**」をクリックし、データをインスタンス化します。

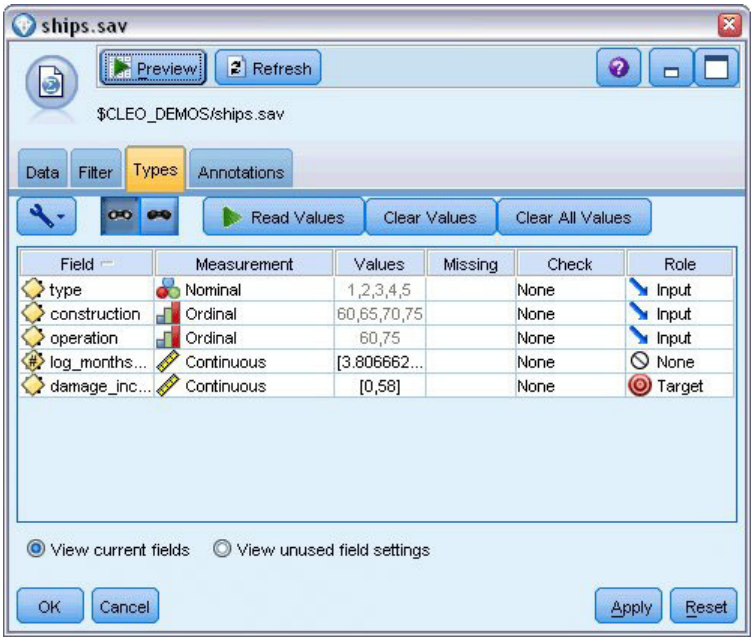


図 316. フィールドの役割の設定

5. 一般化線型ノードをソース・ノードに接続します。一般化線型ノードで「**モデル**」タブをクリックします。

6. *log_months_service* をオフセット変数として選択します。

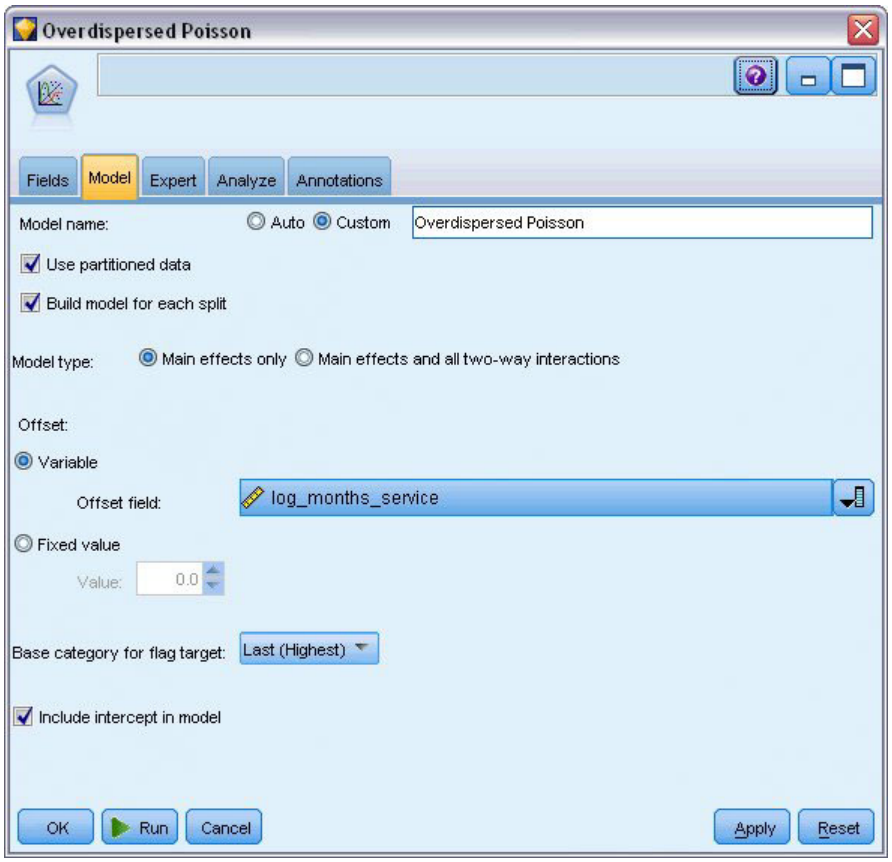


図 317. モデル・オプションの選択

7. 「**エキスパート**」タブをクリックし、「**エキスパート**」を選択して、エキスパート・モデル作成オプションを有効にします。

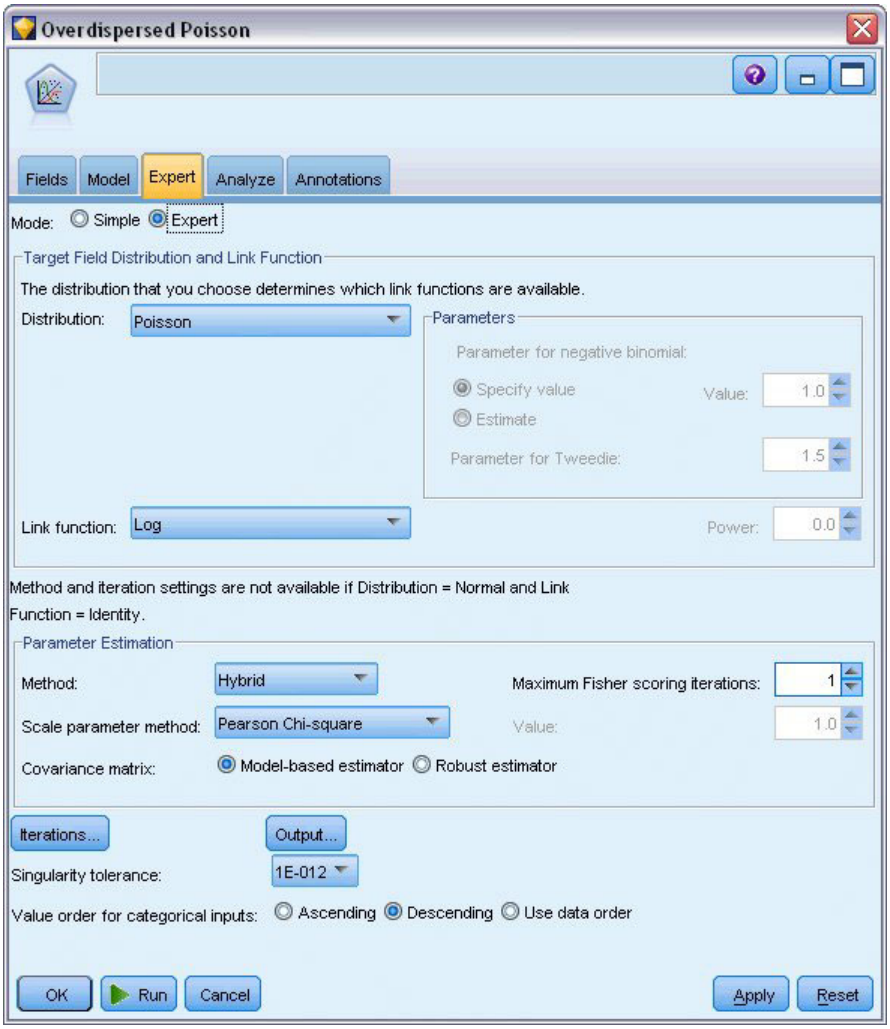


図 318. エキスパート・オプションの選択

8. 「**ポワソン**」を回答の分散として選択し、「**対数**」をリンク関数として選択します。
9. 「**Pearson カイ 2 乗**」をスケール・パラメーターを推定するための方法として選択します。 ポワソン回帰ではスケール・パラメーターを通常は 1 に想定していますが、McCullagh と Nelder は Pearson カイ 2 乗推定を使用してより控えめな分散推定値および有意水準を取得します。
10. 「**降順**」を因子のカテゴリー順として選択します。 これは、それぞれの因子の第 1 のカテゴリーが参照カテゴリーになるということを示します。モデルでのこの選択の効果は、パラメーター推定値の解釈に表れます。
11. 「**実行**」をクリックしてモデル・ナゲットを作成します。これはストリーム領域および右上隅のモデル・パレットに追加されます。 モデルの詳細を表示するには、ナゲットを右クリックし、「**編集**」または「**参照**」を選択して、「**詳細**」タブを選択します。

適合度統計

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	22.883	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	25.000	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

図 319. 適合度統計

適合度統計テーブルは、競合モデルを比較するのに役立つ指標を示しています。さらに、逸脱および Pearson カイ 2 乗統計の *Value/df* は、対応するスケール・パラメーターの推定値を示しています。これらの値は、ポワソン回帰に対してほぼ 1.0 になるはずです。これらの値が 1.0 を上回ると、過分散モデルの適合は妥当であるということになります。

オムニバス検定

Likelihood Ratio Chi-Square	df	Sig.
107.633	8	.000

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. Compares the fitted model against the intercept-only model.

図 320. オムニバス検定

オムニバス検定は、現行モデル対 Null モデル (この場合は切片) の尤度比カイ 2 乗検定です。0.05 未満の有意確率値は、現在のモデルが Null モデルよりも優れていることを示します。

モデル効果の検定

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	2138.657	1	.000
type	15.415	4	.004
construction	17.242	3	.001
operation	6.249	1	.012

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

図 321. モデル効果の検定

モデルの各項であらゆる効果の有無が検定されます。有意確率値が 0.05 未満の項には、明確な効果があります。それぞれの主効果項はモデルに寄与します。

パラメーター推定値

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-6.406	.2828	-6.960	-5.852	513.238	1	.000
[type=5]	.326	.3067	-.276	.927	1.127	1	.288
[type=4]	-.076	.3779	-.817	.665	.040	1	.841
[type=3]	-.687	.4279	-1.526	.151	2.581	1	.108
[type=2]	-.543	.2309	-.996	-.091	5.536	1	.019
[type=1]	0[a]	.	.	.	.	.	.
[construction=75]	.453	.3032	-.141	1.048	2.236	1	.135
[construction=70]	.818	.2208	.386	1.251	13.743	1	.000
[construction=65]	.697	.1946	.316	1.079	12.835	1	.000
[construction=60]	0[a]	.	.	.	.	.	.
[operation=75]	.384	.1538	.083	.686	6.249	1	.012
[operation=60]	0[a]	.	.	.	.	.	.
(Scale)	1.691[b]						

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

図 322. パラメーター推定値

パラメーター推定値テーブルは、各予測値の効果を要約したものです。リンク関数の特性のためにこのモデルでの係数の解釈は困難ですが、共変量の係数の符号および因子レベルの係数の相対値により、モデルにおける予測値の効果に対して重要な洞察を得ることができます。

- 共変量の場合、正 (負) の係数は、予測変数と結果が正 (逆) の関係にあることを示します。正の係数で共変量の値が増加すると、損傷事故率も増加します。
- 因子の場合、より大きな係数の因子レベルは、より高い損傷発生率を示します。因子レベルの係数の符号は、参照カテゴリーに対する因子レベルの効果によって異なります。

パラメーター推定値に基づいて、次のように解釈できます。

- 船舶タイプ *B [type=2]* は、参照カテゴリーのタイプ *A [type=1]* よりも統計的に有意な (*p* 値 0.019) 低い損傷率 (推定係数 –0.543) です。 タイプ *C [type=3]* は実際は *B* よりも小さい推定パラメーターですが、*C* の推定値のばらつきは効果を低下させます。因子レベル間のあらゆる関係については、推定周辺平均を参照してください。
- 1965 年から 1969 年 *[construction=65]* および 1970 年から 1974 年 *[construction=70]* に建造された船舶は、参照カテゴリーの 1960 年から 1964 年 *[construction=60]* に作られたものより統計的に有意な (*p* 値 <0.001) 高い損傷率 (推定係数はそれぞれ 0.697 および 0.818) です。 因子レベル間のあらゆる関係については、推定周辺平均を参照してください。
- 1975 年から 1979 年 *[operation=75]* に運用された船舶は、1960 年から 1974 年 *[operation=60]* に運用されたものより、統計的に有意な (*p* 値 0.012) 高い損傷率 (推定係数 0.384) です。

代替モデルの適合

「過分散」ポワソン回帰の 1 つの問題は、「標準の」ポワソン回帰に対してそれを検定する正式な方法がないということです。ただし、過分散があるかどうかを判断するために提案された 1 つの正式な検定は、その他すべての同じ設定について「標準の」ポワソン回帰と負の 2 項回帰の間で尤度比検定を実行するというものです。ポワソン回帰に過分散がない場合、統計値 –2×(ポワソン・モデルの対数尤度 – 負の 2 項モデルの対数尤度) は、半分は確率質量 0 の混合分布、残りは自由度 1 のカイ 2 乗分布がある必要があります。

1. 「**固定値**」をスケール・パラメーターを推定する方法として選択します。 デフォルトでは、この値は 1 です。

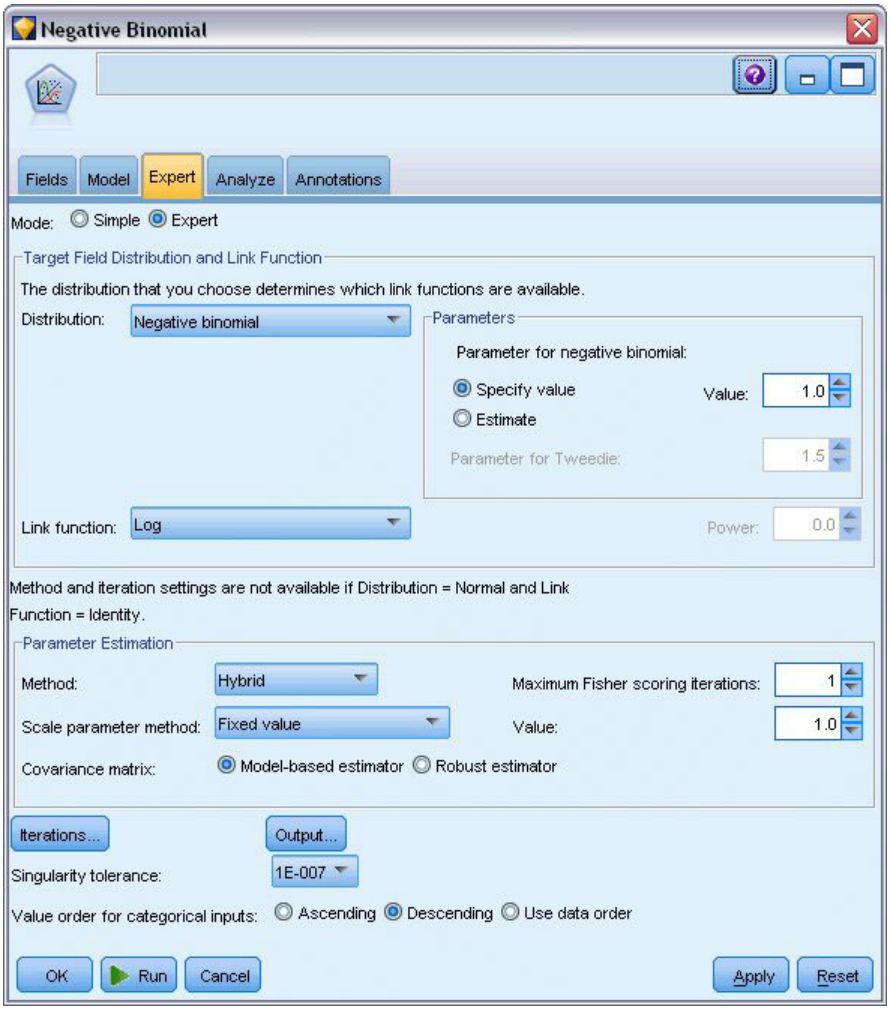


図 323. 「エキスパート」タブ

2. 負の 2 項回帰に適合させるには、一般化線型ノードをコピーして貼り付けし、ソース・ノードに接続し、新規ノードを開いて「**エキスパート**」タブをクリックします。
3. 「**負の 2 項**」を分布として選択します。 補助パラメーターをデフォルト値の 1 のままにします。
4. ストリームを実行して新しく作成されたモデル・ナゲットの「詳細」タブを参照します。

適合度統計

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	38.695	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	42.275	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

図 324. 標準ポワソン回帰の適合度統計

標準ポワソン回帰に報告される対数尤度は –68.281 です。 これを負の 2 項モデルと比較します。

	Value	df	Value/df
Deviance	11.145	25	.446
Scaled Deviance	11.145	25	
Pearson Chi-Square	8.815	25	.353
Scaled Pearson Chi-Square	8.815	25	
Log Likelihood[a]	-83.725		
Akaike's Information Criterion (AIC)	185.450		
Finite Sample Corrected AIC (AICC)	192.950		
Bayesian Information Criterion (BIC)	199.187		
Consistent AIC (CAIC)	208.187		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

図 325. 負の 2 項回帰の適合度統計

負の 2 項回帰に報告される対数尤度は –83.725 です。 これは実際はポワソン回帰における対数尤度より小さく、(尤度比テストは必要ありません) この負の 2 項回帰はポワソン回帰に比べて改善されないことを示しています。

ただし、負の 2 項分布の補助パラメーターに対して 1 の値を選択する場合、このデータ・セットには最適ではない場合があります。別の方法で過分散をテストするには、補助パラメーターが 0 である負の 2 項モデルに適合させて、、「エキスパート」タブの「出力」ダイアログで LaGrange 乗数検定を要求します。検定が有意でない場合、過分散はこのデータ・セットで問題にはなりません。

要約

一般化線型モデルを使用して、度数データに 3 種類のモデルを適合させました。負の 2 項回帰は、ポワソン回帰よりも優れているわけではないということが判明しました。過分散ポワソン回帰は標準ポワソン・モデルの妥当な代替案を示しているようですが、どちらを選択するかを判断するための正式な検定はありません。

IBM SPSS Modeler で使用されるモデル作成手法の数学的な基礎の説明は、「*IBM SPSS Modeler アルゴリズム・ガイド*」を参照してください。

関連手続き

一般化線型モデル手続きは、さまざまなモデルに適合する強力なツールです。

- 一般化推定方程式手続きは、一般化線型モデルを拡張して反復測定を可能にします。
- 線型混合モデル手続きを使用して、ランダム・コンポーネントまたは反復測定、またはその両方を持つスケール従属変数にモデルを適合させることができます。

推奨図書

一般化線型モデルの詳細については、次のテキストを参照してください。

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

第 24 章 自動車保険金請求へのガンマ回帰の適合 (一般化線型モデル)

一般化線型モデルは、正の範囲のデータを分析する場合にガンマ回帰を適合させるのに使用できます。例えば、他の場所[3]を表示および分析するデータ・セットは、自動車の損害請求に関係します。 逆リンク関数を使用して従属変数の平均を予測値の線型の組み合わせに関連付けて、ガンマ分布を示すように平均的な請求額をモデル化することができます。平均的な請求額を計算するために使用する変動する請求数を説明するために、*請求数* をスケーリングの重みとして指定します。

この例では *car-insurance_genlin.str* というストリームを使用し、*car_insurance_claims.sav* というデータ・ファイルを参照します。データ・ファイルは *Demos* フォルダーにあり、ストリーム・ファイルは *streams* サブフォルダーにあります。

ストリームの作成

1. *Demos* フォルダーの *car_insurance_claims.sav* を指し示す Statistics ファイル入力ノードを追加します。

図 326. 自動車保険金請求を予測するサンプル・ストリーム

2. ソース・ノードの「タイプ」タブで、「*claimamt*」フィールドの役割に「**対象**」を設定します。 その他のフィールドの役割は、すべて「**入力**」に設定します。
3. 「**値の読み取り**」をクリックし、データをインスタンス化します。

3. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

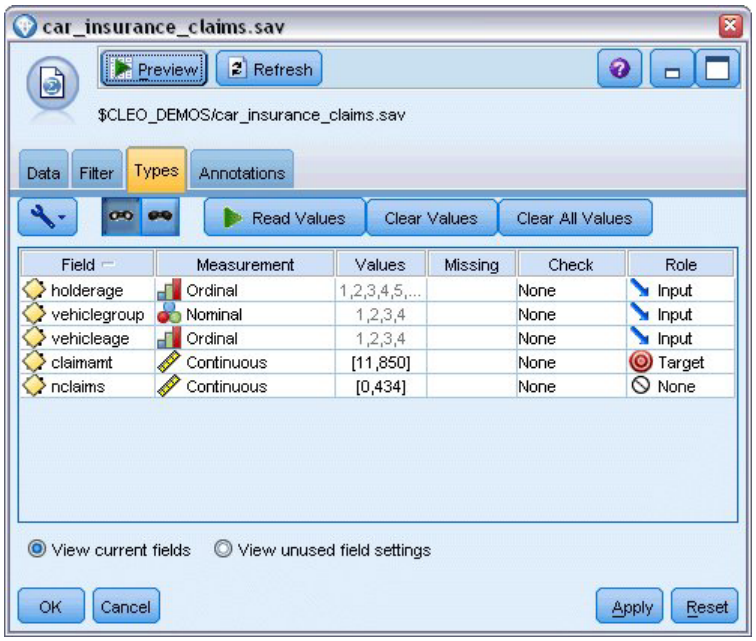


図 327. フィールドの役割の設定

4. Genlin ノードをソース・ノードに接続します。Genlin ノードで、「フィールド」タブをクリックします。
5. 「*nclaims*」をスケール重みフィールドとして選択します。

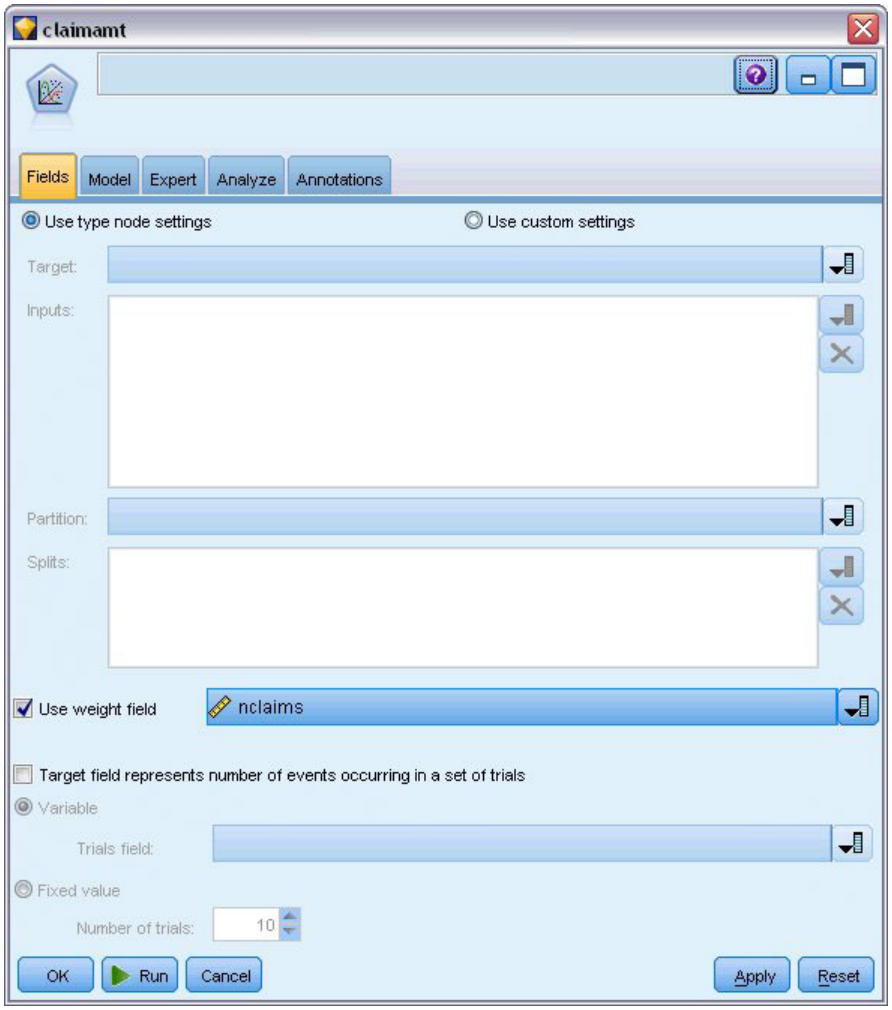


図 328. フィールド・オプションの選択

6. 「エキスパート」タブをクリックし、「**エキスパート**」を選択して、エキスパート・モデル作成オプションを有効にします。

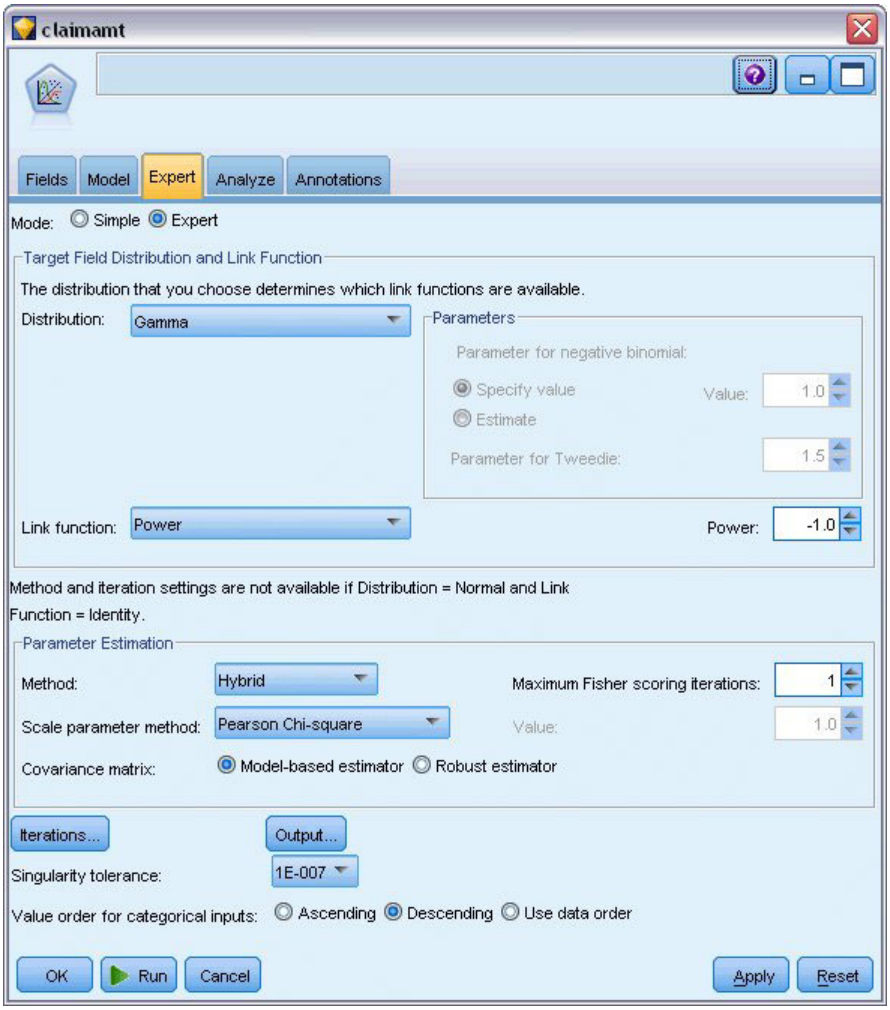


図 329. エキスパート・オプションの選択

7. 「**ガンマ**」を応答分布として選択します。
8. 「**べき乗**」をリンク関数として選択し、「-1.0」をべき乗関数の指数として入力します。これは逆リンクです。
9. 「**Pearson カイ 2 乗**」をスケール・パラメーターを推定するための方法として選択します。 これは McCullagh と Nelder によって使用される方法のため、ここでは結果を複製するためにそれに従います。
10. 「**降順**」を因子のカテゴリー順として選択します。 これは、それぞれの因子の第 1 のカテゴリーが参照カテゴリーになるということを示します。モデルでのこの選択の効果は、パラメーター推定値の解釈に表れます。
11. 「**実行**」をクリックし、モデル・ナゲットを作成します。これはストリーム領域および右上隅のモデル・パレットに追加されます。 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、「**編集**」または「**参照**」を選択して、「詳細」タブを選択します。

パラメーター推定値

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.003411	.000418	.002591	.004230	66.593	1	.000
[holderage=8]	.000920	.000416	.000105	.001735	4.898	1	.027
[holderage=7]	.000916	.000408	.000117	.001716	5.046	1	.025
[holderage=6]	.000969	.000405	.000176	.001763	5.740	1	.017
[holderage=5]	.001370	.000419	.000548	.002192	10.682	1	.001
[holderage=4]	.000462	.000411	-.000342	.001267	1.268	1	.260
[holderage=3]	.000350	.000412	-.000458	.001158	.720	1	.396
[holderage=2]	.000101	.000436	-.000754	.000956	.054	1	.816
[holderage=1]	.000000[a]	.	.	.	.	.	.
[vehiclegroup=4]	-.001421	.000181	-.001775	-.001067	61.883	1	.000
[vehiclegroup=3]	-.000614	.000170	-.000947	-.000281	13.039	1	.000
[vehiclegroup=2]	.000038	.000169	-.000293	.000368	.050	1	.823
[vehiclegroup=1]	.000000[a]	.	.	.	.	.	.
[vehicleage=4]	.004154	.000442	.003287	.005021	88.175	1	.000
[vehicleage=3]	.001651	.000227	.001207	.002096	53.013	1	.000
[vehicleage=2]	.000366	.000101	.000169	.000564	13.191	1	.000
[vehicleage=1]	.000000[a]	.	.	.	.	.	.
(Scale)	1.209[b]			.001	.0004	.000	.002

Dependent Variable: Average cost of claims
Model: (Intercept), holderage, vehiclegroup, vehicleage

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

図 330. パラメーター推定値

オムニバス検定およびモデル効果検定 (非表示) は、そのモデルが Null モデルよりも優れていて、それぞれの主効果の項がモデルに寄与するということを示しています。パラメーターの推定値の表は、McCullagh と Nelder が因子レベルおよびスケール・パラメーターで取得した同じ値を示しています。

要約

一般化線型モデルを使用して、ガンマ回帰を請求データに適合させました。このモデルではガンマ分布の標準リンク関数を使用しましたが、対数リンクでも妥当な結果を得られます。一般に、さまざまなリンク関数を持つモデルを直接比較するのは困難です。ただし、対数リンクは指数が 0 のべき乗リンクの特殊なケースであるため、対数リンクを持つモデルおよびべき乗リンクを持つモデルの逸脱を比較してどちらがより適合しているかを判断できます。

IBM SPSS Modeler で使用されるモデル作成手法の数学的な基礎の説明は、「*IBM SPSS Modeler アルゴリズム・ガイド*」を参照してください。

関連手続き

一般化線型モデル手続きは、さまざまなモデルに適合する強力なツールです。

- 一般化推定方程式手続きは、一般化線型モデルを拡張して反復測定を可能にします。
- 線型混合モデル手続きを使用して、ランダム・コンポーネントまたは反復測定、またはその両方を持つスケール従属変数にモデルを適合させることができます。

推奨図書

一般化線型モデルの詳細については、次のテキストを参照してください。

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

第 25 章 細胞サンプルの分類 (SVM)

Support Vector Machine (SVM) は広範なデータ・セットに特に適した分類および回帰の技術です。広範なデータ・セットとは多数の予測値が含まれたデータ・セットのことであり、バイオインフォマティクス (生化学および生物学データへの情報技術の適用) の分野におけるデータ・セットなどです。

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るとします。元のデータの分析では、良性と悪性のサンプルの間で、多数の特性が有意に異なることが分かりました。研究者は、他の患者から採取したサンプルでこれら細胞の特性の値を使用して、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する SVM モデルを開発しようとしています。

この例では、*Demos* フォルダーの *streams* サブフォルダー内にある *svm_cancer.str* というストリームを使用します。データ・ファイルは *cell_samples.data* です。詳細については、5 ページの『「Demos」フォルダー』を参照してください。

例は UCI マシン学習リポジトリーで公表されているデータ・セットに基づいています。このデータ・セットは何百ものヒト細胞サンプル・レコードで構成されており、それぞれに一連の細胞特性の値が含まれています。各レコードのフィールドは次のとおりです。

フィールド名	説明
ID	患者の ID
Clump	クランプの厚み
UnifSize	細胞の大きさの均一性
UnifShape	細胞の形の均一性
MargAdh	境界の接着性
SingEpiSize	単一の上皮細胞の大きさ
BareNuc	裸核
BlandChrom	ブランド・クロマチン
NormNucl	通常の核小体
Mit	有糸分裂
Class	良性または悪性

この例では、各レコードの予測値が比較的少ないデータ・セットを使用します。

ストリームの作成

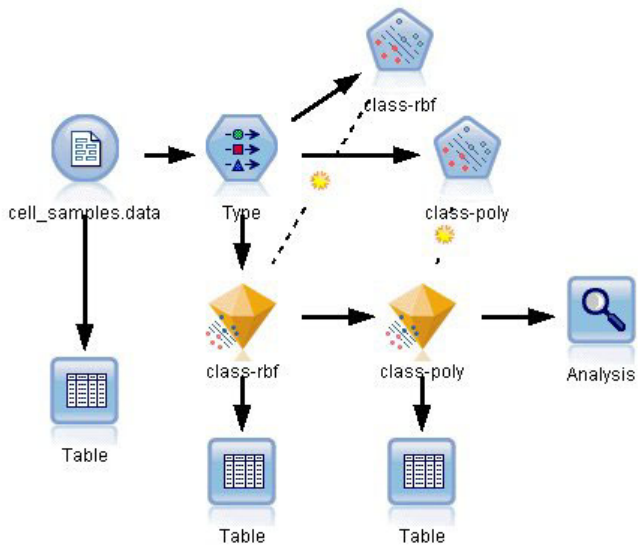
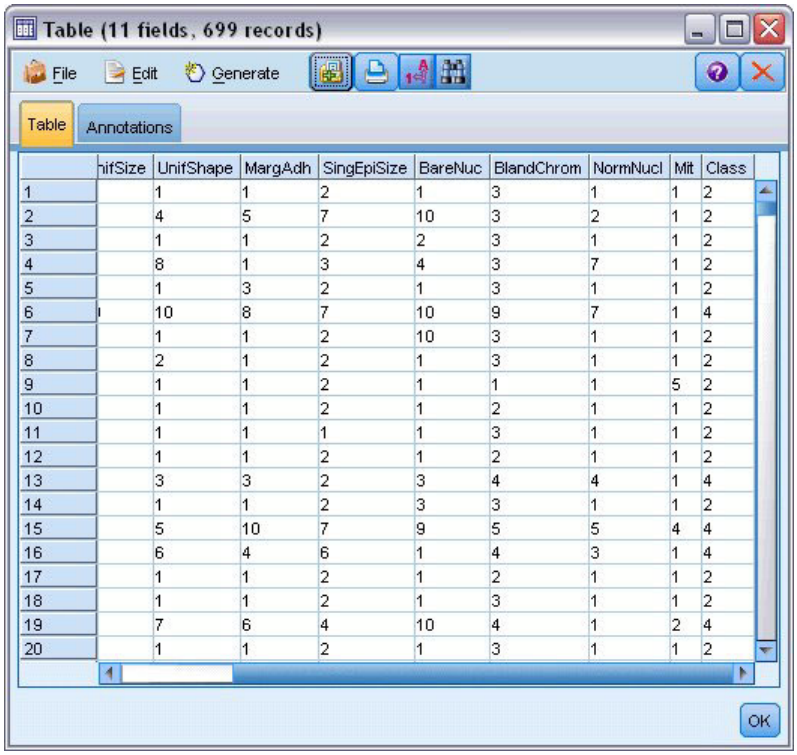


図 331. SVM モデル作成を表示するサンプル・ストリーム

1. 新規ストリームを作成し、IBM SPSS Modeler インストール環境の *Demos* フォルダーにある *cell_samples.data* を指し示す可変長ファイル入力ノードを追加します。

   ソース・ファイルのデータを見てみましょう。
2. テーブル・ノードをストリームに追加します。
3. テーブル・ノードを可変長ファイル・ノードに接続し、ストリームを実行します。



	hifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1		1	1	2	1	3	1	1	2
2		4	5	7	10	3	2	1	2
3		1	1	2	2	3	1	1	2
4		8	1	3	4	3	7	1	2
5		1	3	2	1	3	1	1	2
6		10	8	7	10	9	7	1	4
7		1	1	2	10	3	1	1	2
8		2	1	2	1	3	1	1	2
9		1	1	2	1	1	1	5	2
10		1	1	2	1	2	1	1	2
11		1	1	1	1	3	1	1	2
12		1	1	2	1	2	1	1	2
13		3	3	2	3	4	4	1	4
14		1	1	2	3	3	1	1	2
15		5	10	7	9	5	5	4	4
16		6	4	6	1	4	3	1	4
17		1	1	2	1	2	1	1	2
18		1	1	2	1	3	1	1	2
19		7	6	4	10	4	1	2	4
20		1	1	2	1	3	1	1	2

図 332. SVM のソース・データ

ID フィールドには、患者の ID が含まれています。各患者からの細胞サンプルの特性は、*Clump* から *Mit* のフィールドに含まれています。値は 1 から 10 まで等級分けがされており、1 が良性に一番近い値です。

Class フィールドには、サンプルが良性 (値 = 2) または悪性 (値 = 4) であるかに関する、別の医療処置で確認された診断が含まれています。

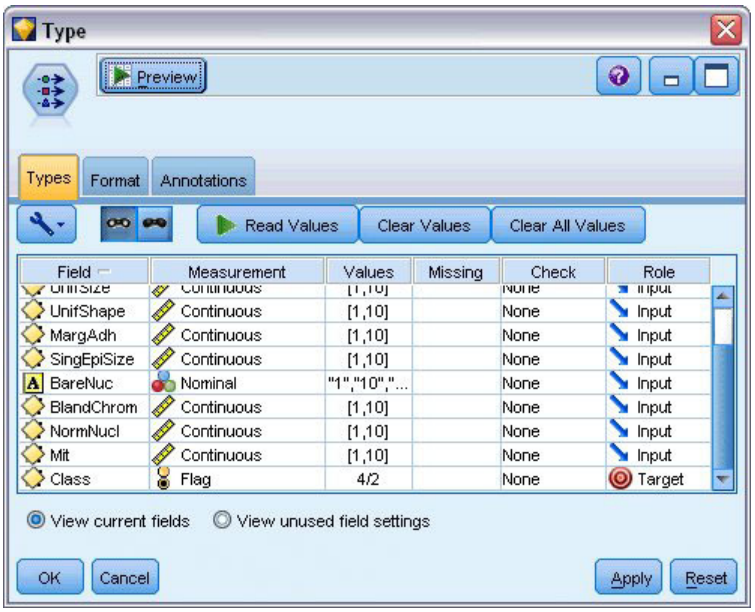


図 333. データ型ノードの設定

4. データ型ノードを追加し、それを可変長ファイル・ノードへ接続します。
5. データ型ノードを開きます。

   Class の値 (つまり、良性 (=2) または悪性 (=4)) を予測するモデルが必要です。このフィールドには 2 つしかない有効値の 1 つが入るため、測定の尺度を変更してこれを反映する必要があります。
6. *Class* フィールド (リストの最後のフィールド) の「**尺度**」列で、値「**連続型**」をクリックして 「**フラグ**」に変更します。
7. 「**値の読み込み**」をクリックします。
8. 「**役割**」列で、*ID* (患者の ID) の役割を「**なし**」に設定します。ID はモデルの予測値としても対象としても使用されないからです。
9. 対象の *Class* の役割を「**対象**」に設定し、その他のすべてのフィールド (予測値) の役割を「**入力**」のままにしておきます。
10. 「**OK**」をクリックします。

SVM ノードは、その処理を実行するためのカーネル関数の選択を提供します。特定のデータ・セットでどの関数のパフォーマンスが最適なのかを知る簡単な手段はないため、順番にさまざまな関数を選択して、結果を比較します。デフォルトの RBF (放射基底関数) から始めます。

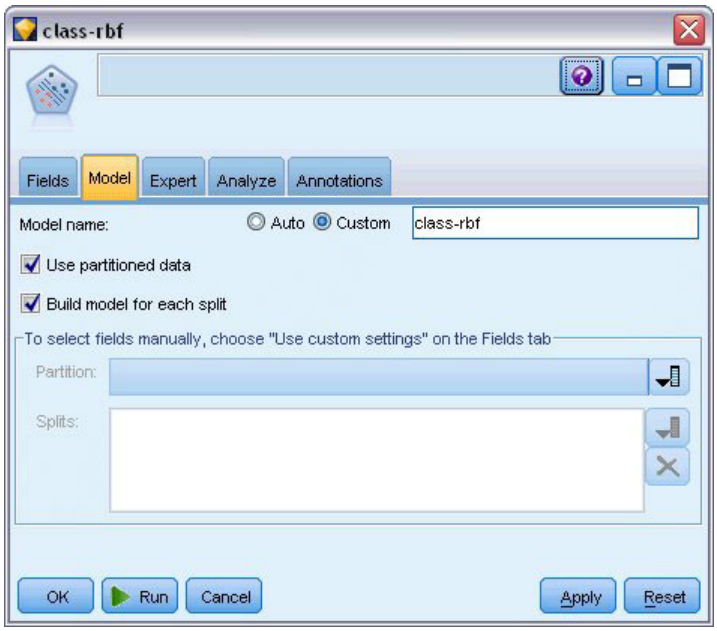


図 334. 「モデル」タブの設定

11. 「モデル生成」パレットから、SVM ノードをデータ型ノードに接続します。
12. SVM ノードを開きます。「**モデル**」タブで、「**モデル名**」の「**カスタム**」オプションをクリックし、隣接するテキスト・フィールドに *class-rbf* と入力します。

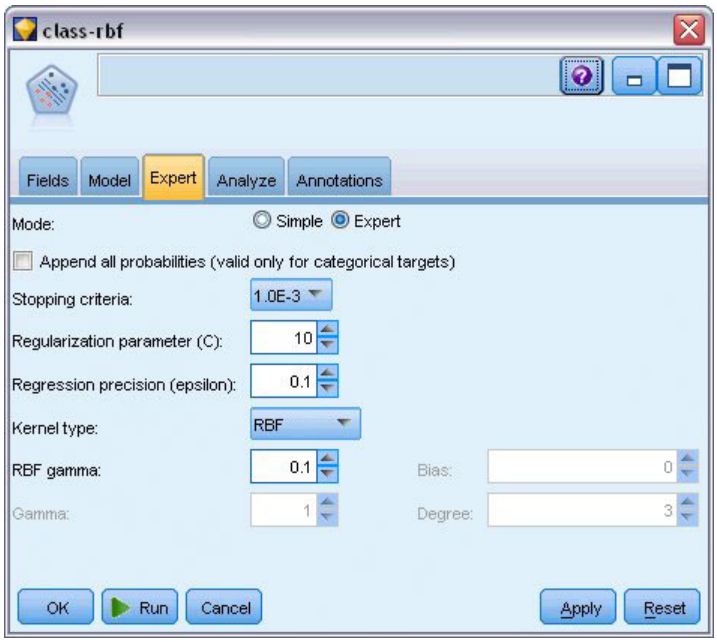


図 335. デフォルトの「エキスパート」タブの設定

13. 「**エキスパート**」タブで、見やすくするために「**モード**」を「**エキスパート**」に設定しますが、すべてのデフォルト・オプションはそのままにします。デフォルトでは「**カーネル・タイプ**」は「**RBF**」に設定されているので注意してください。簡易モードでは、すべてのオプションがグレーアウトされています。

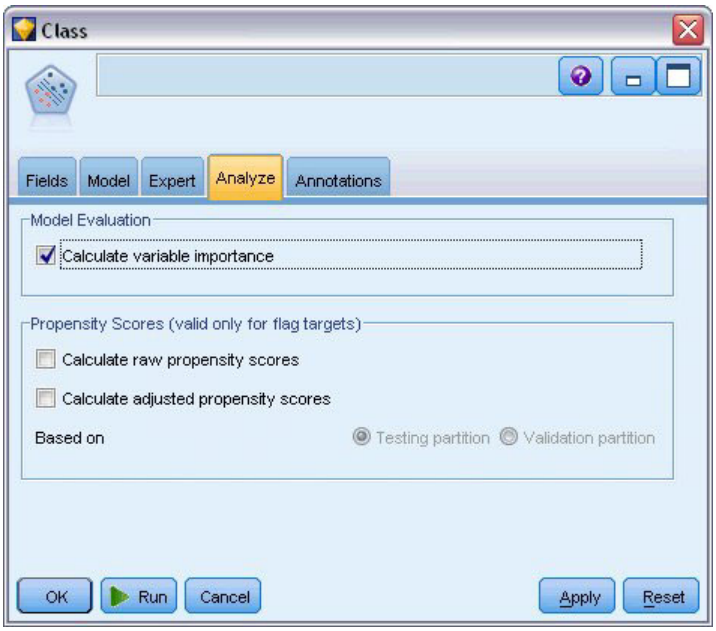


図 336. 「分析」タブの設定

14. 「**分析**」タブで、「**変数重要度を計算**」チェック・ボックスを選択します。
15. 「**実行**」をクリックします。ストリーム内、そして画面右上の「モデル」パレットにモデル・ナゲットが配置されます。
16. ストリーム内のモデル・ナゲットをダブルクリックします。

データの検証

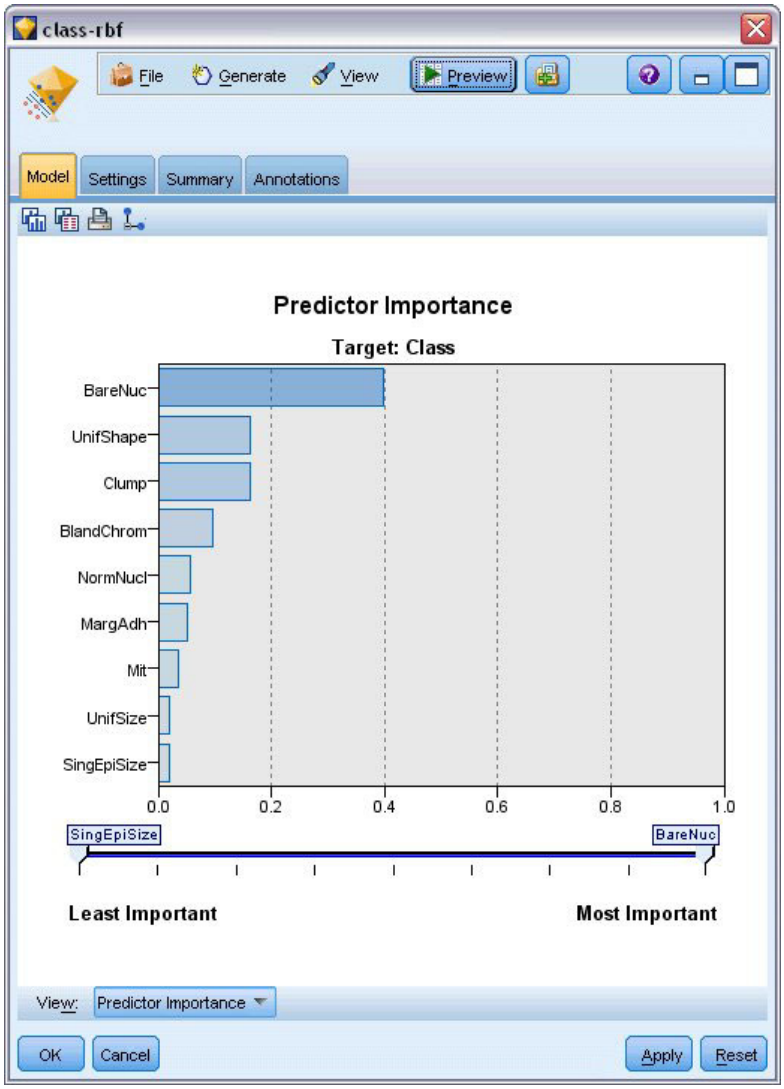
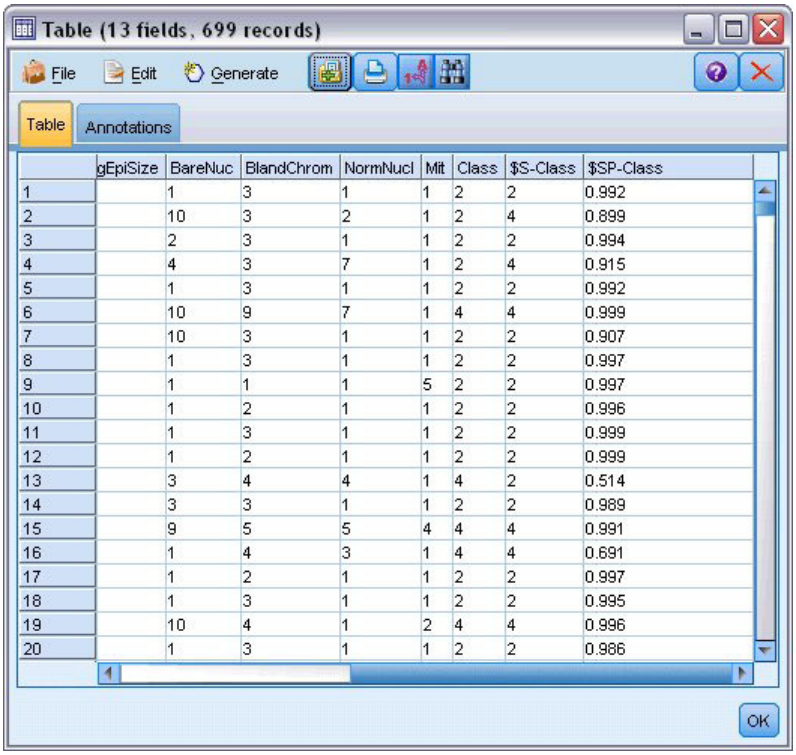


図 337. 予測重要度グラフ

「モデル」タブで、予測重要度グラフは予測でのさまざまなフィールドの相対効果を示しています。これにより、 *BareNuc* が最大効果を持っているのが一目瞭然で、また *UnifShape* および *Clump* もかなりの効果があることが分かります。

1. 「**OK**」をクリックします。
2. テーブル・ノードを *class-rbf* モデル・ナゲットに接続します。
3. テーブル・ノードを開いて、「**実行**」をクリックします。



	gEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class	$S-Class	$SP-Class
1		1	3	1	1	2	2	0.992
2		10	3	2	1	2	4	0.899
3		2	3	1	1	2	2	0.994
4		4	3	7	1	2	4	0.915
5		1	3	1	1	2	2	0.992
6		10	9	7	1	4	4	0.999
7		10	3	1	1	2	2	0.907
8		1	3	1	1	2	2	0.997
9		1	1	1	5	2	2	0.997
10		1	2	1	1	2	2	0.996
11		1	3	1	1	2	2	0.999
12		1	2	1	1	2	2	0.999
13		3	4	4	1	4	2	0.514
14		3	3	1	1	2	2	0.989
15		9	5	5	4	4	4	0.991
16		1	4	3	1	4	4	0.691
17		1	2	1	1	2	2	0.997
18		1	3	1	1	2	2	0.995
19		10	4	1	2	4	4	0.996
20		1	3	1	1	2	2	0.986

図 338. 予測値および信頼度の値に追加するフィールド

4. モデルでは、2 つの追加フィールドを作成しています。テーブル出力を右にスクロールして確認してください。

新規フィールド名	説明
$S-Class	モデルによって予測された *Class* の値。
$SP-Class	この予測の傾向スコア (この予測の尤度は真 (true) で、値は 0.0 から 1.0 です)。

テーブルを見るだけで、レコードの大半の傾向スコア (*$SP-Class* 列内) が十分高いことが分かります。

ただし、顕著な例外がいくつかあります。例えば、13 行目の患者 1041801 のレコードで、0.514 という値は受け入れがたいほど低くなっています。また、*Class* を *$S-Class* と比較すると、傾向スコアが比較的高い場所 (例えば、2 行目と 4 行目) でもこのモデルには間違った予測がたくさんあることが明らかです。

異なる関数タイプを選んで、これを改善できるかを試してみましょう。

異なる関数の試行

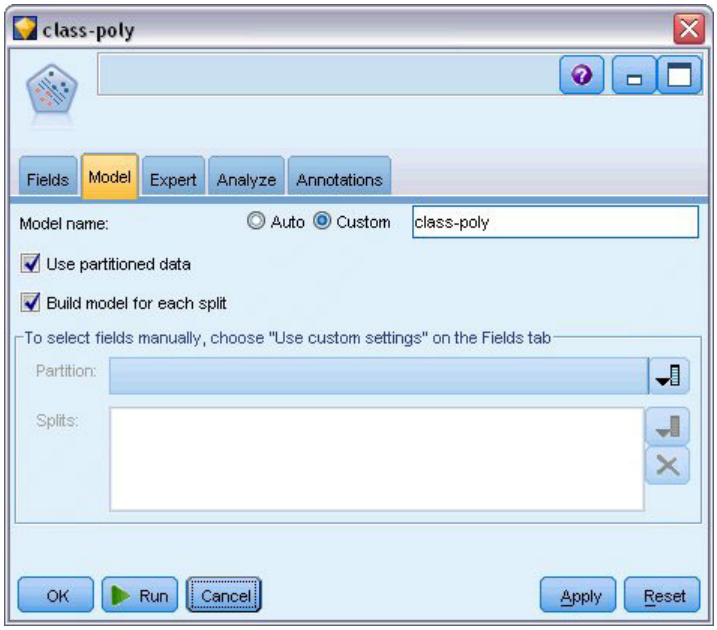


図 339. モデルの新しい名前の設定

1. テーブル出力ウィンドウを閉じます。
2. 2 番目の SVM モデル作成ノードをデータ型ノードに接続します。
3. 新しい SVM ノードを開きます。
4. 「**モデル**」タブで、「カスタム」を選択し、モデル名として *class-poly* と入力します。

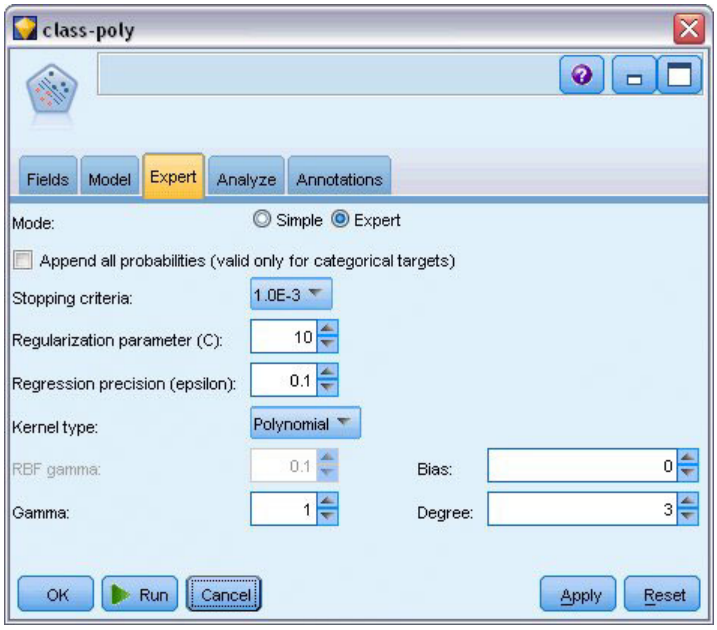
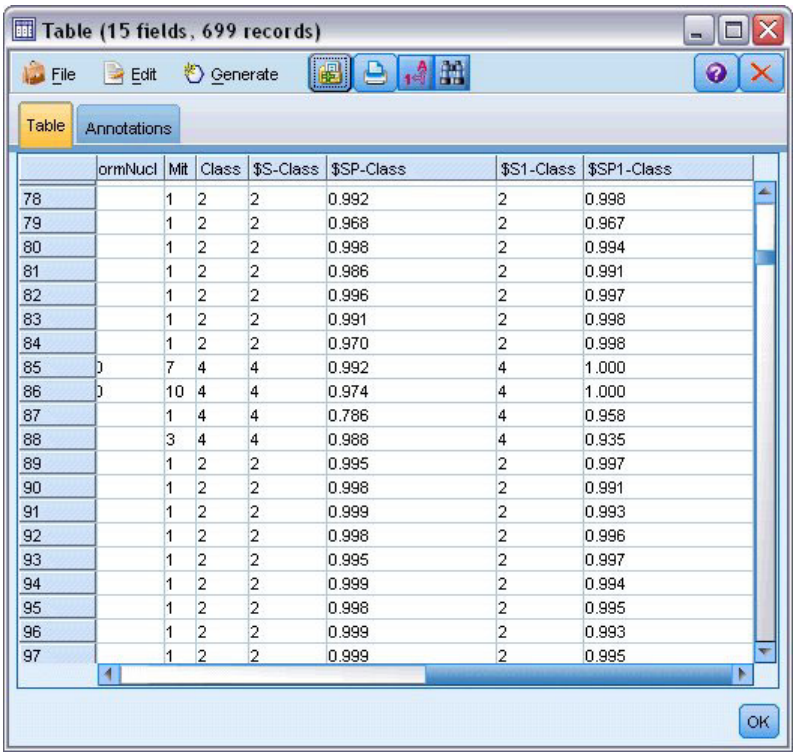


図 340. 多項式の「エキスパート」タブ設定

5. 「**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。

6. 「**カーネル・タイプ**」を「**多項式**」に設定し、「**実行**」をクリックします。*class-poly* モデル・ナゲットが、ストリーム内、そして画面右上の「モデル」パレットに追加されます。
7. *class-rbf* モデル・ナゲットを *class-poly* モデル・ナゲットに接続します (警告ダイアログで「**置換**」を選択します)。
8. テーブル・ノードを *class-poly* ナゲットに接続します。
9. テーブル・ノードを開いて、「**実行**」をクリックします。

結果の比較



	ormNucl	Mit	Class	$S-Class	$SP-Class	$S1-Class	$SP1-Class
78		1	2	2	0.992	2	0.998
79		1	2	2	0.968	2	0.967
80		1	2	2	0.998	2	0.994
81		1	2	2	0.986	2	0.991
82		1	2	2	0.996	2	0.997
83		1	2	2	0.991	2	0.998
84		1	2	2	0.970	2	0.998
85	0	7	4	4	0.992	4	1.000
86	0	10	4	4	0.974	4	1.000
87		1	4	4	0.786	4	0.958
88		3	4	4	0.988	4	0.935
89		1	2	2	0.995	2	0.997
90		1	2	2	0.998	2	0.991
91		1	2	2	0.999	2	0.993
92		1	2	2	0.998	2	0.996
93		1	2	2	0.995	2	0.997
94		1	2	2	0.999	2	0.994
95		1	2	2	0.998	2	0.995
96		1	2	2	0.999	2	0.993
97		1	2	2	0.999	2	0.995

図 341. 多項式関数に追加されたフィールド

1. テーブル出力を右にスクロールして、新しく追加されたフィールドを表示します。

   多項式関数タイプに生成されたフィールドは、*$S1-Class* および *$SP1-Class* という名前です。

   多項式の結果がはるかに改善されたように見えます。傾向スコアの多くが 0.995 以上になっており、これは非常に有望です。
2. モデルでの改善を確認するため、精度分析ノードを *class-poly* モデル・ナゲットへ接続します。

精度分析ノードを開いて、「**実行**」をクリックします。

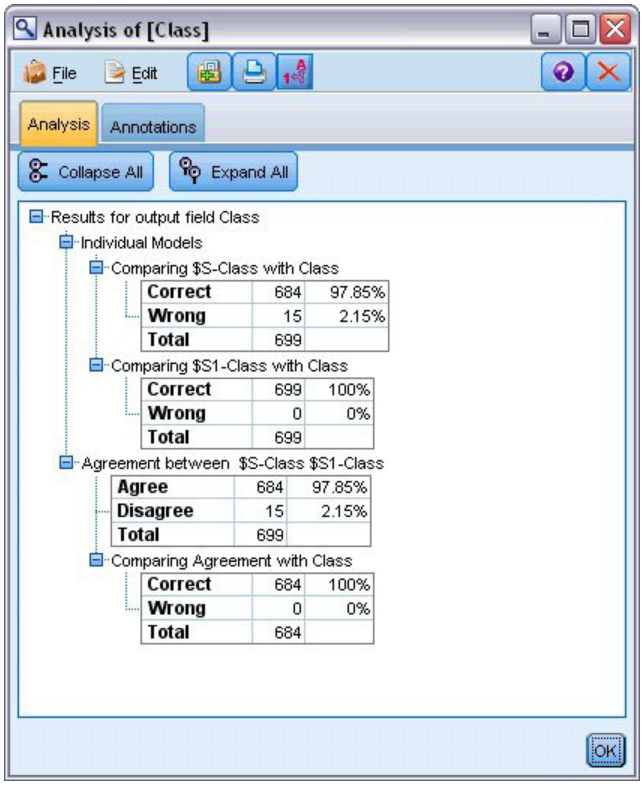


図 342. 精度分析ノード

精度分析ノードを使用したこの技術で、同じタイプの 2 つ以上のモデル・ナゲットを比較することができます。精度分析ノードからの出力で、RBF 関数がケースの 97.85% を正確に予測していることが分かり、これも非常に良好です。ただし、この出力で、多項式関数がどの単一ケースでも診断を正確に予測したことが分かります。実際は 100% の精度はめったに実現しませんが、そのモデルが特定の適用に対して容認できる精度かどうかを判断する上で、精度分析ノードを使用できます。

実際には、他のどの関数タイプも (Sigmoid および線型)、この特定のデータ・セットで多項式ほど良好には機能しません。ただし、別のデータ・セットでは結果が異なる可能性が十分あります。したがって、常にあらゆるオプションを試す価値はあります。

要約

SVM カーネル関数のさまざまなタイプを使用して、多くの属性から分類を予測しました。異なるカーネルがいかに同じデータ・セットに対してさまざまな結果を出すか、そしてモデルごとの改善をどのように測定できるかを学びました。

第 26 章 Cox 回帰を使用した顧客が解約するまでの時間のモデル作成

ある通信会社では、顧客離れを減らすための取り組みの一環として、すぐに他社のサービスに切り替える顧客に関連する要因を特定するために、「解約するまでの期間」のモデル作成に注目しています。このために、顧客のランダム・サンプリングが選択され、顧客としての期間、現在アクティブな顧客かどうか、およびその他のさまざまなフィールドがデータベースから取り出されます。

この例では、*telco_coxreg.str* というストリームを使用します。これは、*telco.sav* というデータ・ファイルを参照します。データ・ファイルは *Demos* フォルダーにあり、ストリーム・ファイルは *streams* サブフォルダーにあります。詳細については、5 ページの『「Demos」フォルダー』を参照してください。

適切なモデルの構築

1. *Demos* フォルダーの *telco.sav* を指し示す Statistics ファイル入力ノードを追加します。

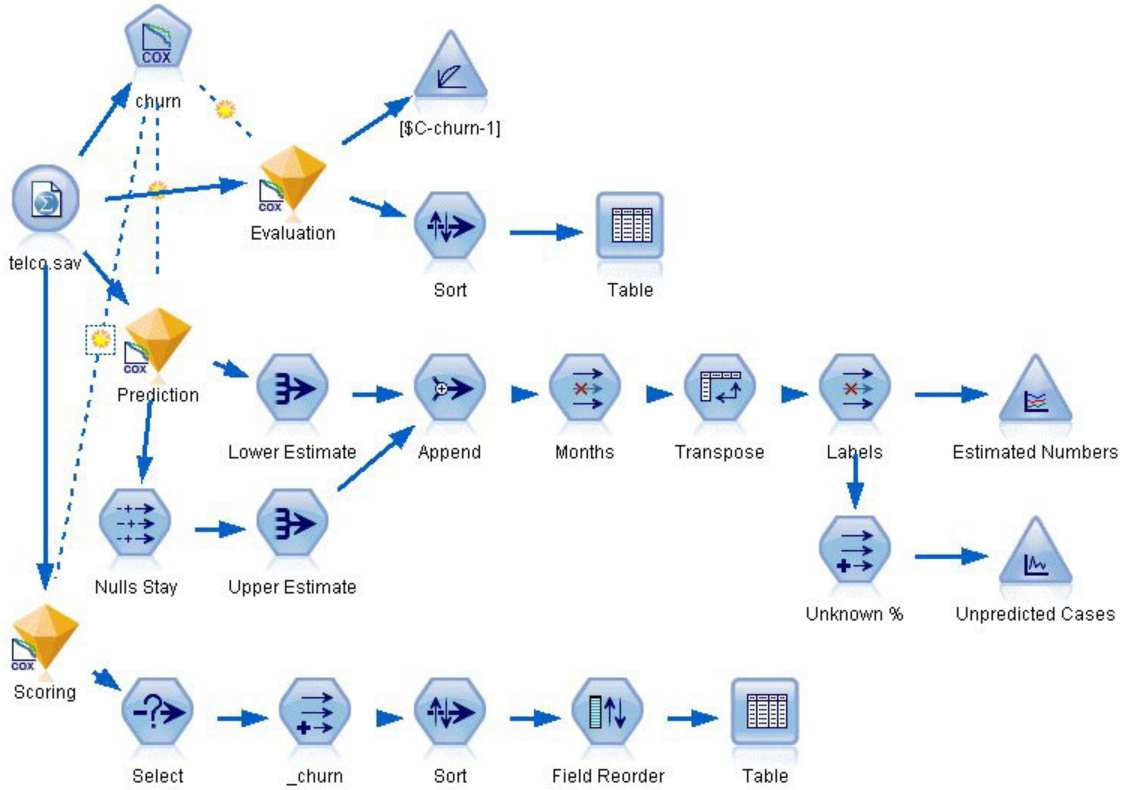


図 343. 解約までの期間を分析するためのサンプル・ストリーム

2. ソース・ノードの「フィルター」タブで、「*region*」、「*income*」、「*longten*」から「*wireten*」まで、および「*loglong*」から「*logwire*」までのフィールドを除外します。

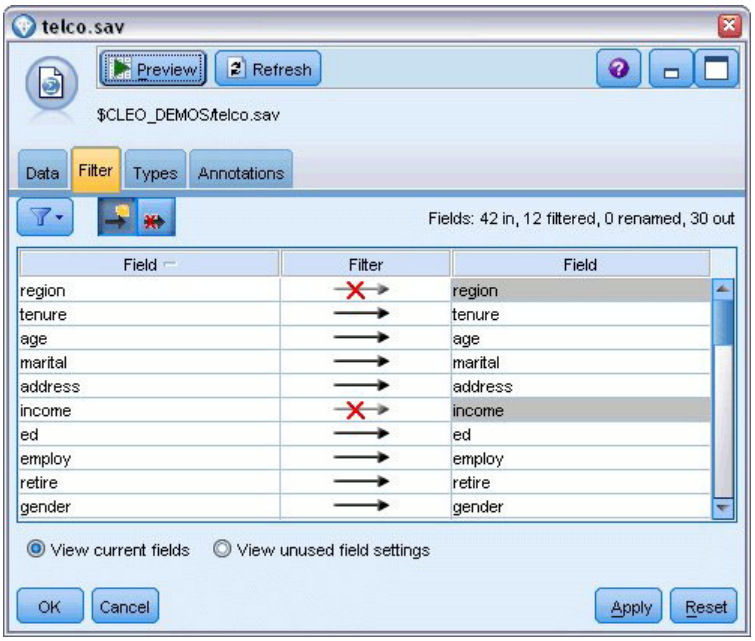


図 344. 不必要なフィールドのフィルタリング

(あるいは、除外する代わりに、「データ型」タブでこれらのフィールドの役割を「**なし**」に変更するか、モデル作成ノードで使用するフィールドを選択することもできます。)

3. ソース・ノードの「データ型」タブで、「*churn*」フィールドの役割を「**対象**」に設定し、その測定の尺度を「**フラグ**」に設定します。その他のフィールドの役割は、すべて「**入力**」に設定します。
4. 「**値の読み取り**」をクリックし、データをインスタンス化します。

図 345. フィールドの役割の設定

5. Cox ノードをソース・ノードに接続します。「**フィールド**」タブで、生存時間変数として「*tenure*」を選択します。

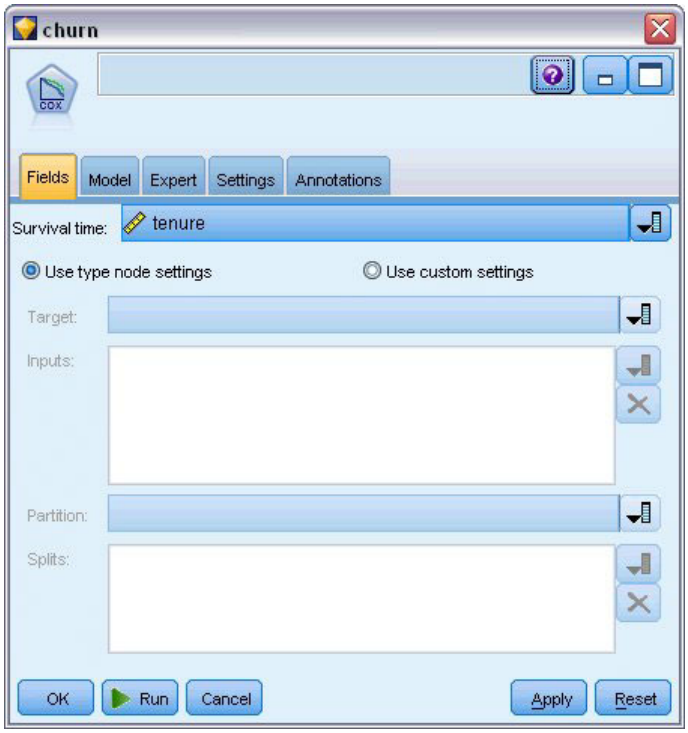


図 346. フィールド・オプションの選択

6. 「**モデル**」タブをクリックします。
7. 変数の選択方法として、「**ステップワイズ法**」を選択します。

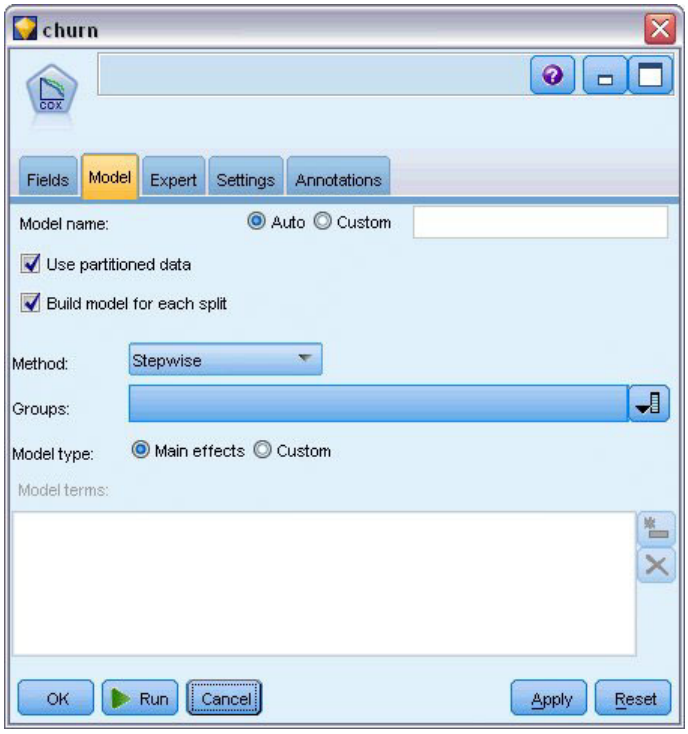


図 347. モデル・オプションの選択

8. 「**エキスパート**」タブをクリックし、「**エキスパート**」を選択して、エキスパート・モデル作成オプションを有効にします。
9. 「**出力**」をクリックします。

図 348. 詳細出力オプションの選択

10. 作成するプロットとして「**生存**」と「**ハザード**」を選択し、「**OK**」をクリックします。
11. 「**実行**」をクリックしてモデル・ナゲットを作成します。これは、ストリームと、右上の「モデル」パレットに追加されます。詳細を表示するには、ストリームのナゲットをダブルクリックします。最初に、「詳細出力」タブを確認します。

打ち切りケース

		N	Percent
Cases available in analysis	Event[a]	274	27.4%
	Censored	726	72.6%
	Total	1000	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		1000	100.0%

a. Dependent Variable: Months with service

図 349. ケース処理の要約

状態変数は、対象のケースにイベントが発生したかどうかを識別します。イベントが発生していない場合、ケースは「打ち切られた」と言われます。打ち切りケースは、回帰係数の計算には使用されませんが、ベースライン・ハザードの計算には使用されます。ケース処理の要約は、726 のケースが打ち切られたことを示しています。これらは、解約していない顧客です。

カテゴリー変数のコード化

		Frequency	(1)[b]	(2)	(3)	(4)
marital[a]	0=Unmarried	505	1			
	1=Married	495	0			
ed[a]	1=Did not complete high school	204	1	0	0	0
	2=High school degree	287	0	1	0	0
	3=Some college	209	0	0	1	0
	4=College degree	234	0	0	0	1
	5=Post-undergraduate degree	66	0	0	0	0
retire[a]	.00=No	953	1			
	1.00=Yes	47	0			
gender[a]	0=Male	483	1			
	1=Female	517	0			
tollfree[a]	0=No	526	1			
	1=Yes	474	0			
equip[a]	0=No	614	1			
	1=Yes	386	0			
callcard[a]	0=No	322	1			
	1=Yes	678	0			
wireless[a]	0=No	704	1			
	1=Yes	296	0			
multline[a]	0=No	525	1			
	1=Yes	475	0			
voice[a]	0=No	696	1			
	1=Yes	304	0			
pager[a]	0=No	739	1			
	1=Yes	261	0			
internet[a]	0=No	632	1			
	1=Yes	368	0			
callid[a]	0=No	519	1			
	1=Yes	481	0			
callwait[a]	0=No	515	1			
	1=Yes	485	0			
forward[a]	0=No	507	1			
	1=Yes	493	0			
confer[a]	0=No	498	1			
	1=Yes	502	0			
ebill[a]	0=No	629	1			
	1=Yes	371	0			
custcat[a]	1=Basic service	266	1	0	0	
	2=E-service	217	0	1	0	
	3=Plus service	281	0	0	1	
	4=Total service	236	0	0	0	

図 350. カテゴリー変数のコード化

カテゴリー変数のコード化は、カテゴリー共変量の回帰係数、特に二分変数を解釈するために参照するのに便利です。デフォルトでは、参照カテゴリーは「最後の」カテゴリーです。したがって、例えば、*既婚* の顧客はデータ・ファイル内の変数値は 1 ですが、回帰で使用するために 0 とコード化されます。

変数選択

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	3392.536	162.303	1	.000	133.828	1	.000	133.828	1	.000
2[b]	3087.314	249.392	2	.000	305.222	1	.000	439.050	2	.000
3[c]	3027.085	328.426	3	.000	60.229	1	.000	499.279	3	.000
4[d]	2990.790	347.197	4	.000	36.294	1	.000	535.574	4	.000
5[e]	2973.790	362.673	5	.000	17.000	1	.000	552.574	5	.000
6[f]	2958.796	376.140	6	.000	14.994	1	.000	567.568	6	.000
7[g]	2945.503	384.717	7	.000	13.293	1	.000	580.861	7	.000
8[h]	2936.993	417.341	8	.000	8.510	1	.004	589.371	8	.000
9[i]	2926.000	423.911	9	.000	10.994	1	.001	600.364	9	.000
10[j]	2917.551	428.078	10	.000	8.449	1	.004	608.813	10	.000
11[k]	2913.308	436.837	11	.000	4.243	1	.039	613.056	11	.000
12[l]	2908.078	440.158	12	.000	5.230	1	.022	618.286	12	.000

a. Variable(s) Entered at Step Number 1: callcard
b. Variable(s) Entered at Step Number 2: longmon
c. Variable(s) Entered at Step Number 3: equip
d. Variable(s) Entered at Step Number 4: employ
e. Variable(s) Entered at Step Number 5: multline
f. Variable(s) Entered at Step Number 6: voice
g. Variable(s) Entered at Step Number 7: address
h. Variable(s) Entered at Step Number 8: equipmon
i. Variable(s) Entered at Step Number 9: ebill
j. Variable(s) Entered at Step Number 10: callid
k. Variable(s) Entered at Step Number 11: internet
l. Variable(s) Entered at Step Number 12: reside
m. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 3526.364
n. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

図 351. オムニバス検定

モデル構築プロセスは、変数増加ステップワイズ法のアルゴリズムを使用します。オムニバス検定は、モデルがどれほどうまく実行されるかを測定します。前のステップからのカイ 2 乗の変化は、前のステップと現在のステップでのモデルの –2 対数尤度の差です。変数を追加するステップでは、変化の有意性が 0.05 より小さい場合、その投入は妥当です。変数を削除するステップでは、変化の有意性が 0.10 より大きい場合、その除外は妥当です。12 ステップの場合、12 個の変数がモデルに追加されます。

		B	SE	Wald	df	Sig.	Exp(B)
Step 12	address	-.035	.009	14.543	1	.000	.966
	employ	-.051	.010	25.767	1	.000	.950
	reside	-.103	.046	5.037	1	.025	.902
	equip	-1.948	.381	26.180	1	.000	.143
	callcard	.777	.151	26.451	1	.000	2.175
	longmon	-.233	.022	115.619	1	.000	.792
	equipmon	-.042	.011	15.377	1	.000	.959
	multline	.612	.145	17.854	1	.000	1.844
	voice	-.501	.157	10.197	1	.001	.606
	internet	-.362	.160	5.114	1	.024	.697
	callid	-.464	.148	9.790	1	.002	.629
	ebill	-.399	.156	6.557	1	.010	.671

図 352. 式内の変数 (ステップ 12 のみ)

最終モデルには、*address*、*employ*、*reside*、*equip*、*callcard*、*longmon*、*equipmon*、*multline*、*voice*、*internet*、*callid*、および *ebill* が含まれます。個々の予測値の効果を理解するために、Exp(B) を見ます。これは、予測値の単位変化に対して予測されるハザードの変化と解釈されます。

- *address* の Exp(B) 値は、同じ場所に住み続けている顧客は毎年解約ハザードが 100%–(100%×0.966)=3.4% ずつ減少することを意味します。同じ場所に 5 年間居住している顧客の解約ハザードは、100%–(100%×0.966^5)=15.88% 減少します。
- *callcard* の Exp(B) 値は、通話カード・サービスに加入していない顧客の解約ハザードは、サービスに加入している顧客より 2.175 倍大きいことを意味します。カテゴリー変数のコード化により、回帰では *No* = 1 であることを思い出してください。
- *internet* の Exp(B) 値は、インターネット・サービスに加入していない顧客の解約ハザードは、サービスに加入している顧客の 0.697 倍であることを意味します。これは、サービスを利用している顧客のほうがサービスを利用していない顧客よりも短期間で解約していることを示唆していることが、若干不安な点です。

		Score	df	Sig.
Step 12	age	.122	1	.726
	marital	.648	1	.421
	income	1.476	1	.224
	ed	6.328	4	.176
	ed(1)	.007	1	.934
	ed(2)	.203	1	.652
	ed(3)	.835	1	.361
	ed(4)	5.773	1	.016
	retire	.013	1	.908
	gender	.214	1	.644
	tollfree	3.243	1	.072
	wireless	.668	1	.414
	tollmon	.000	1	.987
	cardmon	3.163	1	.075
	wiremon	1.084	1	.298
	pager	1.808	1	.179
	callwait	.266	1	.606
	forward	2.201	1	.138
	confer	2.568	1	.109
	custcat	.864	3	.834
	custcat(1)	.466	1	.495
	custcat(2)	.450	1	.502
	custcat(3)	.019	1	.889

図 353. モデル内にない変数 (ステップ 12 のみ)

モデルからはずされた変数のスコア統計量はどれも、有意水準が 0.05 を超えています。ただし、*tollfree* および *cardmon* の有意水準は 0.05 未満ではありませんが、かなり近いと言えます。 これらについて、さらに調べていくと面白いでしょう。

共変量の平均値

	Mean
age	41.684
marital	.505
address	11.551
income	77.535
ed(1)	.204
ed(2)	.287
ed(3)	.209
ed(4)	.234
employ	10.987
retire	.953
gender	.483
reside	2.331
tollfree	.526
equip	.614
callcard	.322
wireless	.704
longmon	11.723
tollmon	13.274
equipmon	14.220
cardmon	13.781
wiremon	11.584
multline	.525
voice	.696
pager	.739
internet	.632
callid	.519
callwait	.515
forward	.507
confer	.498
ebill	.629
custcat(1)	.266
custcat(2)	.217
custcat(3)	.281

図 354. 共変量の平均値

このテーブルには、各予測変数の平均値が表示されます。このテーブルは、平均値に対して作成された生存プロットを見るときに、参照として役立ちます。ただし、カテゴリー予測値の指標変数の平均を見るときに、「平均」の顧客は実際には存在しません。すべてスケール予測値を使用しても、共変量の値がすべて平均に近い顧客を見つけることはほとんどないでしょう。特定のケースの生存曲線を見たい場合は、「プロット」ダイアログ・ボックスで、生存曲線をプロットする共変量の値を変更します。特定のケースの生存曲線を見たい場合は、「詳細出力」ダイアログ・ボックスの「プロット」グループで、生存曲線をプロットする共変量の値を変更します。

生存曲線

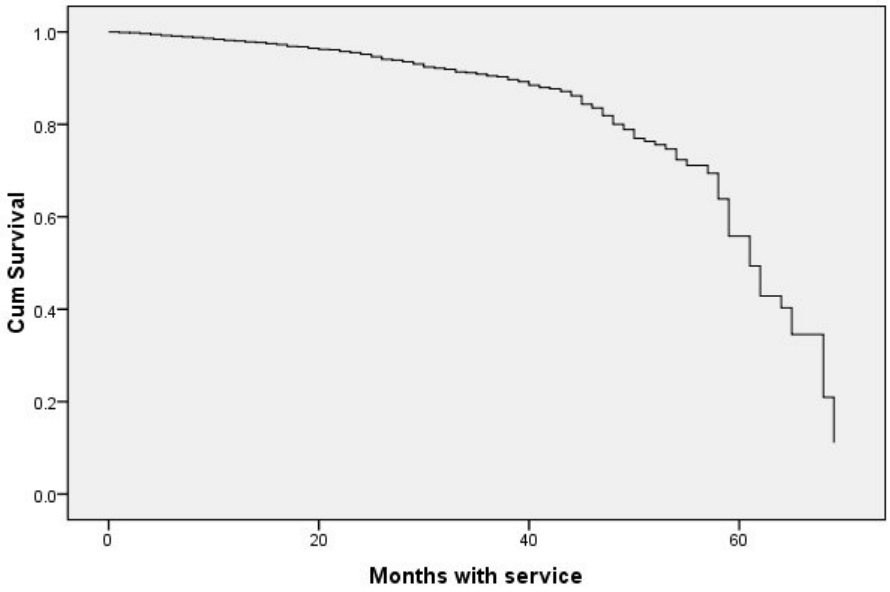


図 355. 「平均」の顧客の生存曲線

基本的な生存曲線は、モデルが予測した「平均」の顧客が解約するまでの時間を視覚的に表示します。横軸はイベントまでの時間を表します。縦軸は生存の確率を表します。したがって、生存曲線のどのポイントも、「平均」の顧客がその時間を経過しても顧客として残っている確率を示します。55 カ月を過ぎると、生存曲線はあまり滑らかではなくなります。それほどの長期にわたって存続する会社の顧客はほとんどいないため、使用できる情報が少なく、そのため曲線が荒くなっています。

ハザード曲線

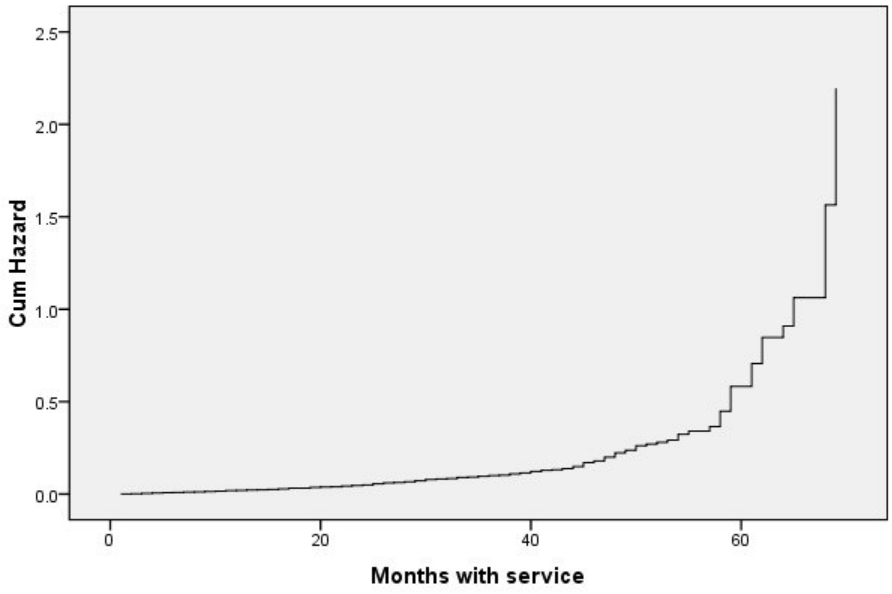


図 356. 「平均」の顧客のハザード曲線

基本的なハザード曲線は、モデルが予測した、「平均」の顧客が解約する累積的な可能性を視覚的に表示します。横軸はイベントまでの時間を表します。縦軸は、生存確率の負の対数と等しい累積ハザードを示します。55 カ月を過ぎると、生存曲線の場合と同じ理由により、ハザード曲線はあまり滑らかではなくなります。

評価

ステップワイズ選択法により、モデルに「統計的に有意な」予測値だけが含まれることは保証されますが、そのモデルが実際に対象を予測するのに適していることを保証するものではありません。そのためには、スコアリングされたレコードを分析する必要があります。

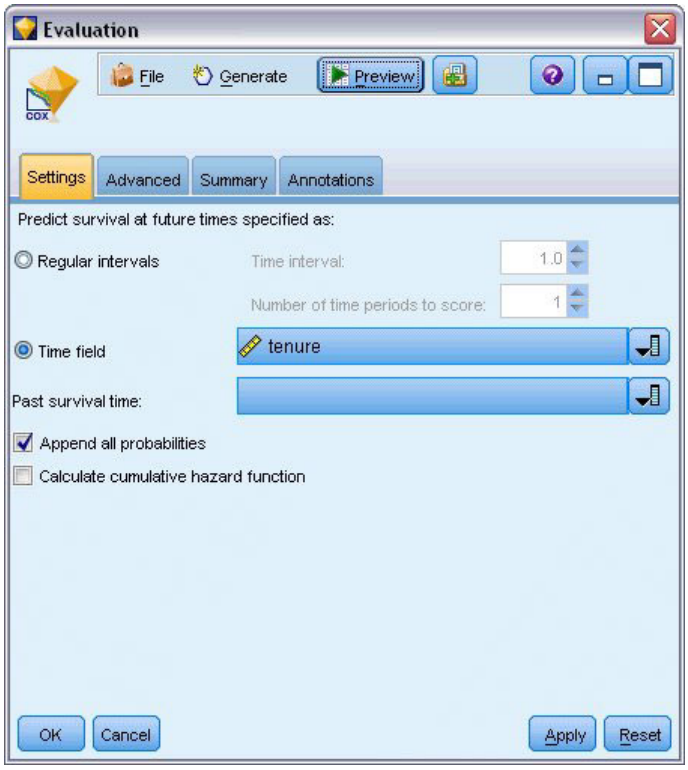


図 357. Cox ナゲット: 「設定」タブ

1. モデル・ナゲットを領域に配置し、それをソース・ノードに接続します。ナゲットを開いて、「設定」タブをクリックします。
2. **「時間フィールド」**を選択し、「*tenure*」を指定します。各レコードは、その tenure の長さでスコアリングされます。
3. **「すべての確率を追加」**を選択します。

   これにより、顧客が解約するかどうかのカットオフとして 0.5 を使用してスコアが作成されます。顧客の解約傾向が 0.5 より大きい場合、解約者としてスコアリングされます。この数字には何の魔力もなく、異なるカットオフを使えば、もっと理想的な結果が得られる可能性もあります。カットオフの選択について考慮する 1 つの方法は、評価ノードを使用することです。

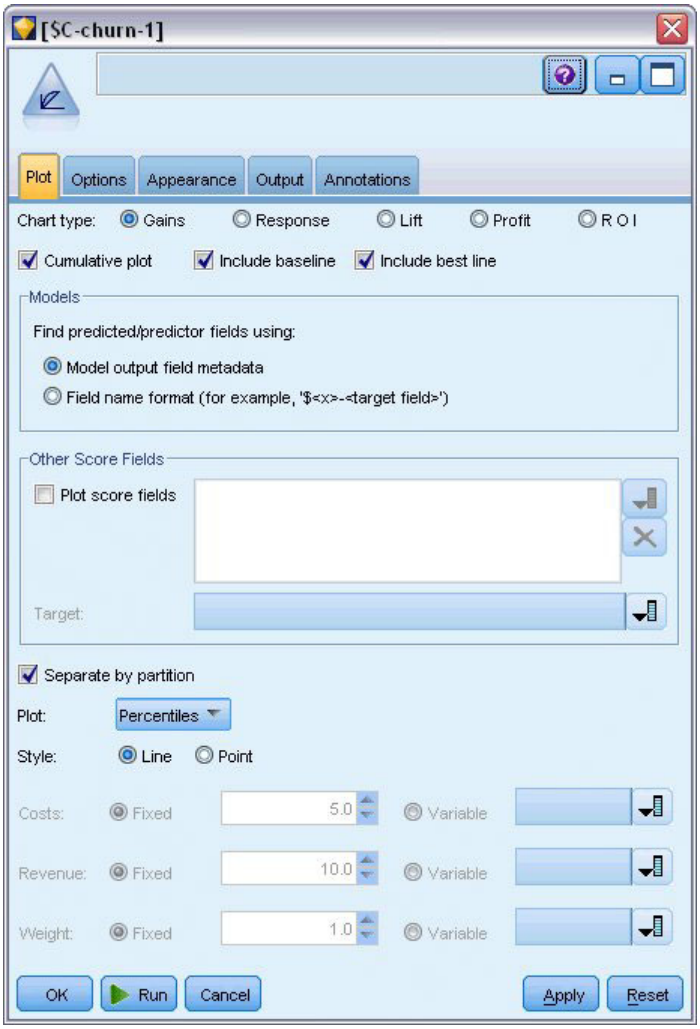


図 358. 評価ノード: 「プロット」タブ

4. 評価ノードをモデル・ナゲットに接続します。「プロット」タブで、「**ベストラインを含める**」を選択します。
5. 「**オプション**」タブをクリックします。

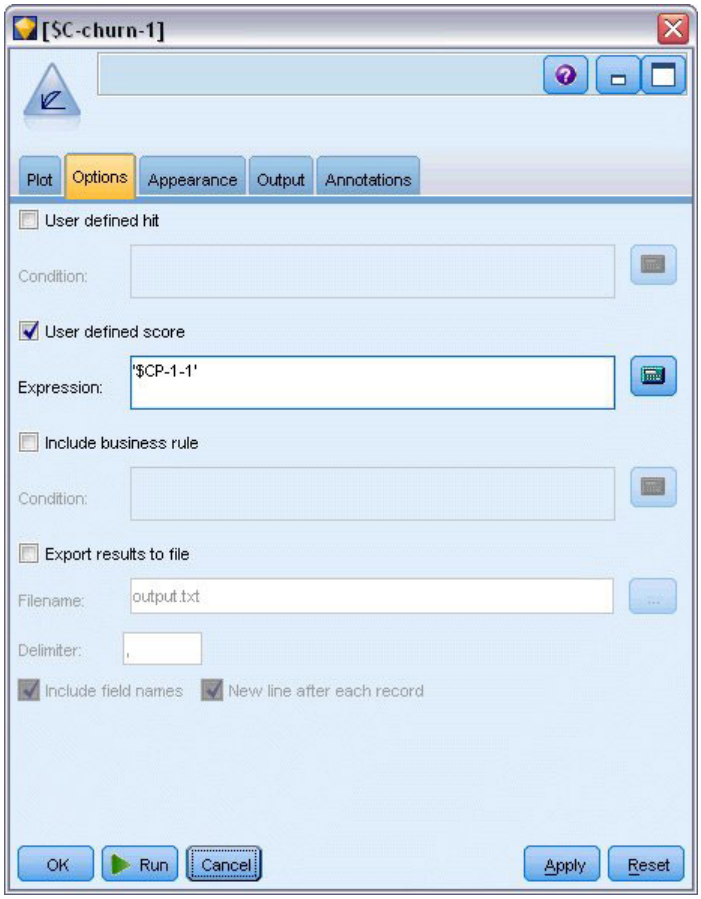


図 359. 評価ノード: 「オプション」タブ

6. 「**ユーザー定義のスコア**」を選択し、式として `'$CP-1-1'` を入力します。これがモデルによって生成されたフィールドで、解約の傾向と対応します。
7. 「**実行**」をクリックします。

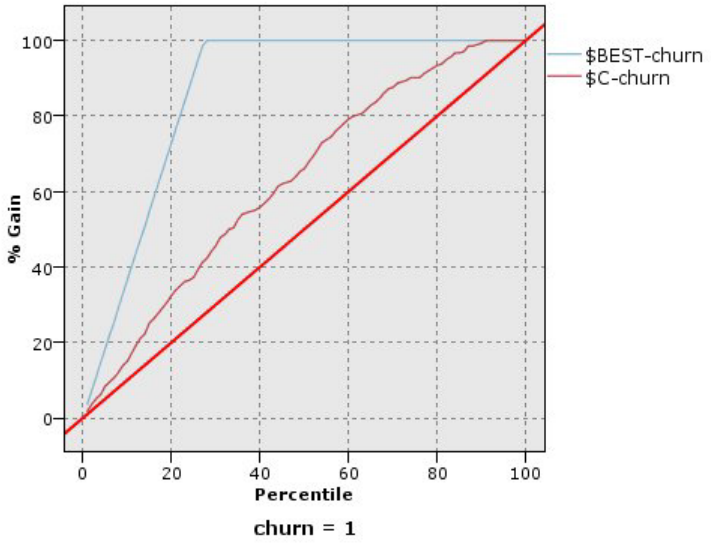


図 360. ゲイン・グラフ

累積ゲイン・グラフは、ケースの合計数のパーセントを対象にすることで、特定のカテゴリー「ゲイン」に含まれるケースの総数のパーセントを示します。例えば、曲線上の 1 点 (10%、15%) は、このモデルを使用してデータ・セットをスコアリングし、解約の予測傾向で全ケースをソートした場合、上位 10% には実際にカテゴリー *1* (解約者) と見なされる全ケースの約 15% が含まれることを意味します。同様に、上位 60% には解約者の約 79.2% が含まれます。スコアリングされたデータ・セットの 100% を選択した場合、データ・セット内のすべての解約者を取得できます。

対角線は「ベースライン」曲線です。スコアリングされたデータ・セットからランダムに 20% のレコードを選択する場合、実際にカテゴリー *1* と見なされる全レコードの約 20% が「ゲイン」になると予測できます。曲線がベースラインより上にあるほど、ゲインが大きくなります。「最適な」ラインは、非解約者よりも解約者により高い解約の傾向スコアを割り当てる「完璧な」モデルの曲線を示します。累積ゲイン・グラフを使用すると、希望するゲインに対応するパーセントを選択し、そのパーセントを適切なカットオフ値にマッピングすることで、分類のカットオフを選択しやすくなります。

「希望する」ゲインが何で構成されるかは、タイプ I およびタイプ II の誤差にかかるコストに依存します。つまり、解約者を非解約者として分類したことによるコストが何か (タイプ I) と、非解約者を解約者として分類したことによるコストが何か (タイプ II) です。顧客の保持が最も重要な関心事なので、タイプ I の誤差を小さくする必要があります。累積ゲイン・グラフでは、これは、予測傾向が *1* の上位 60% に入る顧客に対するカスタマー・ケアを向上させることに対応します。これによって、解約の可能性のある顧客の 79.2% を獲得できますが、それにかかる時間とリソースは、新規顧客の獲得に費やすこともできます。現在の顧客ベースを維持するためのコストの抑制が優先される場合は、タイプ II の誤差を小さくします。グラフでは、上位 20% に対するカスタマー・ケアを向上させることに対応します。これにより解約者の 32.5% を獲得できます。通常はどちらも重要な考慮事項です。そのため、重要度と特異性の最適な組み合わせで顧客を分類できるような決定ルールを選択する必要があります。

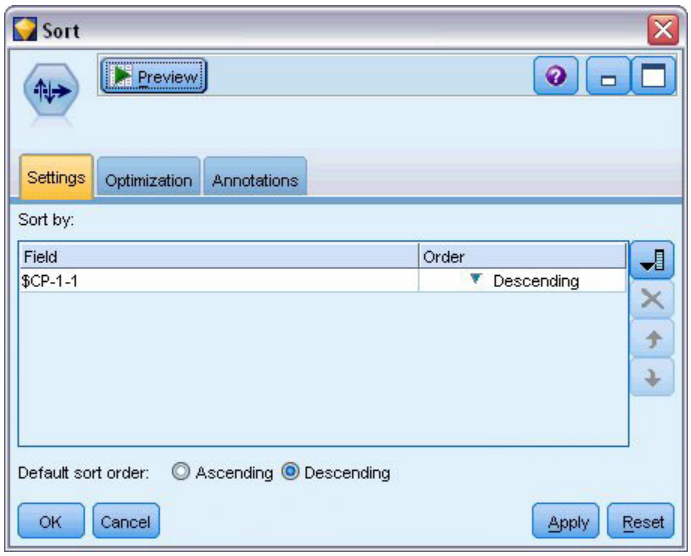


図 361. ソート・ノード: 「設定」タブ

8. 例えば、希望するゲインを 45.6% に決定したとします。これは、レコードの上位 30% に該当します。適切な分類カットオフを見つけるには、ソート・ノードをモデル・ナゲットに接続します。
9. 「設定」タブで、*$CP-1-1* による降順のソートを選択し、「**OK**」をクリックします。

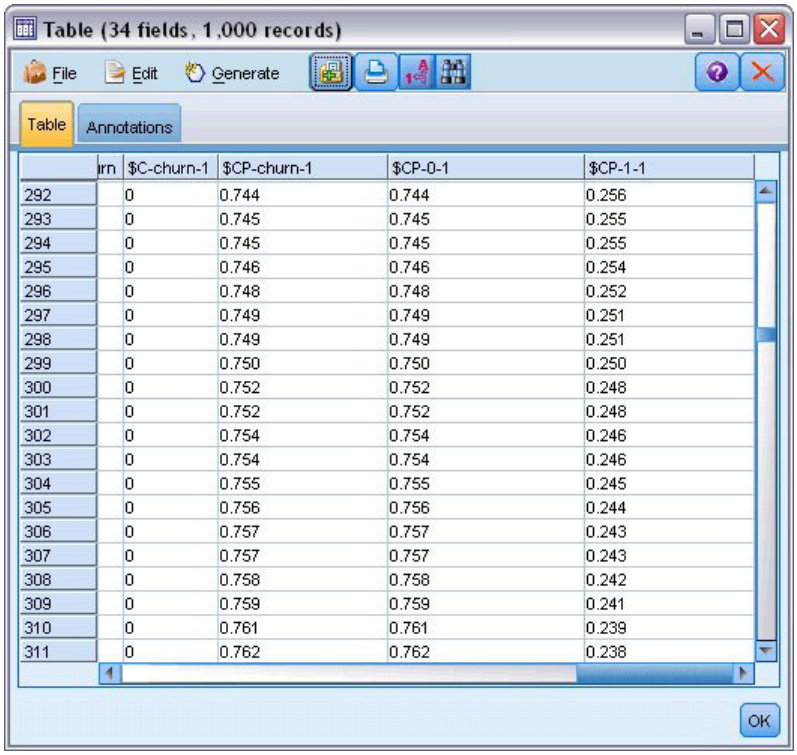


図 362. テーブル

10. テーブル・ノードをソート・ノードに接続します。
11. テーブル・ノードを開いて、「**実行**」をクリックします。

出力を下へスクロールすると、*$CP-1-1* の値が 300 番目のレコードでは 0.248 であることがわかります。0.248 を分類カットオフとして使用した結果は、解約者としてスコアリングされた顧客の約 30% になり、実際の全解約者の約 45% を獲得できます。

予測固定客数の追跡

満足するモデルを得られたら、データ・セット内で、今後 2 年間にわたって維持できると予測される顧客数を追跡します。Null 値の顧客は、合計保有期間 (将来の時間 + *tenure*) がモデルの学習に使用されたデータの生存時間の範囲外にあり、興味深い難題を提示しています。これに対処するための 1 つの方法は、Null 値を解約済みと仮定する予測値と、固定客化したと仮定する予測値の、2 つの予測値セットを作成することです。この方法によって、固定客の予測数に上限と下限を設定できます。

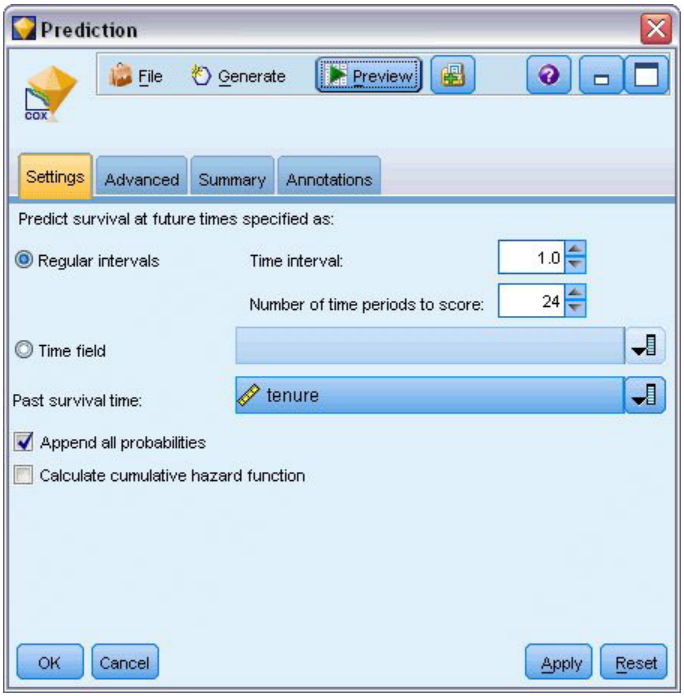


図 363. Cox ナゲット: 「設定」タブ

1. 「モデル」パレットでモデル・ナゲットをダブルクリック (またはストリーム領域でナゲットをコピーして貼り付け) し、新しいナゲットをソース・ノードに接続します。
2. ナゲットを「設定」タブに開きます。
3. 「**一定の間隔**」が選択されていることを確認してから、時間間隔として「1.0」を、さらにスコアリングする期間数として「24」を指定します。これにより、各レコードが以降 24 カ月間、毎月スコアリングされるように指定されます。
4. 過去の生存時間を指定するためのフィールドとして「*tenure*」を選択します。スコアリング・アルゴリズムは、各顧客の会社の顧客としての時間の長さを考慮します。
5. 「**すべての確率を追加**」を選択します。

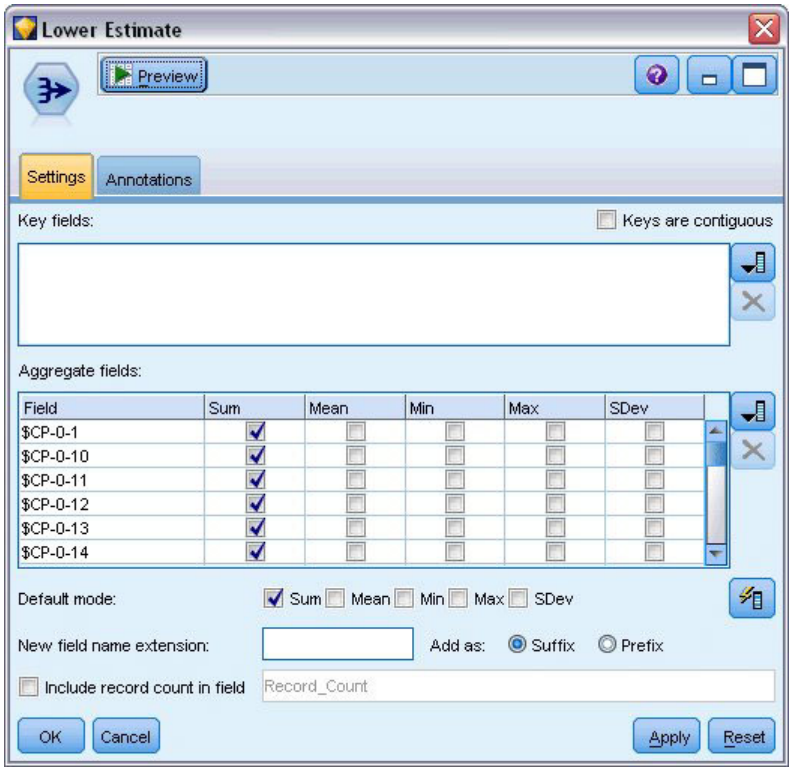


図 364. 集計ノード: 「設定」タブ

6. 集計ノードをモデル・ナゲットに接続します。「設定」タブで、デフォルト・モードとしての「**平均**」を選択解除します。
7. 集計するフィールドとして、「*$CP-0-1*」から「*$CP-0-24*」まで、形式が *$CP-0-n* のフィールドを選択します。これは、「フィールドの選択」ダイアログ・ボックスで、フィールドを名前 (つまり、アルファベット順) でソートしていれば、最も簡単に行えます。
8. 「**フィールドにレコード数を含める**」を選択解除します。
9. 「**OK**」をクリックします。このノードは、「下限」の予測値を作成します。

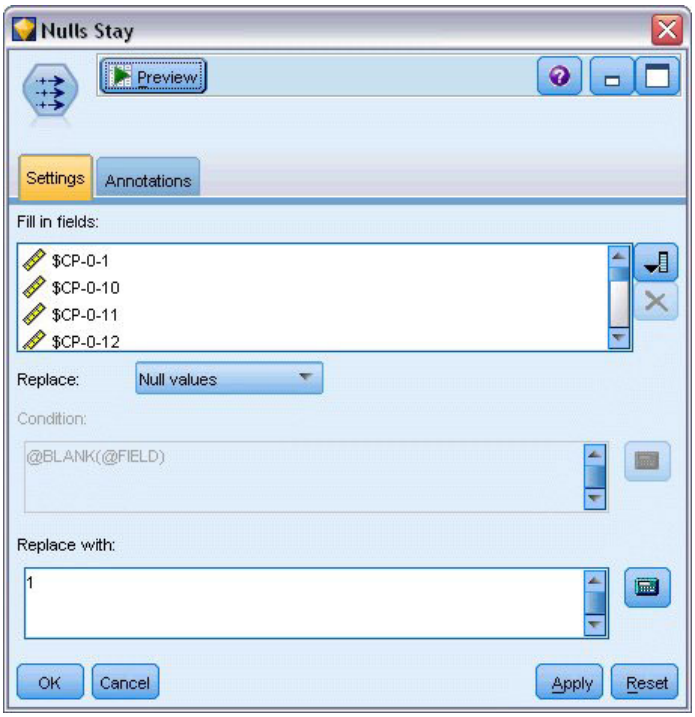


図 365. 置換ノード: 「設定」タブ

10. 先に集計ノードを接続した Cox 回帰ナゲットに置換ノードを接続します。「設定」タブで、置換するフィールドとして、「*$CP-0-1*」から「*$CP-0-24*」まで、形式が *$CP-0-n* のフィールドを選択します。これは、「フィールドの選択」ダイアログ・ボックスで、フィールドを名前 (つまり、アルファベット順) でソートしていれば、最も簡単に行えます。
11. 「**Null 値**」を値 1 で置換することを選択します。
12. 「**OK**」をクリックします。

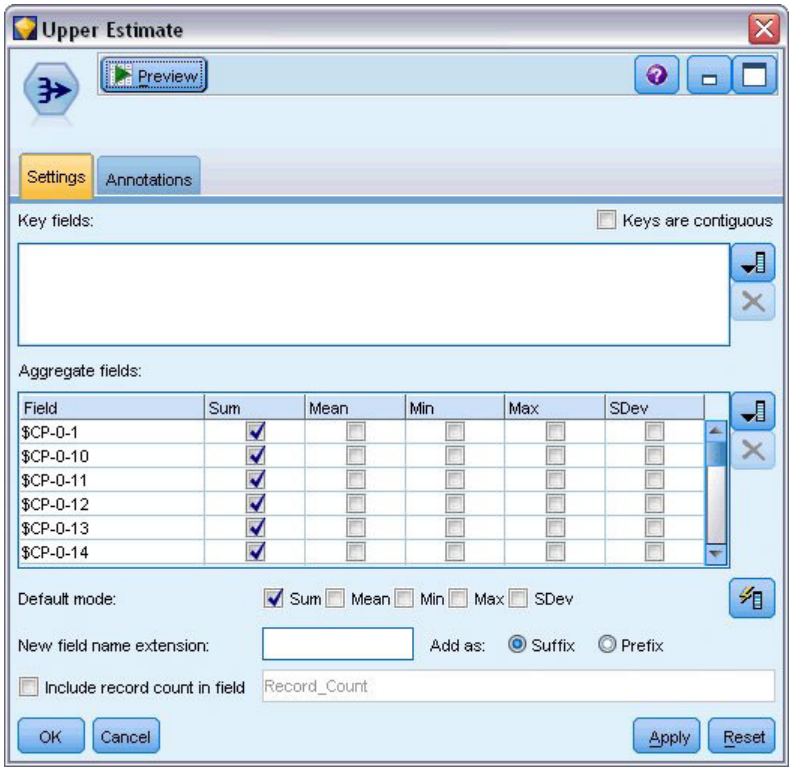


図 366. 集計ノード: 「設定」タブ

13. 集計ノードを置換ノードに接続します。「設定」タブで、デフォルト・モードとしての「**平均**」を選択解除します。
14. 集計するフィールドとして、「*$CP-0-1*」から「*$CP-0-24*」まで、形式が *$CP-0-n* のフィールドを選択します。これは、「フィールドの選択」ダイアログ・ボックスで、フィールドを名前 (つまり、アルファベット順) でソートしていれば、最も簡単に行えます。
15. 「**フィールドにレコード数を含める**」を選択解除します。
16. 「**OK**」をクリックします。このノードは、「上限」の予測値を作成します。

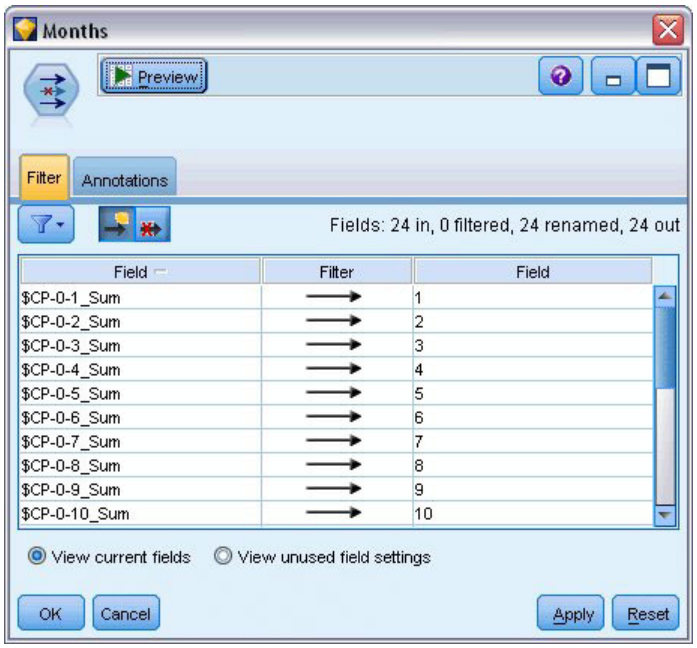


図 367. フィルター・ノード: 「設定」タブ

17. 1 つの追加ノードを 2 つの集計ノードに接続してから、フィルター・ノードを追加ノードに接続します。
18. フィルター・ノードの「設定」タブで、フィールドの名前を *1* から *24* までに変更します。入れ替えノードを使用すると、これらのフィールド名は、下流のグラフで *x* 軸の値になります。

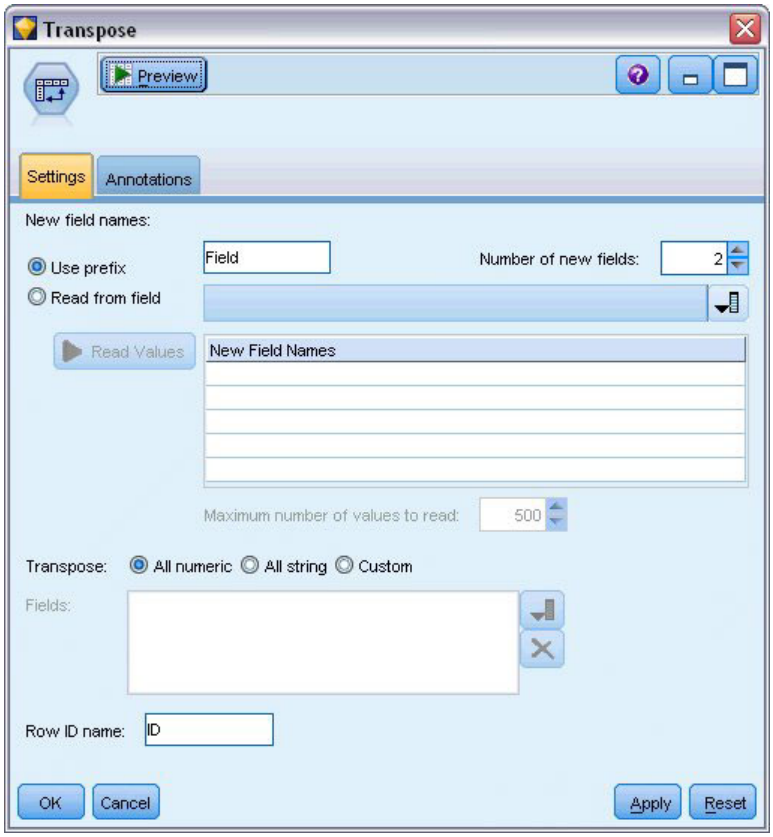


図 368. 入れ替えノード: 「設定」タブ

19. 入れ替えノードをフィルター・ノードに接続します。
20. 新しいフィールドの数として、2 を入力します。

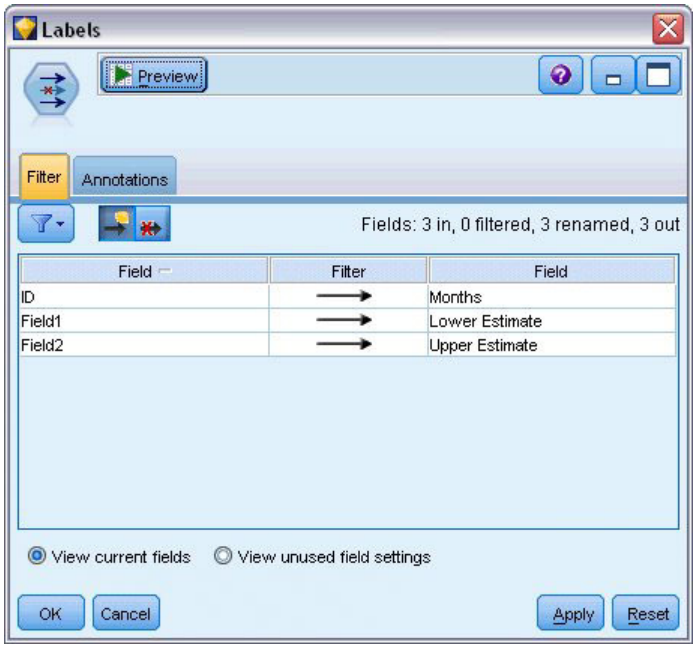


図 369. フィルター・ノード: 「フィルター」タブ

21. フィルター・ノードを入れ替えノードに接続します。
22. フィルター・ノードの「設定」タブで、「*ID* 」を「*月*」に、「*Field1*」を「*推定値の下限*」に、「*Field2*」を「*推定値の上限*」に名前を変更します。

図 370. 線グラフ・ノード: 「プロット」タブ

23. 線グラフ・ノードをフィルター・ノードに接続します。
24. 「プロット」タブで、「X フィールド」に「*月*」を、「Y フィールド」に「*推定値の下限*」と「*推定値の上限*」を設定します。

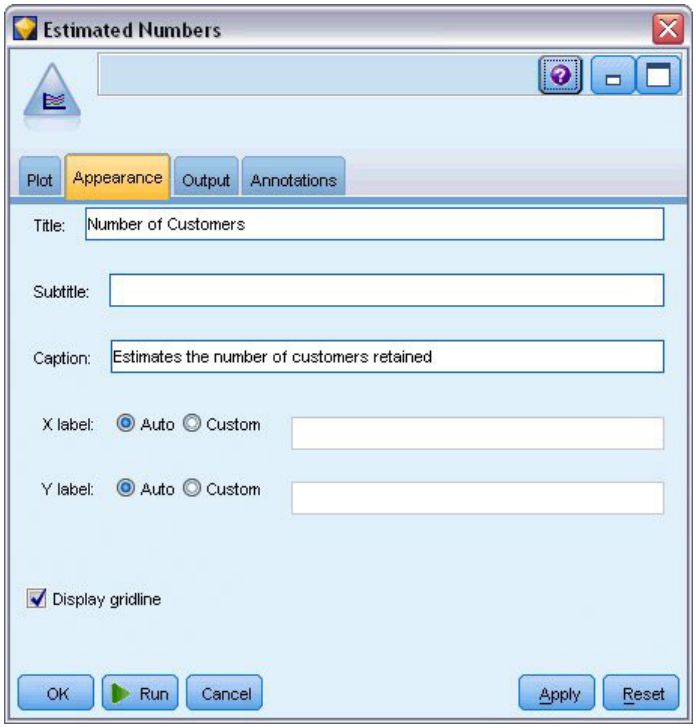


図 371. 線グラフ・ノード: 「外観」タブ

25. 「外観」タブをクリックします。
26. タイトルとして、「**顧客数**」と入力します。
27. キャプションとして、「**固定客数の予測**」と入力します。
28. 「**実行**」をクリックします。

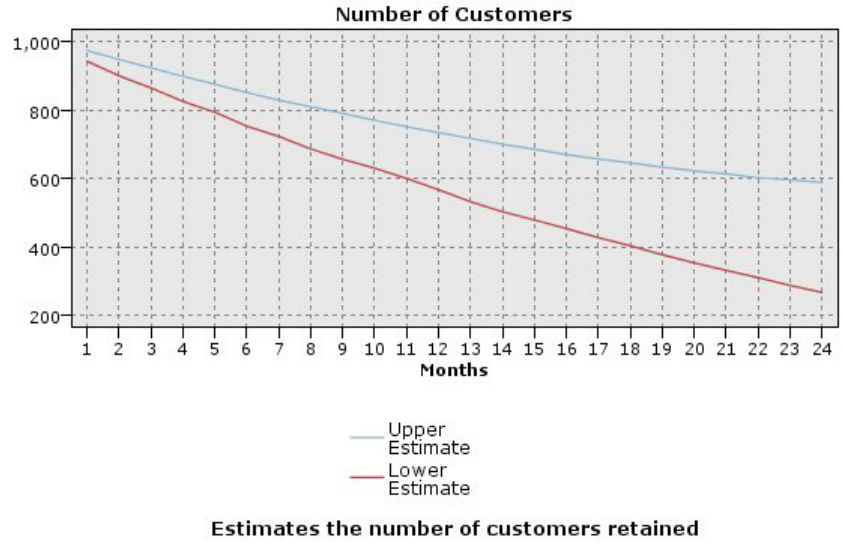


図 372. 固定客数を予測する線グラフ

予測される固定客数の上限と下限が作図されます。 2 本の線の間の差は Null とスコアリングされた顧客数です。そのため、その状況はかなり不確実です。時間とともに、これらの顧客数は増加します。12 カ月後にはデータ・セット内の元の顧客の 601 ～ 735 人が固定客であると期待されますが、24 カ

月後には 288 ～ 597 人になります。

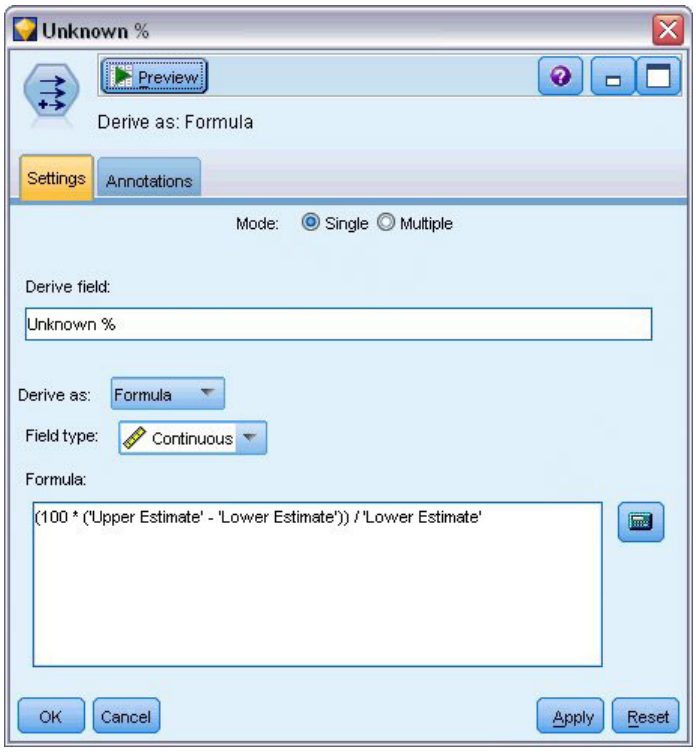


図 373. フィールド作成ノード: 「設定」タブ

29. 固定客数の推定がどれほど不確実であるかを見直すには、フィールド作成ノードをフィルター・ノードに接続します。
30. フィールド作成ノードの「設定」タブで、派生フィールドとして「*Unknown %*」と入力します。
31. フィールドのデータ型として「**連続型**」を選択します。
32. 式として、`(100 * ('Upper Estimate' - 'Lower Estimate')) / 'Lower Estimate'` と入力します。「*Unknown %*」は、「疑わしい」顧客の数を推定値の下限のパーセントで表します。
33. 「**OK**」をクリックします。

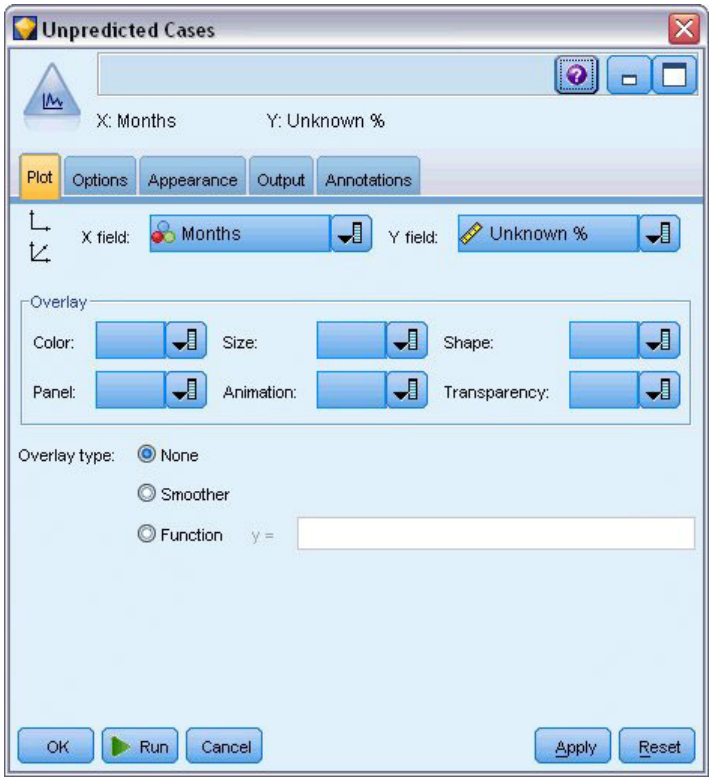


図 374. 作図ノード: 「プロット」タブ

34. 作図ノードをフィールド作成ノードに接続します。
35. 作成ノードの「プロット」タブで、X フィールドとして「*月*」を、Y フィールドとして「*Unknown %*」を選択します。
36. 「**外観**」タブをクリックします。

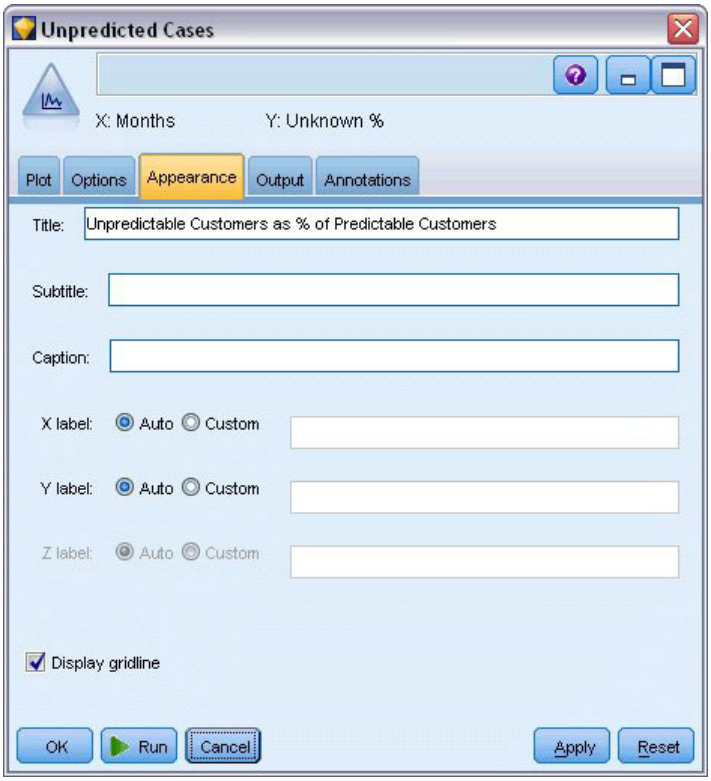


図 375. 作図ノード: 「外観」タブ

37. タイトルとして、「予測可能な顧客に対する予測不可能な顧客の割合」と入力します。
38. ノードを実行します。

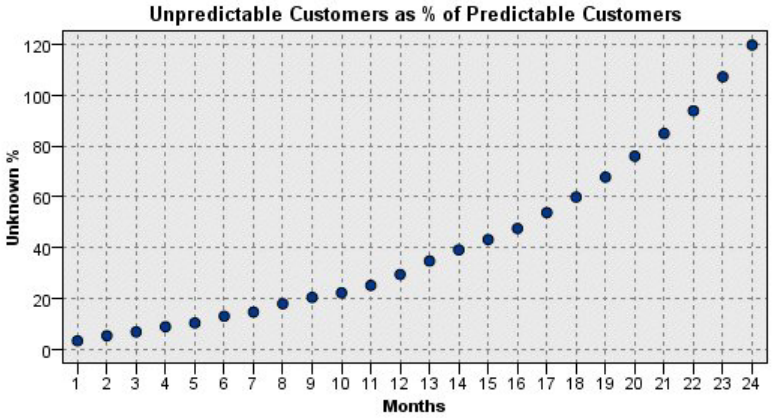


図 376. 予測不可能な顧客のプロット

最初の 1 年間は予測不可能な顧客のパーセントはほぼ線型速度で増加しますが、2 年目には増加率が急増し、23 カ月までに、Null 値を持つ顧客の数が、予測される固定客数を顧客の数を上回ります。

スコアリング

満足するモデルを得られたら、顧客をスコアリングして、来年中に解約すると予測される個人を四半期ごとに識別します。

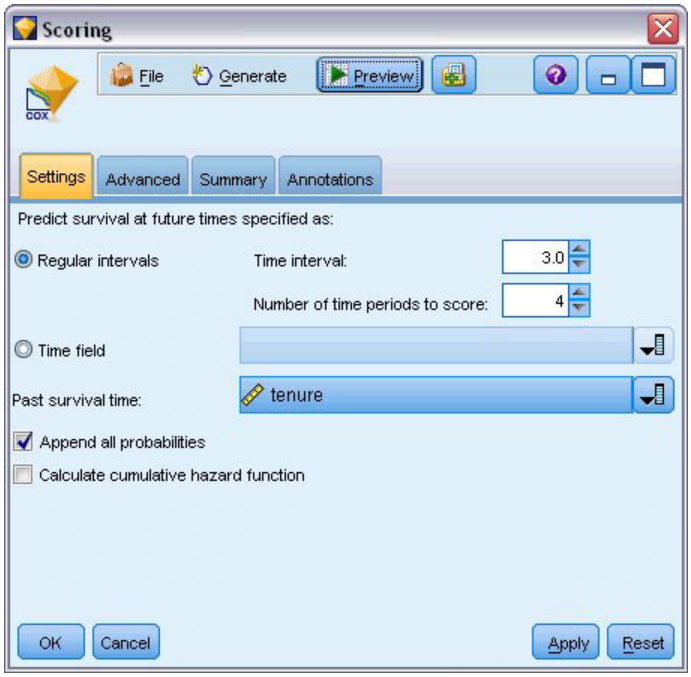


図 377. Coxreg ナゲット: 「設定」タブ

1. 3 番目のモデル・ナゲットをソース・ノードに接続して、そのモデル・ナゲットを開きます。
2. 「**一定の間隔**」が選択されていることを確認してから、時間間隔として「3.0」を、さらにスコアリングする期間数として「4」を指定します。これにより、各レコードが以降 4 四半期の間、スコアリングされるように指定されます。
3. 過去の生存時間を指定するためのフィールドとして「*tenure*」を選択します。スコアリング・アルゴリズムは、各顧客の会社の顧客としての時間の長さを考慮します。
4. 「**すべての確率を追加**」を選択します。これらの追加のフィールドにより、テーブルに表示する際のレコードのソートがより簡単になります。

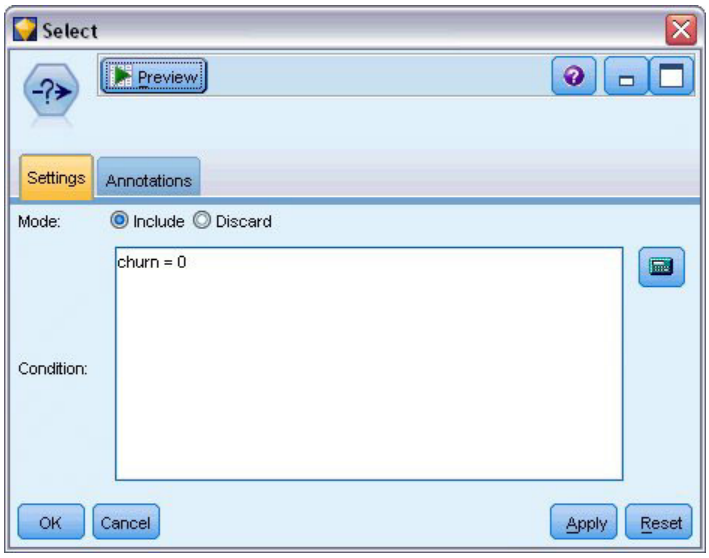


図 378. 条件抽出ノード: 「設定」タブ

5. 条件抽出ノードをモデル・ナゲットに接続します。「設定」タブで、条件として「churn=0」と入力します。これにより、すでに解約した顧客が結果テーブルから削除されます。

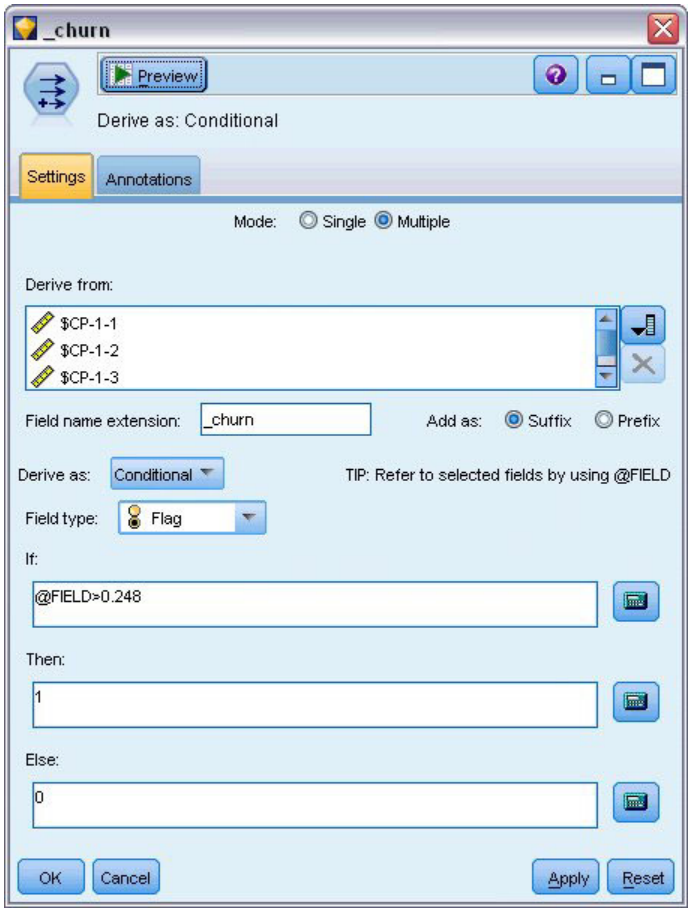


図 379. フィールド作成ノード: 「設定」タブ

6. フィールド作成ノードを条件抽出ノードに接続します。「設定」タブで、モードとして「**複数**」を選択します。
7. 派生元として「*$CP-1-1*」から「*$CP-1-4*」まで、形式が *$CP-1-n* のフィールドを選択し、追加する接尾辞として「`_churn`」を入力します。これは、「フィールドの選択」ダイアログ・ボックスで、フィールドを名前 (つまり、アルファベット順) でソートしていれば、最も簡単に行えます。
8. フィールドを「**条件付き**」として作成することを選択します。
9. 測定の尺度として「**フラグ**」を選択します。
10. 「**If**」 条件として、「`@FIELD>0.248`」を入力します。これは評価のときに識別された分類カットオフであることを思い出してください。
11. 「**Then**」式として「`1`」を入力します。
12. 「**Else**」式として「`0`」を入力します。
13. 「**OK**」をクリックします。

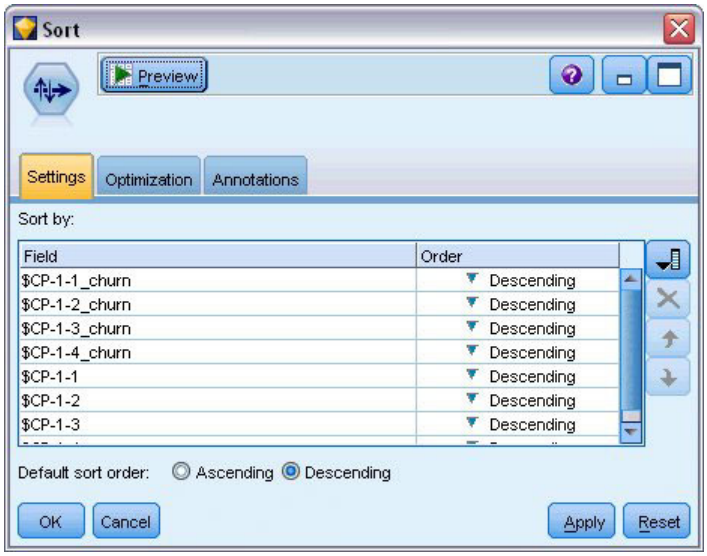
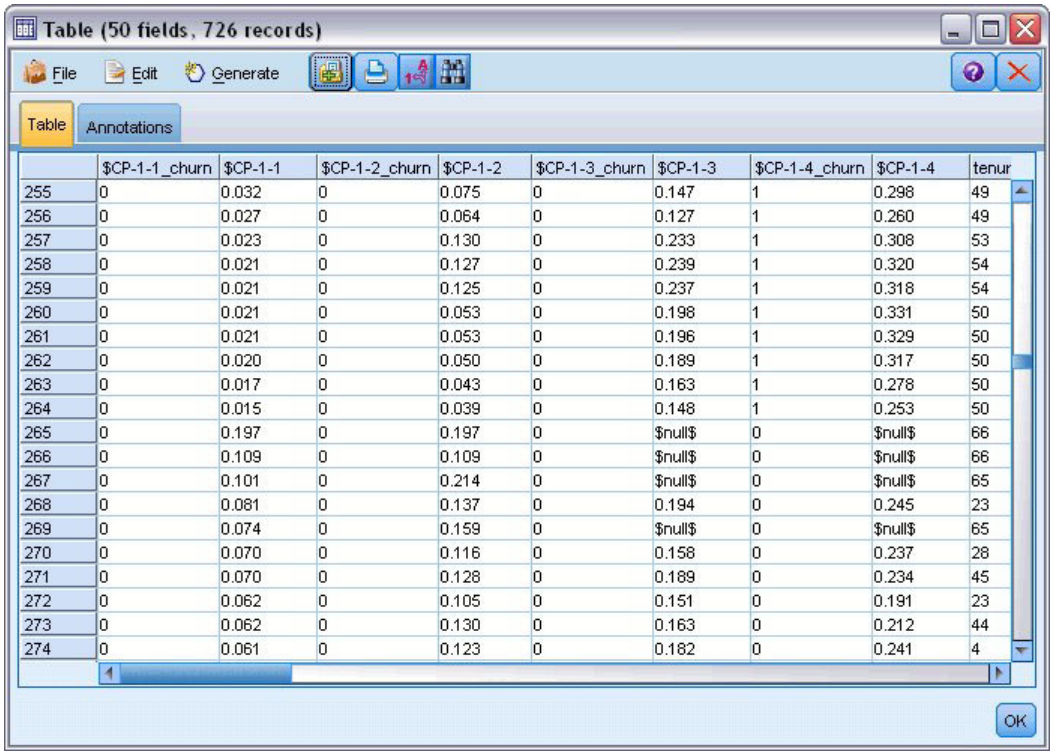


図 380. ソート・ノード: 「設定」タブ

14. ソート・ノードをフィールド作成ノードに接続します。「設定」タブで、*$CP-1-1_churn* から *$CP-1-4-churn* まで、および *$CP-1-1* から *$CP-1-4* までで、すべて降順でソートすることを選択します。解約が予測される顧客は上位に表示されます。

図 381. フィールドの並べ替えノード: 「並べ替え」タブ

15. フィールドの並べ替えノードをソート・ノードに接続します。「並べ替え」タブで、*$CP-1-1_churn* から *$CP-1-4* までを他のフィールドの前に配置されるように選択します。これは、ただ結果テーブルを見やすくするためなので、オプションです。図に示す位置にフィールドを移動するには、ボタンを使用します。

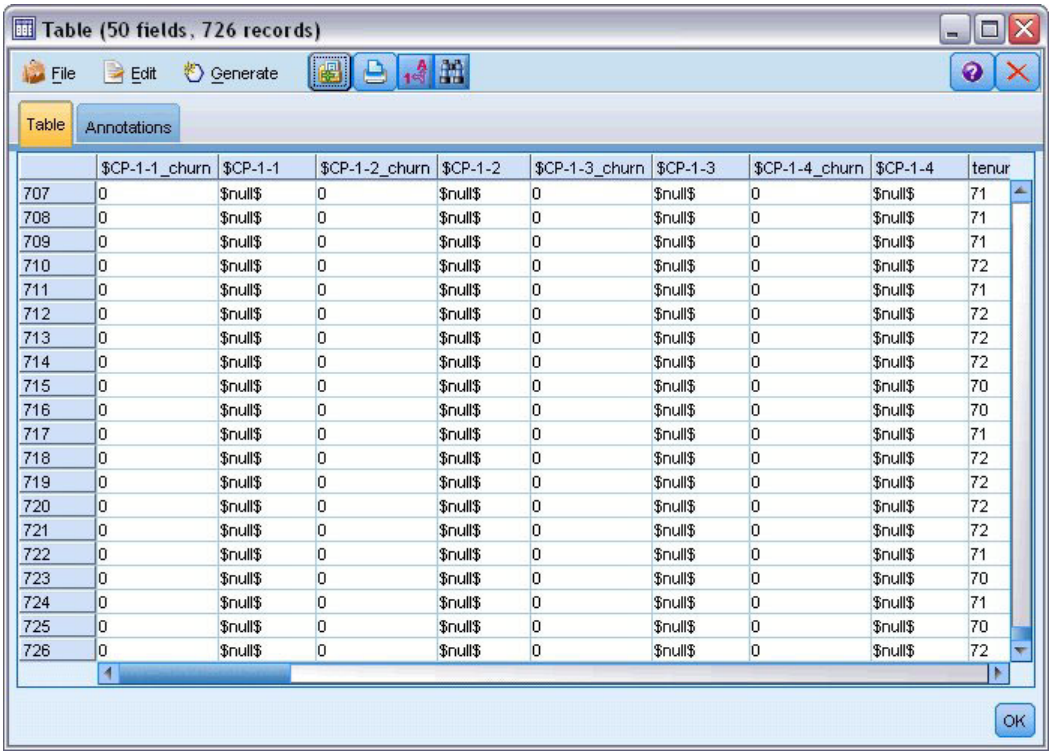


	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
255	0	0.032	0	0.075	0	0.147	1	0.298	49
256	0	0.027	0	0.064	0	0.127	1	0.260	49
257	0	0.023	0	0.130	0	0.233	1	0.308	53
258	0	0.021	0	0.127	0	0.239	1	0.320	54
259	0	0.021	0	0.125	0	0.237	1	0.318	54
260	0	0.021	0	0.053	0	0.198	1	0.331	50
261	0	0.021	0	0.053	0	0.196	1	0.329	50
262	0	0.020	0	0.050	0	0.189	1	0.317	50
263	0	0.017	0	0.043	0	0.163	1	0.278	50
264	0	0.015	0	0.039	0	0.148	1	0.253	50
265	0	0.197	0	0.197	0	$null$	0	$null$	66
266	0	0.109	0	0.109	0	$null$	0	$null$	66
267	0	0.101	0	0.214	0	$null$	0	$null$	65
268	0	0.081	0	0.137	0	0.194	0	0.245	23
269	0	0.074	0	0.159	0	$null$	0	$null$	65
270	0	0.070	0	0.116	0	0.158	0	0.237	28
271	0	0.070	0	0.128	0	0.189	0	0.234	45
272	0	0.062	0	0.105	0	0.151	0	0.191	23
273	0	0.062	0	0.130	0	0.163	0	0.212	44
274	0	0.061	0	0.123	0	0.182	0	0.241	4

図 382. 顧客のスコアを示すテーブル

16. テーブル・ノードをフィールドの並べ替えノードに接続して、実行します。

年末までに 264 人の顧客が解約し、第 3 四半期の終わりまでに 184 人、第 2 四半期までに 103 人、第 1 四半期までに 31 人が解約すると予測されます。特定の 2 人の顧客に注目すると、第 1 四半期に解約の傾向がより強かった顧客が、後の四半期でも解約の傾向が強いとは限りません。例えば、レコード 256 と 260 を参照してください。これは、顧客の現在の保有期間の後の数カ月のハザード関数の形状に起因するようです。例えば、販売促進活動の際に加入した顧客は、人から勧められて加入した顧客よりも早期に他社に切り替える可能性が高くなりますが、切り替えなかった場合、そのような顧客は、実際のところ残りの保有期間はより忠実な顧客になる傾向があります。顧客をソートし直して、解約する可能性が最も高い顧客について別の見方をすることもできます。



	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
707	0	$null$	0	$null$	0	$null$	0	$null$	71
708	0	$null$	0	$null$	0	$null$	0	$null$	71
709	0	$null$	0	$null$	0	$null$	0	$null$	71
710	0	$null$	0	$null$	0	$null$	0	$null$	72
711	0	$null$	0	$null$	0	$null$	0	$null$	71
712	0	$null$	0	$null$	0	$null$	0	$null$	72
713	0	$null$	0	$null$	0	$null$	0	$null$	72
714	0	$null$	0	$null$	0	$null$	0	$null$	72
715	0	$null$	0	$null$	0	$null$	0	$null$	70
716	0	$null$	0	$null$	0	$null$	0	$null$	70
717	0	$null$	0	$null$	0	$null$	0	$null$	71
718	0	$null$	0	$null$	0	$null$	0	$null$	72
719	0	$null$	0	$null$	0	$null$	0	$null$	72
720	0	$null$	0	$null$	0	$null$	0	$null$	72
721	0	$null$	0	$null$	0	$null$	0	$null$	72
722	0	$null$	0	$null$	0	$null$	0	$null$	71
723	0	$null$	0	$null$	0	$null$	0	$null$	70
724	0	$null$	0	$null$	0	$null$	0	$null$	71
725	0	$null$	0	$null$	0	$null$	0	$null$	70
726	0	$null$	0	$null$	0	$null$	0	$null$	72

図 383. Null 値の顧客を示すテーブル

テーブルの一番下に、予測値が Null の顧客が表示されています。これらの顧客は、合計保有期間 (将来の時間 + *tenure*) がモデルの学習に使用されたデータの生存時間の範囲外にあります。

要約

Cox 回帰を使用して、解約するまでの時間に関する妥当なモデルを見つけ、今後 2 年間に固定客でい続けると予想される顧客の数をプロットし、来年中に解約する可能性が最も高い顧客を個々に識別しました。これは妥当なモデルではあるものの、最適なモデルではない可能性もあります。理想的には、少なくとも、変数増加ステップワイズ法を使用して得られたこのモデルを、変数減少ステップワイズ法を使用して作成されたモデルと比較する必要があります。

IBM SPSS Modeler で使用されるモデル作成手法の数学的な基礎の説明は、「*IBM SPSS Modeler アルゴリズム・ガイド* 」を参照してください。

第 27 章 マーケット・バスケット分析 (ルール帰納/C5.0)

この例では、スーパーマーケットのバスケット (つまり同時に購入される品目の集まり) の内容を説明する架空のデータと、購入者の関連個人データを扱います。この個人データは、ロイヤルティー・カード・スキームから取得される可能性があります。目的は、類似した製品を購入し、年齢や収入などによって人口統計的に特徴付けることができる顧客のグループを発見することです。

この例では、次のようなデータ・マイニングの 2 つのフェーズについて説明します。

- アソシエーション・ルール・モデル作成と Web グラフ表示によって、購入されるアイテム間のリンクを明らかにします。
- C5.0 ルール帰納によって、識別された製品グループの購入者のプロファイルを作成します。

注: この適用では予測モデル作成を直接使用しません。そのため、結果のモデルの精度は測定されず、またデータ・マイニング・プロセスにおける関連付けられた学習/検定の区別もありません。

この例では、*BASKETS1n* というデータ・ファイルを参照する *baskrule* というストリームを使用します。これらのファイルは、IBM SPSS Modeler インストール環境の *Demos* ディレクトリーにあります。このディレクトリーには、Windows の「スタート」メニューの IBM SPSS Modeler プログラム・グループからアクセスできます。*baskrule* ファイルは、*streams* ディレクトリー内にあります。

データへのアクセス

可変長ファイル・ノードを使用して、データ・セット *BASKETS1n* に接続し、ファイルからフィールド名を読み込むことを選択します。データ型ノードをデータ・ソースに接続し、次にそのノードをテーブル・ノードに接続します。フィールド *cardid* の測定の尺度を「*データ型不明*」に設定します (各ロイヤルティー・カード ID は、データ・セット内で 1 回しか発生しないため、モデル作成では役に立ちません)。フィールド *sex* の測定の尺度に「*名義型*」を選択します (Apriori モデル・アルゴリズムが *sex* をフラグ型として処理しないようにするため)。

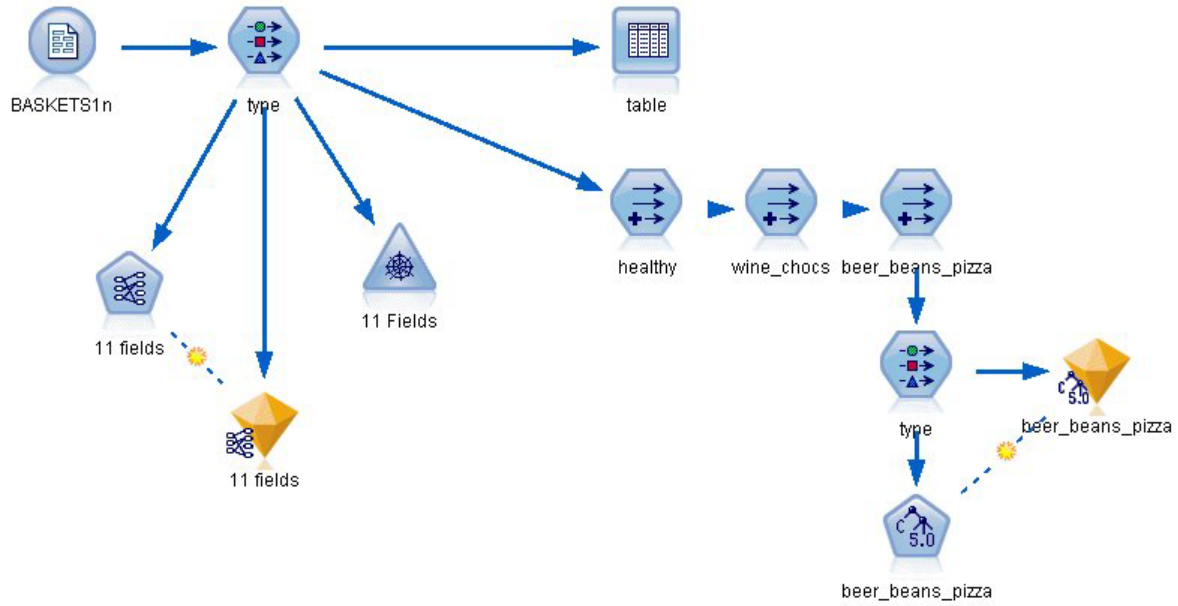


図 384. baskrule ストリーム

次に、ストリームを実行して、データ型ノードをインスタンス化し、テーブルを表示します。データ・セットには 18 のフィールドが含まれており、各レコードはバスケットを表します。

18 のフィールドは、それぞれ次の見出しで表されています。

バスケットの要約:

- *cardid*。このバスケットを購入する顧客のロイヤルティー・カードの ID です。
- *value*。バスケットの合計購入金額です。
- *pmethod*。バスケットの支払方法です。

カード所有者の個人情報の詳細:

- *sex*
- *homeown*。カード所有者が家屋を所有しているかどうか。
- *income*
- *age*

バスケットの内容 - 製品カテゴリーの有無を示すフラグ:

- *fruitveg*
- *freshmeat*
- *dairy*
- *cannedveg*
- *cannedmeat*
- *frozenmeal*
- *beer*
- *wine*
- *softdrink*

- *fish*
- *confectionery*

バスケットの内容における密接な関係の発見

始めに、アソシエーション・ルールを生成するための Apriori を使用してバスケットの内容における密接な関係 (相関) の概観を把握する必要があります。データ型ノードを編集し、すべての製品カテゴリーの役割を*両方* に設定し、他のすべての役割をなし に設定することで、このモデル作成プロセスで使用するフィールドを選択します。(*両方* とは、フィールドが結果モデルの入力または出力のいずれにすることもできることを意味します)。

注: 列からオプションを指定する前に、Shift キーを押しながら複数のフィールドを選択することで、それらの複数のフィールドに対してオプションを設定することができます。

図 385. モデル作成用フィールドの選択

モデル作成用フィールドを指定したら、Apriori ノードをデータ型ノードに接続して、Apriori ノードを編集し、「**フラグは真の値のみ**」オプションを選択して、Apriori ノードの実行をクリックします。結果は、マネージャー・ウィンドウの右上の「モデル」タブにモデルとして表示されます。この結果には、アソシエーション・ルールが含まれており、コンテキスト・メニューで「**参照**」を選択して表示することができます。

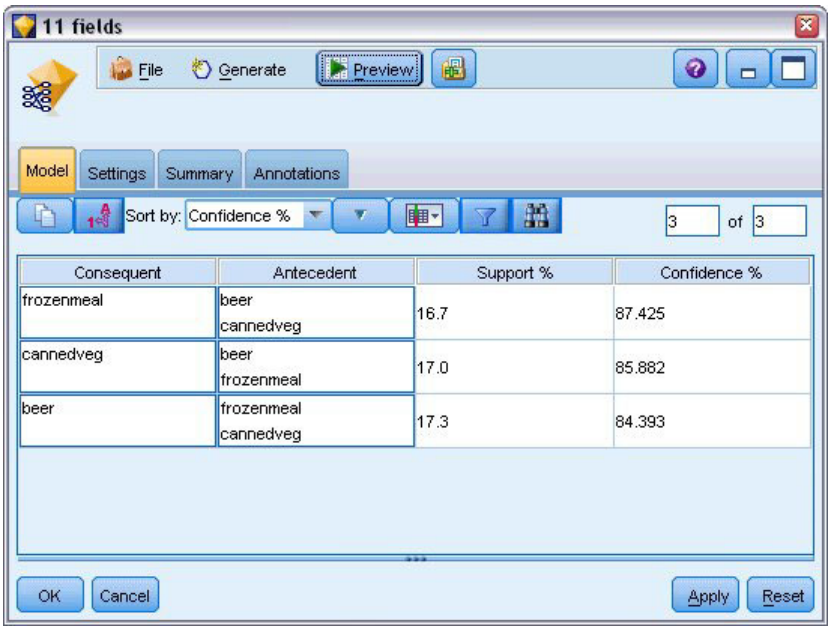


図 386. アソシエーション・ルール

これらのルールは、冷凍食品、缶詰野菜、およびビールのさまざまな関連性を示します。次のような双方向のアソシエーション・ルールが存在します。

```
frozenmeal -> beer
beer -> frozenmeal
```

これは、Web 表示 (双方向の関連だけを表示します) がこのデータ内のいくつかのパターンを強調する場合があることを示します。

Web ノードをデータ型ノードに接続して、Web ノードを編集し、すべてのバスケットの内容フィールドを選択します。次に、「**フラグは真の値のみ**」を選択し、Web ノードの実行をクリックします。

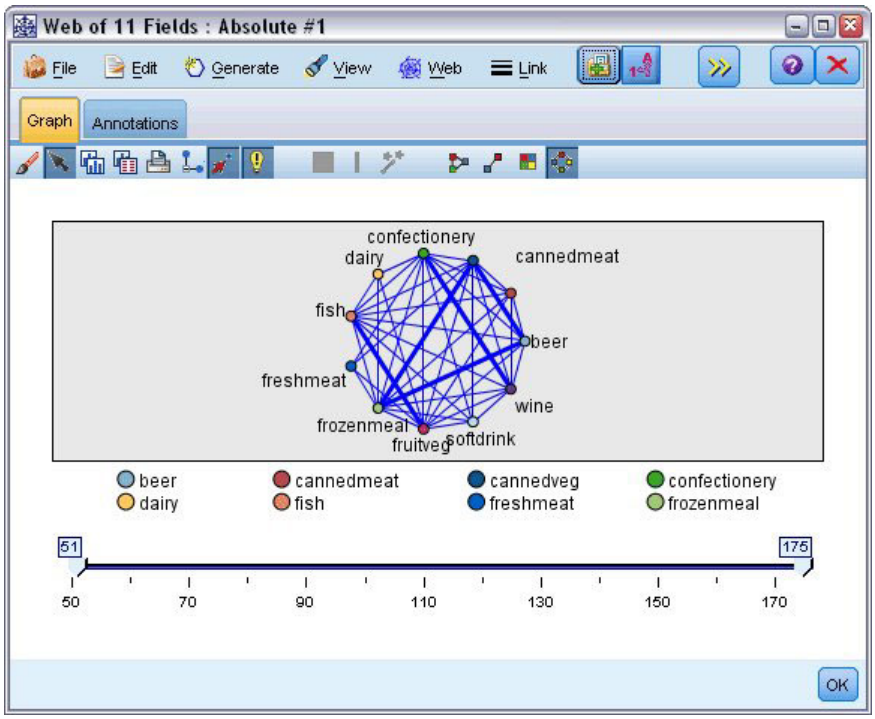


図 387. 製品の連関の Web グラフ表示

大半の製品カテゴリーの組み合わせが複数のバスケット内で発生しているため、この Web グラフ上の密接なリンクが多すぎて、モデルで示された顧客のグループを示すことができません。

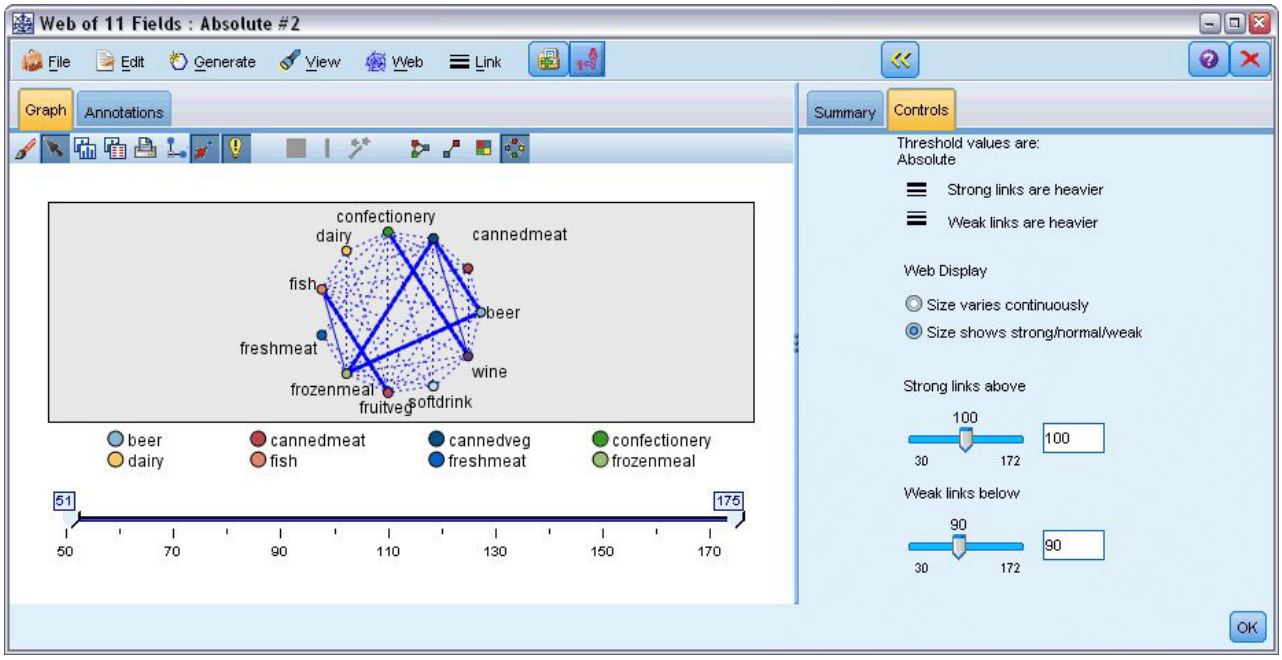


図 388. 制限された Web グラフ表示

1. 弱い相関と強い相関を指定するには、ツールバーの黄色い二重矢印ボタンをクリックします。ダイアログ・ボックスが展開され、Web 出力の概要およびコントロールが表示されます。
2. 「**サイズで強い/通常/弱いを表示**」を選択します。

3. 弱いリンクを 90 未満に設定します。
4. 強いリンクを 100 超に設定します。

結果のグラフ表示内では、次の 3 つの顧客のグループが目立っています。

- 魚と果物と野菜を購入するグループ (「健康的なグループ」と呼ぶことができます)
- ワインと菓子類を購入するグループ
- ビール、冷凍食品、および缶詰野菜を購入するグループ (「ビールと豆とピザのグループ」)

顧客グループのプロファイル作成

これで、購入した製品のタイプを基に、3 つの顧客のグループを識別しました。しかし、さらにこれらの顧客がどのような顧客か (つまりこれらの顧客の人口統計的なプロファイル) を知る必要が生じる場合もあります。これは、各顧客にこれらの各グループを示すフラグを使用してタグを付け、さらにルール帰納 (C5.0) を使用して、これらのフラグのルール・ベースのプロファイルを作成することによって実現できます。

はじめに、各グループのフラグを派生させる必要があります。これは、先ほど作成した Web グラフ表示を使って自動的に生成することができます。右マウス・ボタンで、*fruitveg* と *fish* 間のリンクをクリックして強調表示してから、右クリックし、「**リンクのフィールド作成ノード生成**」を選択します。

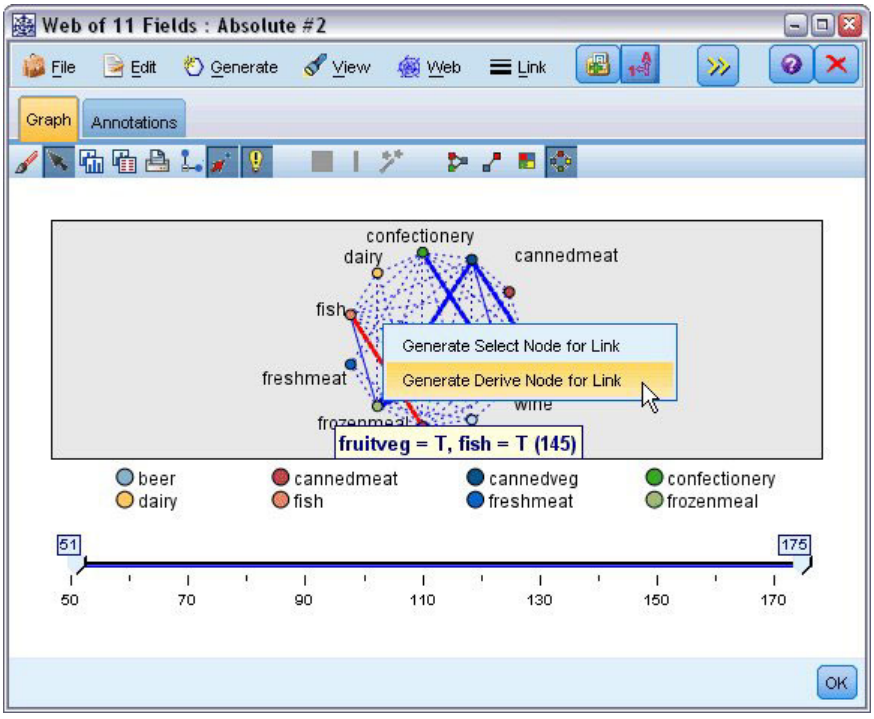


図 389. 各顧客グループのフラグの派生

結果のフィールド作成ノードを編集して、派生フィールド名を *healthy* に変更します。 *wine* から *confectionery* へのリンクを使用してこの作業を繰り返し、結果の派生フィールドに *wine_chocs* という名前を付けます。

第 3 のグループでは (3 つのリンクが関与します)、最初にリンクが選択されていないことを確認します。次に、左マウス・ボタンをクリックしながらシフト・キーを押して、*cannedveg*、*beer*、および *frozenmeal*

の三角形の 3 つすべてのリンクを選択します。 (必ず、編集モードではなくインタラクティブ・モードにしてください)。それから、Web グラフ表示メニューから、次のオプションを選択します。

「生成」 > 「フィールド作成ノード (AND)」

結果の派生フィールドの名前を *beer_beans_pizza* に変更します。

これらの顧客グループのプロファイルを作成するには、既存のデータ型ノードをこれらの 3 つのフィールド作成ノードに直列に接続し、次に別のデータ型ノードに接続します。新規データ型ノードで、すべてのフィールドの役割をなし に設定します。ただし、例外として、*value*、*pmethod*、*sex*、*homeown*、*income*、および *age* は入力 に設定する必要があり、関連する顧客グループ (例えば、*beer_beans_pizza*) は対象 に設定する必要があります。C5.0 ノードを接続し、出力タイプを**「ルール・セット」**に設定し、ノードの実行をクリックします。以下のように、結果のモデル (*beer_beans_pizza*) には、この顧客グループの明確な人口統計プロファイルが含まれています。

```
Rule 1 for T:
if sex = M
and income <= 16,900
then T
```

2 番目のデータ型ノードで出力として選択することで、同じ方法を他の顧客グループ・フラグに適用できます。このコンテキストで C5.0 ではなく Apriori を使用することで、より幅広い代替プロファイルを生成できます。Apriori は、1 つの出力フィールドに制限されないため、すべての顧客グループのフラグのプロファイルを同時に作成するために使用することもできます。

要約

この例では、IBM SPSS Modeler を使用して、モデル作成 (Apriori を使用) とビジュアル化 (Web グラフ表示を使用) の両方によって、データベース内の密接な関係またはリンクを発見する方法を示しています。これらのリンクはデータ内のケースのグループ化に対応し、これらのグループはモデル作成 (C5.0 ルール・セットを使用) によって、詳しく調べて、プロファイルを作成できます。

小売業分野では、例えばこのように顧客をグループ分けし、特別オファーの対象にしてダイレクト・メールに対する反応率を向上させたり、人口統計ベースの需要に合わせて支店に在庫として置く製品の範囲を調整したりできます。

第 28 章 新しい自動車製品の評価 (KNN)

最近隣分析は、その他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。機械学習で、保存されたパターンまたはケースに完全一致させることを必要とせず、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。類似したケースは互いに近く、類似していないケースは互いに離れています。つまり、2 つのケース間の距離は、それらの非類似性の尺度です。

互いに近いケースは「近隣」であると考えられます。新規ケース (ホールドアウト) が示された場合、モデル内の各ケースからの距離が計算されます。最も類似している (最近隣) ケースの分類が集計され、新規のケースは最大数の最近隣を含むカテゴリーに配置されます。

検証する最近隣の数を指定することができ、この値は *k* と呼ばれます。図では、2 つの異なる値 *k* を使用して、新規のケースがどのように分類されるかを示します。*k* = 5 の場合、最近隣の大多数がカテゴリー *1* に属しているため、新規のケースはカテゴリー *1* に配置されます。ただし、*k* = 9 の場合、新規のケースは最近隣の大多数がカテゴリー *0* に属しているため、カテゴリー *0* に配置されます。

また、最近隣分析を使用して、連続型対象の値を計算することもできます。この場合、最近隣の平均値または対象の中央値を使用して、新規のケースの予測値を取得します。

自動車メーカーが、2 つの新しい自動車 (乗用車およびトラック) のプロトタイプを開発しています。新しいモデルを範囲に導入する前に、メーカーは市場にある既存の自動車でどれが最もプロトタイプに近いのか、つまりどの自動車が「最近隣」なのか、そのためどのモデルが競争相手となるのかを判断する必要があります。

メーカーはさまざまなカテゴリーの既存のモデルに関するデータを収集し、そのプロトタイプの詳細情報を追加しました。モデルを比較するカテゴリーには、価格 (単位: 千) (*price*)、エンジンのサイズ (*engine_s*)、馬力 (*horsepow*)、ホイールベース (*wheelbas*)、幅 (*width*)、全長 (*length*)、重量 (*curb_wgt*)、燃料積載量 (*fuel_cap*) および燃料効率 (*mpg*) があります。

この例では、*Demos* フォルダーの *streams* サブフォルダー内にある *car_sales_knn.str* というストリームを使用します。データ・ファイルは *car_sales_knn_mod.sav* です。詳細については、5 ページの『「Demos」フォルダー』を参照してください。

ストリームの作成

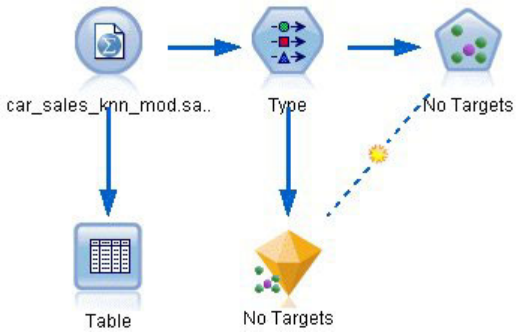
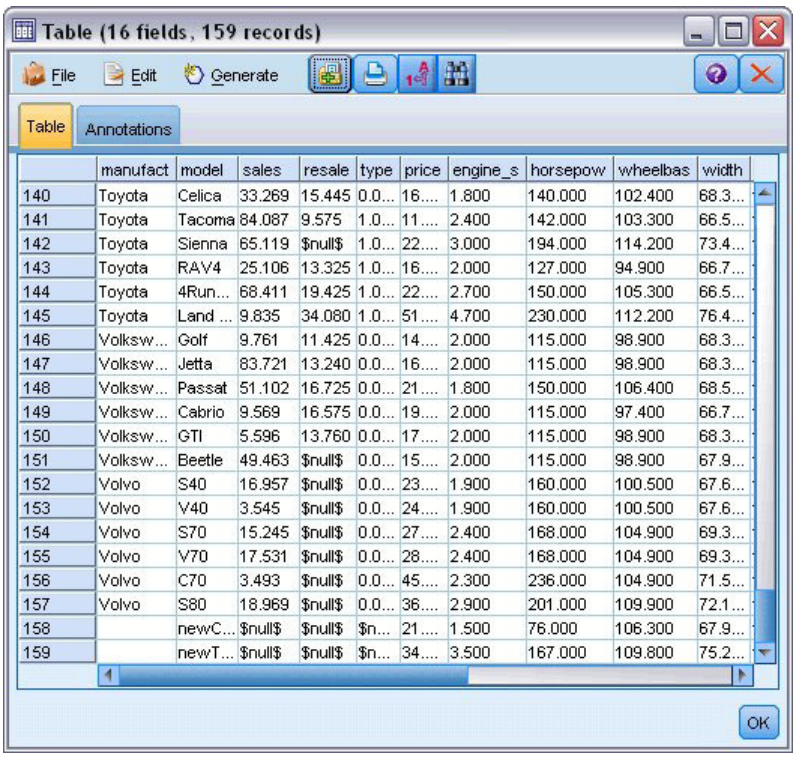


図 390. KNN モデル作成のサンプル・ストリーム

新規ストリームを作成し、IBM SPSS Modeler インストール環境の *Demos* フォルダーにある *car_sales_knn_mod.sav* を指し示す Statistics ファイル入力ノードを追加します。

まず、メーカーが収集したデータについて見てみましょう。

1. テーブル・ノードを Statistics ファイル入力ノードに接続します。
2. テーブル・ノードを開いて、「**実行**」をクリックします。



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

図 391. 乗用車およびトラックのソース・データ

newCar と *newTruck* という 2 つのプロトタイプの詳細が、ファイルの最後に追加されています。

ソース・データから、メーカーはかなりあいまいな「トラック」の分類 (*type* 列の値 1) を使用していて、自動車以外の種類の車両を意味していることが分かります。

最近隣を特定する場合に 2 つのプロトタイプをホールドアウトの順序で指定できるようにするには、最後の列 *partition* が必要です。このように、これらのデータは、検討に入れる市場の残りの部分であるため、計算に影響しません。2 つのホールドアウト・レコードの *partition* の値を 1 に設定し、このフィールドの他のすべてのレコードを 0 に設定すると、重要レコード (最近隣を計算する対象レコード) を設定する際に後でこのフィールドを使用できます。

後から参照するので、テーブル出力ウィンドウは開いたままにします。

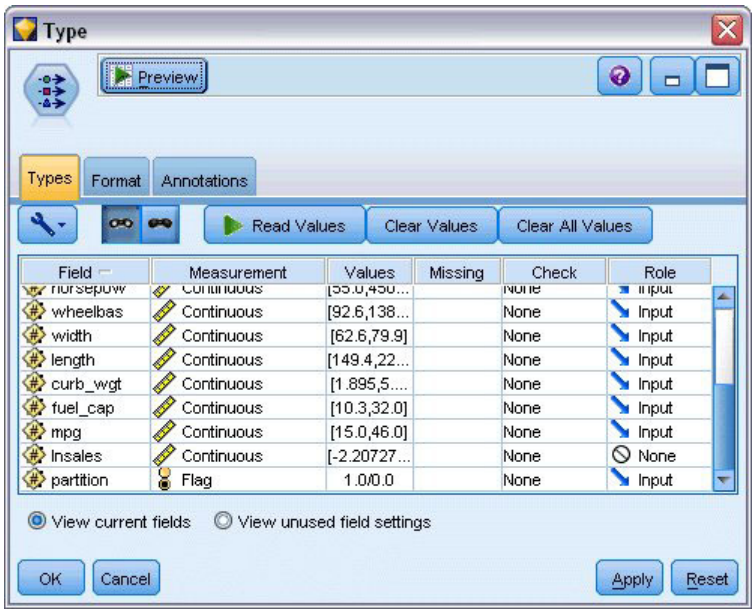


図 392. データ型ノードの設定

3. データ型ノードをストリームに追加します。
4. データ型ノードを Statistics ファイル入力ノードに接続します。
5. データ型ノードを開きます。

   フィールド *price* から *mpg* までのみを比較するため、これらのすべてのフィールドの役割は「**入力**」の設定のままにします。
6. その他のすべてのフィールド (*manufact* から *type* と *lnsales*) の役割を「**なし**」に設定します。
7. 最後のフィールド *partition* の測定の尺度を「**フラグ**」に設定します。その役割が「**入力**」に設定されていることを確認してください。
8. 「**値の読み込み**」をクリックしてデータ値をストリームに読み込みます。
9. 「**OK**」をクリックします。

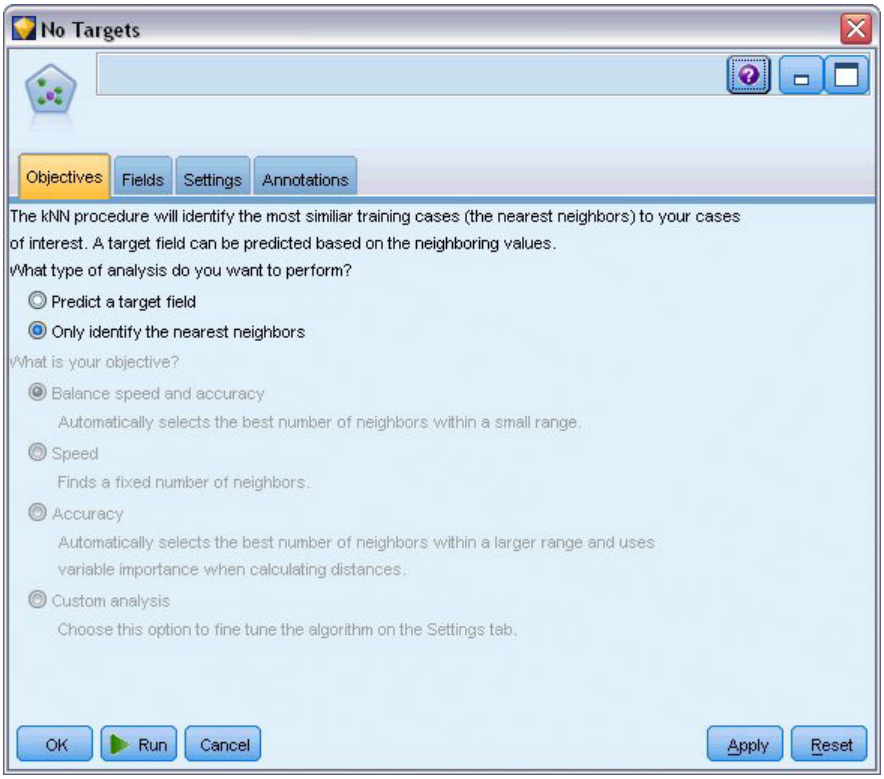


図 393. 最近隣の特定の選択

10. KNN ノードをデータ型ノードに接続します。
11. KNN ノードを開きます。

    2 つのプロトタイプの最近隣を見つけるだけなので、今回は対象フィールドの予測は行っていません。
12. 「**目的**」タブで、「**最近隣のみを特定**」を選択します。
13. 「**設定**」タブをクリックします。

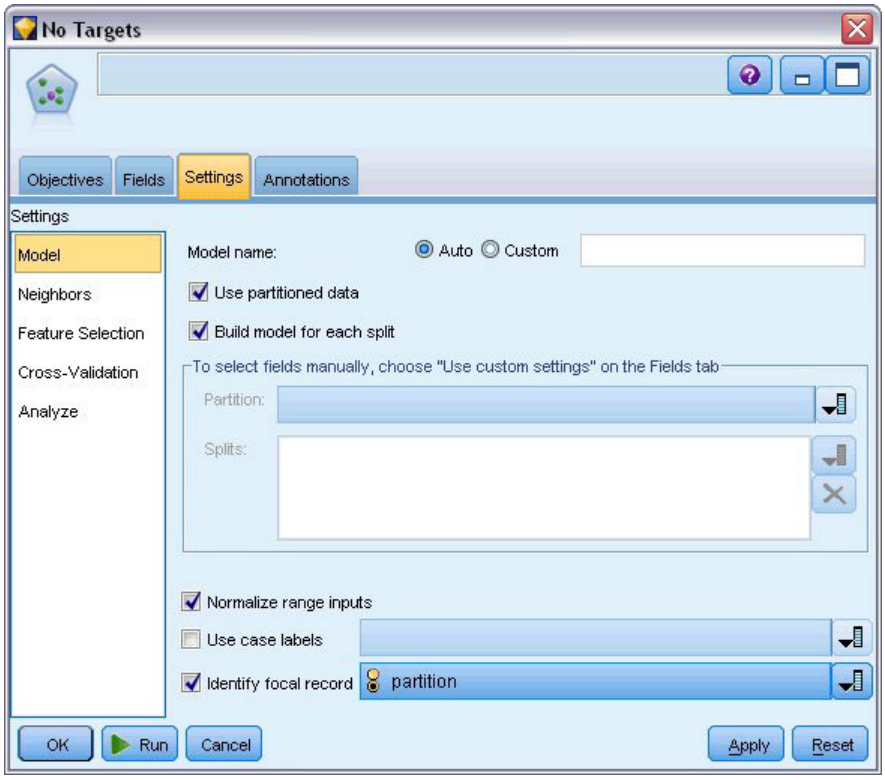


図 394. データ区分フィールドを使用した重要レコードの特定

ここでは、データ区分 フィールドを使用して重要レコード (最近隣を特定する対象のレコード) を特定できます。フラグ型フィールドを使用して、このフィールドの値が 1 に設定されたレコードが重要レコードとなっていることを確認します。

前述のとおり、このフィールドの値が 1 になっているレコードは *newCar* および *newTruck* のみであるため、それらのレコードが重要レコードになります。

14. 「**設定**」タブの「**モデル**」パネルで、「**重要レコードの特定**」チェック・ボックスを選択します。
15. このフィールドのドロップダウン・リストから、「**データ区分**」を選択します。
16. 「**実行**」ボタンをクリックします。

出力の調査

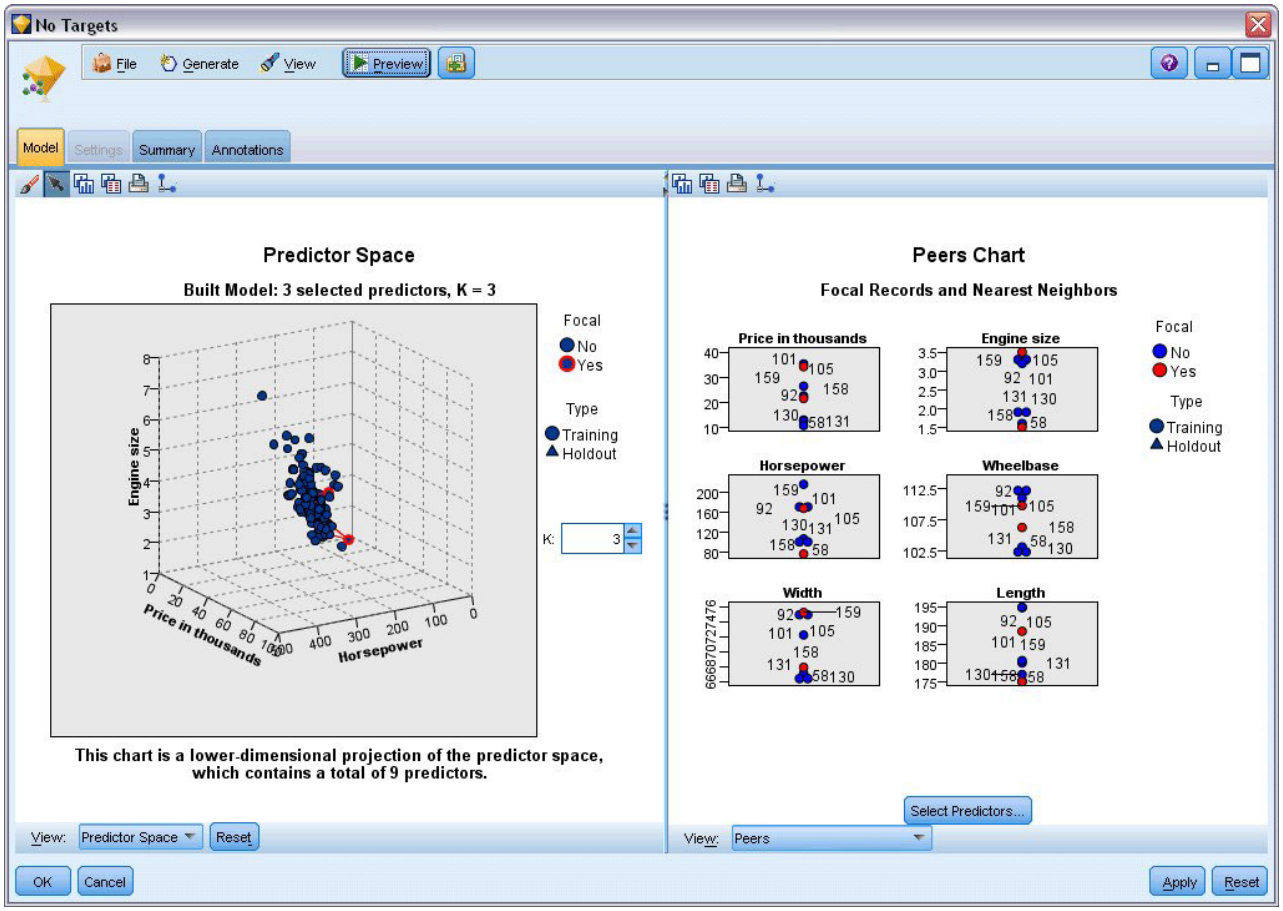


図 395. モデル・ビューアー・ウィンドウ

モデル・ナゲットがストリーム領域およびモデル・パレットに作成されています。ナゲットのいずれかを開くとモデル・ビューアーが表示されます。モデル・ビューアーには次の 2 つのパネルから成るウィンドウがあります。

- 1 つ目のパネルはメイン・ビューと呼ばれ、モデルの概要が表示されます。最近隣モデルのメイン・ビューは、**予測領域**と呼ばれます。
- 2 つ目のパネルには、次の 2 種類のビューのいずれかが表示されます。

  補助的モデル・ビューには、モデルの詳細が表示されますが、モデル自体に焦点は当たっていません。

  リンク・ビューは、メイン・ビューの一部についてドリルダウンしたときの、モデルのある特徴についての詳細を示すビューです。

予測値の領域

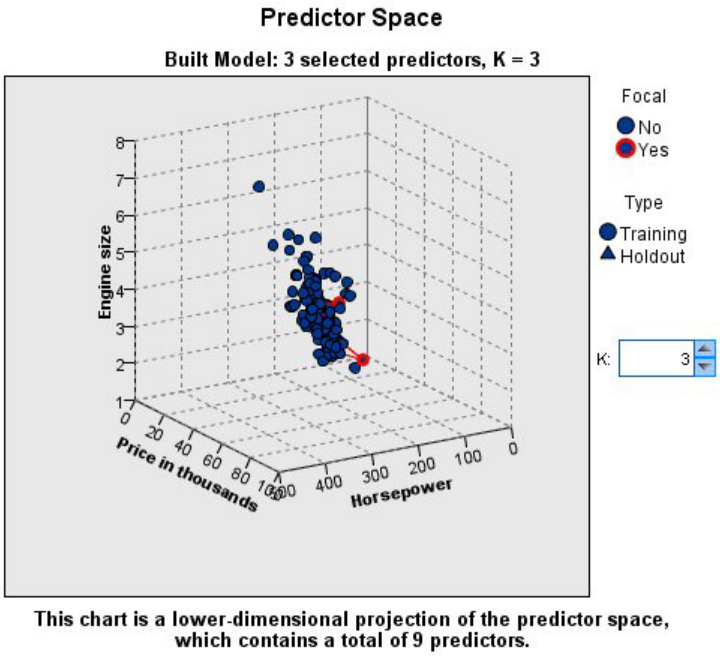


図 396. 予測領域のグラフ

予測領域のグラフは、3 つの特徴 (実際はソース・データの最初の 3 つの入力フィールド) のデータ・ポイントをプロットするインタラクティブな 3 次元のグラフで、価格、エンジンのサイズ、馬力を示します。

2 つの中心レコードは赤く強調表示され、それらと *k* の最近隣が線でつながれています。

グラフをクリックしてドラッグし、予測領域のポイントの分布をより分かりやすく表示するために、グラフを回転させることができます。「**リセット**」ボタンをクリックすると、デフォルト・ビューに戻ります。

同位図

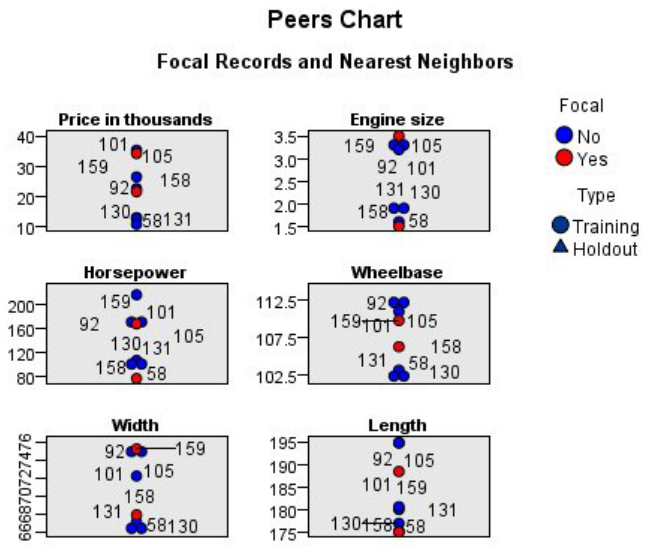


図 397. 同位図

デフォルトの補助ビューは同位図です。同位図では、予測領域で選択した 2 つの中心レコードと、6 つの特徴 (ソース・データの最初の 6 つの入力フィールド) それぞれの *k* の最近隣が強調表示されます。

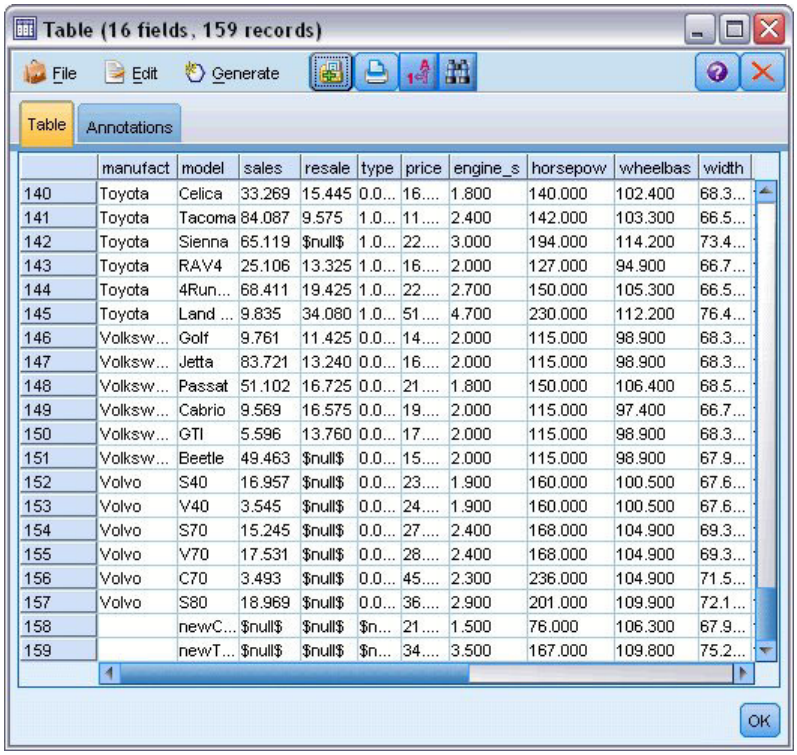
自動車は、ソース・データのレコード番号で表されます。ここで、自動車を特定するために、テーブル・ノードからの出力が必要です。

テーブル・ノード出力がまだ使用可能である場合、次の手順を実行します。

1. メイン IBM SPSS Modeler ウィンドウの右上にあるマネージャー・ペインの「**出力**」タブをクリックします。
2. 項目「**テーブル (16 フィールド、159 レコード)**」をダブルクリックします。

テーブル出力が使用できなくなっている場合、次の手順を実行します。

3. メイン IBM SPSS Modeler ウィンドウで、テーブル・ノードを開きます。
4. 「**実行**」をクリックします。



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

図 398. レコード番号によるレコードの特定

テーブルの下部にスクロールすると、*newCar* および *newTruck* がデータの最後の 2 つのレコードであり、それぞれ 158 および 159 の番号が付いていることが分かります。

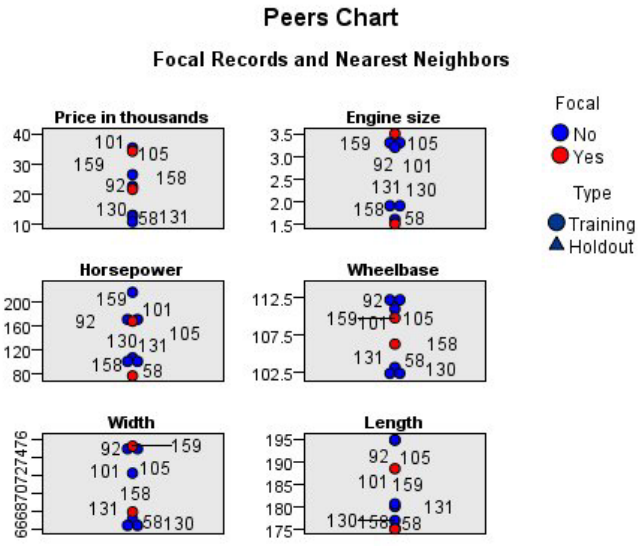


図 399. 同位図の特徴の比較

このことから同位図では、例えば、*newTruck* (159) のエンジンのサイズはいずれの最近隣よりも大きく、*newCar* (158) のエンジンは、その 最近隣のどれよりも小さいことが分かります。

6 つの特徴それぞれについて、特定の点の上にマウスを移動し、その特定のケースについて各特徴の実際の値を表示できます。

しかし、どの自動車が *newCar* と *newTruck* の最近隣なのでしょうか。

同位図は、若干混雑しているため、より単純な表示に変更しましょう。

5. 同位図の下部にある「**表示**」ドロップダウン・リスト (現在「**同位**」となっている項目) をクリックします。
6. 「**近隣および距離の表**」を選択します。

近隣および距離のテーブル

k Nearest Neighbors and Distances

Displayed for Initial Focal Records

Focal Record	Nearest Neighbors			Nearest Distar	
	1	2	3	1	2
158	131	130	58	0.979	0.990
159	105	92	101	0.580	0.634

図 400. 近隣および距離のテーブル

これで見やすくなりました。これで、2 つのプロトタイプのそれぞれに対し、市場で最も近い 3 つのモデルを確認できます。

newCar (中心レコード 158) の場合、Saturn SC (131)、Saturn SL (130)、および Honda Civic (58) です。

驚くべきことはありません。3 つの車すべてが中型のセダン型自動車であるため、特に燃料効率に優れており、*newCar* がよく適合します。

newTruck (中心レコード 159) の場合、最近隣は Nissan Quest (105)、Mercury Villager (92) および Mercedes M-Class (101) となります。

前述のとおり、従来の意味ではこれらは必ずしもトラックではありませんが、自動車として分類されていない車両です。最近隣のテーブル・ノード出力を見ると、*newTruck* が比較的価格が高く、またこの種類で最重量の車両の 1 つであることが分かります。ただし、この場合でも、燃料効率が最も近い競合他社の製品より優れているため、利点の 1 つに数えられます。

要約

最近隣分析をどのように使用して、特定のデータ・セットのケースの一連の幅広い特徴を比較するのかについて説明しました。また、2 つのまったく異なるホールドアウト・レコードについて、これらのホールドアウトに最も似ているケースの計算も行いました。

特記事項

本書は IBM が世界各国で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。

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

以下の保証は、国または地域の法律に沿わない場合は、適用されません。 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 Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

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。

この文書に含まれるいかなるパフォーマンス・データも、管理環境下で決定されたものです。そのため、他の操作環境で得られた結果は、異なる可能性があります。一部の測定が、開発レベルのシステムで行われた可能性がありますが、その測定値が、一般に利用可能なシステムのものと同じである保証はありません。さらに、一部の測定値が、推定値である可能性があります。実際の結果は、異なる可能性があります。お客様は、お客様の特定の環境に適したデータを確かめる必要があります。

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 の将来の方向または意向に関する記述については、予告なしに変更または撤回される場合があり、単に目標を示しているものです。

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、名称や住所が類似する企業が実在しているとしても、それは偶然にすぎません。

この情報をソフトコピーでご覧になっている場合は、写真やカラーの図表は表示されない場合があります。

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。

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。

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。

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]

C

D

E

I

M

O

S

W

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