IBM SPSS Modeler 15 アプリケーション ガイド

注 : この情報とサポートされている製品をご使用になる前に、「注意事項」(p.) の一般情報をお読みください。

本版は IBM SPSS Modeler 15 ,および新版で指示されるまで後続するすべてのリリースおよび変更に対して適用されます。

Adobe 製品のスクリーンショットは Adobe Systems Incorporated の許可を得て転載しています。

Microsoft 製品のスクリーンショットは Microsoft 社の許可を得て転載しています。

はじめに

IBM® SPSS® Modeler は、IBM Corp. が開発した企業強化用のデータ マイニング ワークベンチです。SPSS Modeler を使用すると、企業はデータを詳しく調べることにより顧客および一般市民とのリレーションシップを強化することができます。企業は、SPSS Modeler を使って得られた情報に基づいて利益をもたらす顧客を獲得し、抱き合わせ販売の機会を見つけ、新規顧客を引き付け、不正を発見し、リスクを減少させ、政府機関へのサービスの提供を改善することができます。

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

IBM Business Analytics について

IBM Business Analytics ソフトウェアは、意思決定者がビジネス パフォーマンスを向上させるために信頼する完全で、一貫した正確な情報を提供します。ビジネス インテリジェンス、予測分析、財務実績および戦略管理、および分析アプリケーションの包括的なポートフォリオを利用することによって、現在の実績を明確、迅速に理解し、将来の結果を予測することができます。豊富な業界のソリューション、実績ある実例、専門サービスと組み合わせ、さまざまな規模の組織が、高い生産性を実現、意思決定を自信を持って自動化し、より良い決定をもたらします。

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

テクニカル サポート

お客様はテクニカル サポートをご利用いただけます。IBM Corp. 製品の使用方法、または対応するハードウェア環境へのインストールについてサポートが必要な場合は、テクニカル サポートにご連絡ください。テクニカ

ル サポートの詳細は、IBM Corp. Web ページ http://www.ibm.com/support を参照してください。ご本人、組織、サポートの同意を確認できるものをご用意ください。

内容

パート II: データ準備の例

パート III: モデル作成の例

付録

章 1

IBM SPSS Modeler について

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

SPSS Modeler ではさまざまなモデル作成方法を提供しています。[モデル作成] パレットを利用して、データから新しい情報を引き出したり、予測モデルを作成することができます。各手法によって、利点や適した問題の種類が異なります。

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

IBM SPSS Modeler 製品

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

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server の IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler はこの製品のすべての機能を搭載したバージョンであり、コンピュータにインストールし、そのコンピュータで実行します。スタンドアロン製品としてローカル モードで SPSS Modeler を実行するか、大規

模なデータ セットを使用する場合にパフォーマンスを向上させるために IBM® SPSS® Modeler Server と組み合わせて実行することができます。

SPSS Modeler を使用して、プログラミングの必要なく、正確な予測モデルを迅速かつ直感的に構築することができます。独自のビジュアル インターフェイスを使用すると、データ マイニング プロセスを簡単に視覚化することができます。製品に組み込まれている高度な分析の支援を受けて、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、ビジネスチャンスを活用してリスクを軽減できるようになり、それらに影響を与える要因を理解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。 詳細は、 IBM SPSS Modeler エディション in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。

IBM SPSS Modeler Server

SPSS Modeler は、クライアント/サーバー アーキテクチャを使用し、リソース主体の操作が必要な要求を、強力なサーバー ソフトウェアへ分散されるようになりました。その結果、規模が比較的大きいデータ セットを処理するパフォーマンスを実現しました。

SPSS Modeler Server は、1 つまたは複数の IBM® SPSS® Modeler のインストールと組み合わせてサーバー ホストで分散分析モードで継続的に実行する、別途ライセンスが必要な製品です。このように、メモリー集中型の操作は、クライアントコンピュータにデータをダウンロードせずにサーバー上で実行することができるため、SPSS Modeler Server は大きなデータセットに対し優れたパフォーマンスを示すことができます。IBM® SPSS® Modeler Server は、パフォーマンスと自動化のさらなる利点を提供し、SQLの最適化とデータベース内のモデリング機能をサポートしています。

IBM SPSS Modeler Administration Console

Modeler Administration Console は多くの SPSS Modeler Server 設定オプションを管理し、オプション ファイルによって設定可能なグラフィカルアプリケーションです。アプリケーションには、SPSS Modeler Server のインストールを監視、構成するコンソール ユーザー インターフェイスが用意されており、しかも、現在の SPSS Modeler Server のお客様には無料で提供されます。アプリケーションは Windows コンピュータにのみインストールできますが、サポートされる任意のプラットフォームにインストールされたサーバーを管理できます。

IBM SPSS Modeler Batch

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

IBM SPSS Modeler Solution Publisher

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

IBM SPSS Modeler Server の IBM SPSS Collaboration and Deployment Services

さまざまな IBM® SPSS® Collaboration and Deployment Services アダプタを使用すると、SPSS Modeler および SPSS Modeler Server が IBM SPSS Collaboration and Deployment Services リポジトリとインタラクティブに機能させることができます。このように、リポジトリにデプロイされた SPSS Modeler ストリームは、複数のユーザーで共有したり、またはシンクライアントアプリケーション IBM SPSS Modeler Advantage からアクセスできます。リポジトリをホストするシステム上のアダプタをインストールします。

IBM SPSS Modeler エディション

SPSS Modeler は次のエディションで使用できます。

SPSS Modeler Professional

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

SPSS Modeler Premium

SPSS Modeler Premium は、エンティティの分析やソーシャル ネットワーキングなどの特化したデータ、又は構造化されていないテキスト データを処理するために SPSS Modeler Professional を拡張する、別途ライセンスが必要な製品です。SPSS Modeler Premium は次のコンポーネントで構成されています。

IBM® SPSS® Modeler Entity Analytics が新しい次元を IBM® SPSS® Modeler の予測分析に追加します。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、エンティティ分析ではレコードの中でアイデンティティの競合を解決することで現在のデータの干渉性と一貫性を改善することに焦点を当てます。アイデンティティは、個人、組織、オブジェクトまたは曖昧さの存在する他のエンティティとなります。アイデンティティの解決は、顧客関係の管理、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などの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、関係に関する情報を、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。ソーシャル ネットワークの基底となる関係を説明するデータを使用して、IBM® SPSS® Modeler Social Network Analysis はネットワークの他の人の行動に影響を与えるソーシャル リーダーを識別します。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、予測モデルの基準となる個人の包括的なプロフィールを作成できます。この社会的情報を含むモデルは、含まないモデルに比べてパフォーマンスが高くなります。

Text Analytics for IBM® SPSS® Modeler は、高度な言語技術と Natural Language Processing (NLP) を使用して、多様な未構築のテキスト データを急速に処理し、重要なコンセプトを抽出および組織化、そしてそのコンセプトをカテゴリ別に分類します。抽出されたコンセプトとカテゴリを、人口統計のような既存の構造化データと組み合わせ、SPSS Modeler の豊富なデータ マイニング ツールを適用する方法で、焦点を絞ったより良い決定を下すことができます。

IBM SPSS Modeler ドキュメント

オンライン ヘルプ形式のドキュメントは、SPSS Modeler の［ヘルプ］メニューから使用できます。SPSS Modeler、SPSS Modeler Server、および SPSS Modeler Solution Publisher のアプリケーション ガイドやその他サポート資料が含まれています。

各製品の PDF 形式の完全なドキュメント（インストール手順を含む）は、各製品 DVD の ¥Documentation フォルダにもあります。インストール マニュアルは、Web サイト http://www-01.ibm.com/support/docview.wss?uid=swg27023172 からもダウンロードできます。

これらの形式のドキュメントは、SPSS Modeler インフォメーション センター http://publib.boulder.ibm.com/infocenter/spssmodl/v15r0m0/ から入手できます。

SPSS Modeler Professional ドキュメント

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

- **IBM SPSS Modeler ユーザー ガイド:**SPSS Modeler の使用方法への全体的な入門で、データ ストリームの構築方法、欠損地の処理方法、CLEM 式の処理方法、プロジェクトおよびレポートの処理方法、IBM SPSS Collaboration and Deployment Services、予測アプリケーション製品、または IBM SPSS Modeler Advantage へ展開するストリームのパッケージ化方法が含まれています。
- **IBM SPSS Modeler 入力ノード、プロセス ノード、出力ノード:** さまざまな形式のデータを読み込み、処理し、出力するために使用するすべてのノードの説明があります。これは、モデル作成ノード以外のすべてのノードについての説明です。
- **IBM SPSS Modeler モデル作成ノード:** データ マイニング モデルの作成に使用するすべてのノードの説明。IBM® SPSS® Modeler には、マシン学習、人工知能、および統計に基づいたさまざまなモデル作成手法が用意されています。 詳細は、 3 章 モデル作成ノードの概要 in IBM SPSS Modeler 15 モデル作成ノード を参照してください。
- **IBM SPSS Modeler アルゴリズム ガイド:**SPSS Modeler で使用されている手法の数学的な基礎の説明があります。このガイドは、PDF 形式のみです。
- **IBM SPSS Modeler アプリケーション ガイド :** 本ガイドの例では、特定のモデル作成手法および技術に関する簡単で、目的に沿った説明を行います。本ガイドのオンライン バージョンは、［ヘルプ］メニューからも利用できます。 詳細は、 アプリケーションの例 in IBM SPSS Modeler 15 ユーザー ガイド 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:** スクリプトの実行によるシステムのオートメーションの情報で、ノードおよびストリームを操作するために使用することができるプロパティが含まれています。
- **IBM SPSS Modeler 展開ガイド:**SPSS Modeler のストリームやシナリオを IBM® SPSS® Collaboration and Deployment Services Deployment Manager のジョブを処理するステップとしての実行についての情報。
- **IBM SPSS Modeler CLEF 開発者ガイド:**CLEF では、 SPSS Modeler のノードとしてデータ処理ルーチンやモデル作成アルゴリズムなどのサードパーティ製のプログラムを統合します。
- **IBM SPSS Modeler データベース内 マイニング ガイド:** ユーザーのデータベースを最大限に活用して、パフォーマンスを改善する方法と、サードパーティー製のアルゴリズムを使用して分析可能な範囲を拡大する方法についての情報があります。
- **IBM SPSS Modeler Server 管理およびパフォーマンス ガイド**IBM® SPSS® Modeler Server の設定と管理の方法について説明します。
- **IBM SPSS Modeler 管理コンソール ユーザー ガイド:**SPSS Modeler Server を監視して設定するためのコンソール ユーザー インターフェイスのインストールおよび使用に関する情報。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 Solution Publisherガイド:** SPSS Modeler Solution Publisher はアドオン コンポーネントです。組織はこれを使用すると、標準的な SPSS Modeler 環境の外部へストリームを公開できます。
- **IBM SPSS Modeler CRISP-DM Guide:** CRISP-DM 手法を使用して SPSS Modeler によるデータ マイニングを行う段階的なガイドです。
- **IBM SPSS Modeler Batch ユーザー ガイド:** バッチ モードの実行およびコマンドラインの引数の詳細を含む、IBM SPSS Modeler をバッチ モードで使用するための完全ガイド。このガイドは、PDF 形式のみです。

SPSS Modeler Premium ドキュメント

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

- **IBM SPSS Modeler Entity Analytics ユーザー ガイド:** リポジトリのインストールと設定、エンティティ分析ノード、管理タスクについて説明した、SPSS Modeler でのエンティティ分析の使用に関する情報。
- **IBM SPSS Modeler Social Network Analysis ユーザー ガイド:** グループ分析および拡散分析を含む SPSS Modeler によるソーシャル ネットワーク分析を実行するためのガイド。

- **Text Analytics for SPSS Modeler ユーザー ガイド:** 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、その他のリソースについて説明した、SPSS Modeler でのテキスト分析の使用に関する情報。
- **Text Analytics for IBM SPSS Modeler 管理コンソール ユーザー ガイド:**Text Analytics for SPSS Modeler と使用するために IBM® SPSS® Modeler Server を監視して設定するためのコンソール ユーザー インターフェイスのインストールおよび使用に関する情報。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

SPSS Modeler のデータ マイニング ツールは、多様なビジネスおよび組織の問題解決を支援しますが、アプリケーションの例では、特定のモデル作成手法および技術に関する簡単で、目的に沿った説明を行います。ここで使用されるデータセットは、データ マイニング作業者によって管理された巨大なデータ ストアよりも非常に小さいですが、関係するコンセプトや方法は実際のアプリケーションに対して大規模です。

SPSS Modeler の［ヘルプ］メニューから［**アプリケーションの例**］を選択すると、例にアクセスすることができます。データ ファイルとサンプル ストリームは、製品のインストール ディレクトリの Demos フォルダにインストールされています。詳細は、 Demos フォルダ in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。

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

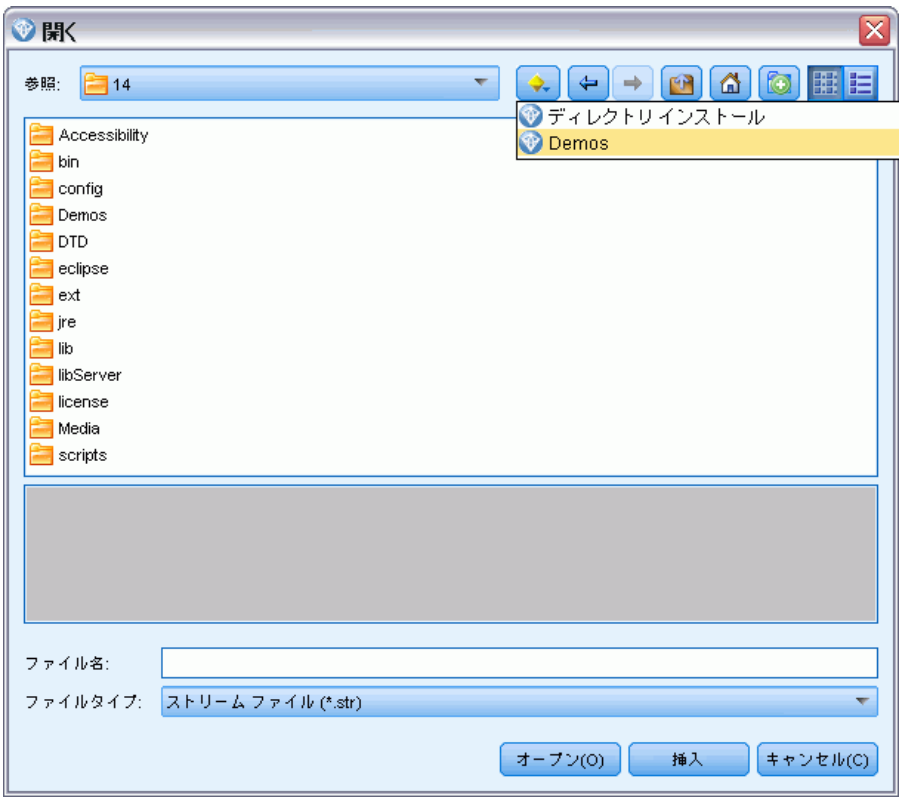
スクリプトの例 : 例は、『IBM SPSS Modeler スクリプトとオートメーション ガイド』を参照してください。

Demos フォルダ

アプリケーションの例で使用されるデータ ファイルとサンプル ストリームは、製品のインストール ディレクトリの Demos フォルダにインストールされています。このフォルダには、Windows［スタート］メニューの IBM SPSS Modeler 15 プログラム グループから、または［ファイルを開く］ダイ

アログ ボックスの最近使ったディレクトリの一覧から［Demos］を選択してアクセスすることもできます。

図 1-1
最近使用されたディレクトリの一覧から Demos フォルダを選択



パート I:
概要およびはじめに

章
2

IBM SPSS Modeler の概要

はじめに

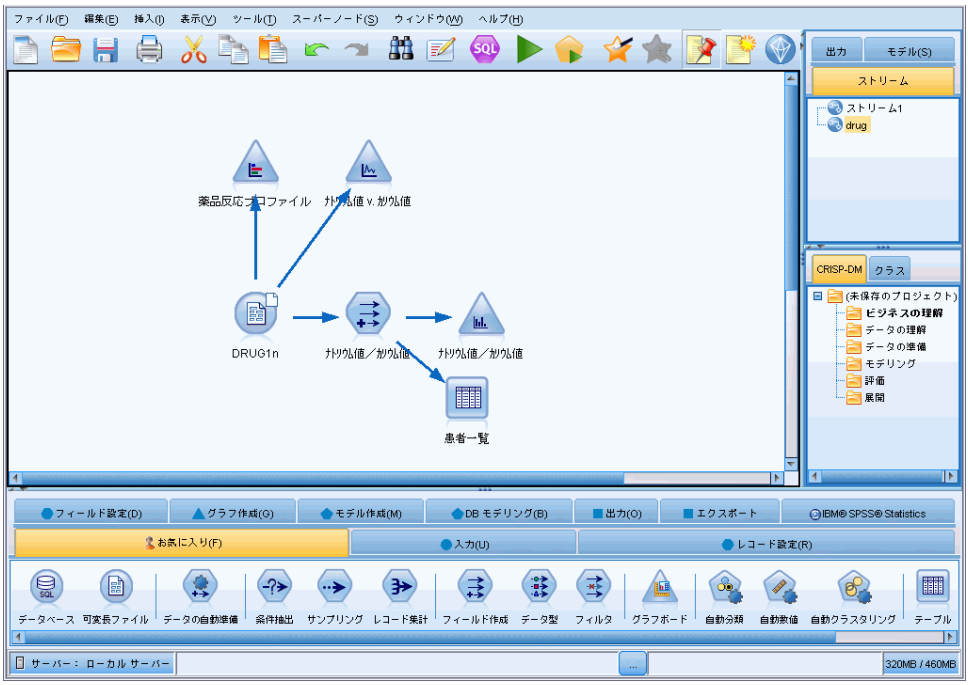
IBM® SPSS® Modeler はデータ マイニング アプリケーションで、大きなデータ セット中の有益なリレーションシップを見つけだすための、戦略的なアプローチ手段を提供しています。従来の統計的な手法とは対照的に、作業開始時に何を見つけだそうとしているのかがわからなくてもかまいません。データの探索、さまざまなモデルの適合、およびさまざまなデータの関連の調査などの作業を行いながら、有益な情報を見つけだしていくことができます。

IBM SPSS Modeler の起動

アプリケーションを起動するには、以下のメニューをクリックします。
[スタート] > [すべてのプログラム] > IBM SPSS Modeler15 > IBM SPSS Modeler15

数秒後にメイン ウィンドウが表示されます。

図 2-1
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、次のようにして IBM® SPSS® Modeler を起動できます。

- IBM® SPSS® Modeler がインストールされているコンピュータで、DOS つまりコマンド プロンプト ウィンドウを開きます。
- SPSS Modeler インターフェイスをインタラクティブ モードで起動するには、**modelerclient** コマンドを入力し、続いてたとえば次のような適切な引数を入力します。

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

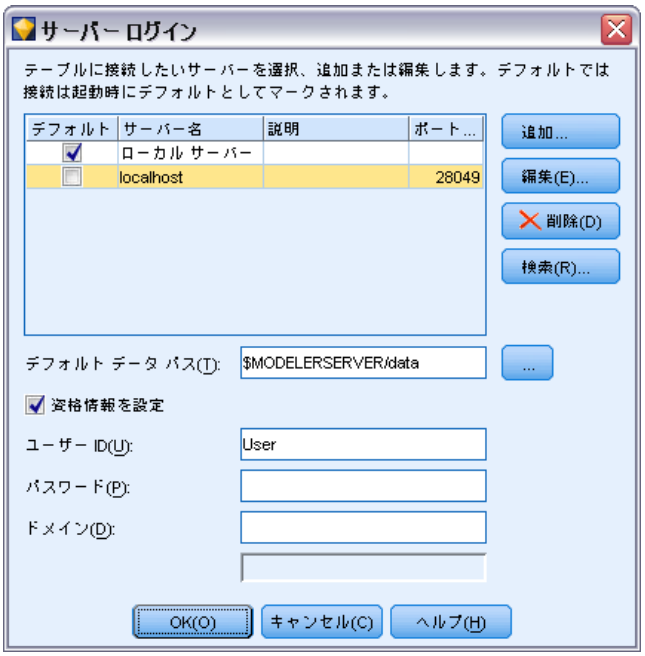
使用可能な引数（フラグ）により、サーバーへの接続、ストリームのロード、スクリプトの実行、または必要に応じて他のパラメータの指定を行うことができます。

IBM SPSS Modeler Server への接続

IBM® SPSS® Modeler は、スタンドアロンのアプリケーションとして、または、IBM® SPSS® Modeler Server に直接または IBM® SPSS® Collaboration and Deployment Services から Coordinator of Processes プラグインを介して SPSS Modeler Server またはサーバー クラスタに接続されたクライアントとして、実行できます。現在の接続ステータスは、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場合は、接続する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できます。ただし、IBM SPSS Collaboration and Deployment Services を使用する場合、[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 クラスタのリストを使用して検索することができます。ネットワーク上で実行する Statistics サービスを介して参照する機能は、Coordinator of Processes で使用できます。 詳細は、 D 付録 サーバー クラスタでの負荷バランシング in IBM SPSS Modeler Server 15 管理およびパフォーマンス ガイド を参照してください。

図 2-2
[サーバー ログイン] ダイアログ ボックス



サーバーに接続するには

▶ [ツール] メニューの **[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、SPSS Modeler ウィンドウの接続ステータス領域をダブル クリックします。

▶ ダイアログ ボックスで、ローカル サーバーのコンピュータ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。詳細は、 IBM SPSS Modeler Server 接続の追加および編集 in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。
- [検索] をクリックして、SPSS COP のサーバーまたはサーバー クラスタにアクセスします。 詳細は、 IBM SPSS Collaboration and Deployment Services のサーバーの検索 in IBM SPSS Modeler 15 ユーザー ガイド 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、定義されたサーバー接続のセットが含まれています。このテーブルには、デフォルト接続、サーバー名、説明、ポート番号が表示されています。既存の接続を選択または検索、あるいは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 コンピュータ上のデータへのパスを指定します。[...] ボタンをクリックして、目的の場所を指定することもできます。

資格情報の設定: このボックスのチェックを解除すると、**シングル サインオン**機能を有効にし、ローカル コンピュータのユーザー名とパスワードの詳細を入力してサーバーにログインします。この でシングル サインオンを使用できない場合、またはこのボックスをチェックしてシングル サインオンを無効にした場合（たとえば、管理者アカウントにログインした場合）、資格情報を入力するための次のフィールドが表示されます。

ユーザー ID: サーバーにログイン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対応するパスワードを入力します。

ドメイン: サーバーにログインする際に使用するドメイン名を指定します。サーバー コンピュータが クライアント コンピュータとは異なる Windows ドメインにある場合にのみ、ドメイン名が必要です。

▶ [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

▶ ［ツール］ メニューの **[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、SPSS Modeler ウィンドウの接続ステータス領域をダブル クリックします。

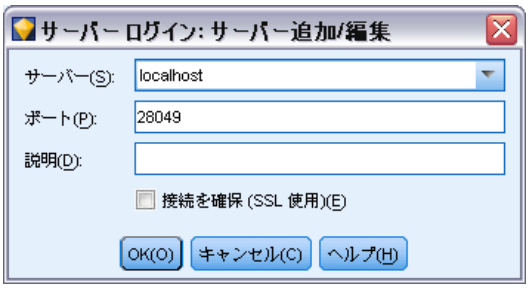
▶ ダイアログ ボックスで、［ローカル サーバー］ を選択し、[OK] をクリックします。

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することができます。［追加］をクリックすると、サーバー接続の詳細を入力できる空の［サーバーの追加/編集］ダイアログ ボックスにアクセス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、その接続を変更することができます。

注 ： IBM® SPSS® Collaboration and Deployment Services から追加されたサーバー接続は、名前、ポート、およびそのほかの詳細が IBM SPSS Collaboration and Deployment Services で定義されているため、編集することができません。

図 2-3
［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



サーバー接続を追加するには

- ［ツール］メニューの **［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。
- 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 ボックスが表示されます。
- サーバー接続の詳細を入力して **［OK］** をクリックします。接続が保存され、［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:** 利用できるサーバーを指定するか、またはリストから選択します。サーバー コンピュータは、英数字の名前（たとえば、myserver）、または、サーバー コンピュータに割り当てられた IP アドレス（たとえば、202.123.456.78）で識別できます。
  - **ポート:** サーバーが待機しているポート番号を入力します。デフォルトのポート番号がうまく動作しない場合は、システム管理者に問い合わせて正しいポート番号を取得してください。

- **説明：** サーバー接続の説明をオプションで入力します。
- **接続を確保 (SSL 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、ネットワークを介してセキュアなデータ送信を行うために一般的に使用されているプロトコルです。この機能を使用するには、IBM® SPSS® Modeler Server をホストするサーバー側で SSL を有効にする必要があります。必要な場合、詳細を各サイトの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

▶ ［ツール］メニューの **[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。

▶ 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 ボックスが表示されます。

▶ サーバー接続の詳細を変更して [OK] をクリックします。変更が保存され、［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 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 (COP) 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 クラスタを選択できます。サーバー クラスタは、 Coordinator of Processes が処理要求に応答するのに最適なサーバーを決定するサーバーのグループです。 詳細は、 D 付録 サーバー クラスタでの負荷バランシング in IBM SPSS Modeler Server 15 管理およびパフォーマンス ガイド を参照してください。

［サーバーへのログイン］ダイアログ ボックスでサーバーを手動で追加できますが、使用できるサーバーを検索して、正しいサーバー名やポート番号を知らなくてもサーバーに接続することができます。この情報は、自動的に提供されます。ただし、ユーザー名、ドメインおよびパスワードなどの、正しいログオン情報が必要です。

注 ： Coordinator of Processes 機能へアクセスしていない場合、接続するサーバー名を手動で入力したり、以前定義した名前を選択することができます。 詳細は、 IBM SPSS Modeler Server 接続の追加および編集 in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。

図 2-4
[サーバーの検索] ダイアログ ボックス

サーバーおよびクラスタを検索するには

- ▶ [ツール] メニューの **[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。
- ▶ ダイアログ ボックスで **[検索]** をクリックすると、[サーバーの検索] ダイアログ ボックスが表示されます。Coordinator of Processes をブラウズしようとする際に IBM SPSS Collaboration and Deployment Services にログオンしていない場合、ログオンを指示するメッセージが表示されます。 詳細は、 9 章 リポジトリへの接続 in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。
- ▶ リストからサーバーまたはサーバー クラスタを選択します。
- ▶ [OK] をクリックしてダイアログ ボックスが閉じられ、選択した接続が [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されます。

一時ディレクトリの変更

IBM® SPSS® Modeler Server が行う処理や操作の中には、一時ファイルを作成する必要があるものもあります。IBM® SPSS® Modeler のデフォルトでは、システムの一時ファイル用ディレクトリに一時ファイルが作成されます。一時ディレクトリの場所を変更するには、次の手順に従ってください。

- ▶ 新規ディレクトリ spss およびそのサブディレクトリ servertemp を作成します。
- ▶ SPSS Modeler のインストール ディレクトリ中の /config にある、options.cfg を編集します。次のファイルの temp_directory パラメータを編集して読み込みます。temp_directory, "C:/spss/servertemp"

- この作業を行った後は、SPSS Modeler Server サーバー サービスを再起動する必要があります。再起動するには、Windows コントロール パネルの［管理ツール］にある **[サービス]** タブをクリックしてください。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、マシンを再起動しても、サービスが再開始されます。

  これで、新しいディレクトリに一時ファイルが作成されるようになります。

  注： この作業を行う際に、スラッシュの種類を間違えて指定することのないように注意してください。SPSS Modeler のように、普通のスラッシュを使用しています。

複数の IBM SPSS Modeler セッションの起動

複数の IBM® SPSS® Modeler を一度に起動する必要がある場合、IBM® SPSS® Modeler および Windows の設定を変更する必要があります。たとえば、2 つの個別のサーバー ライセンスを持ち、同じクライアント マシンから 2 つの異なるサーバーに対して 2 つのストリームを実行する場合に、変更を行う必要があります。

複数の SPSS Modeler セッションを有効化する手順は、次のとおりです。

- 以下のメニューをクリックします。
  [スタート] > [すべてのプログラム] > IBM SPSS Modeler15
- IBM SPSS Modeler15 のショートカット（アイコンで表示）で、右クリックして **[プロパティ]** を選択します。
- **[対象]** テキスト ボックスで、文字列の終わりに **-noshare** を追加します。
- Windows の Explorer で、次の項目を選択します。
  ツール > フォルダ オプション...
- ［ファイル タイプ］タブで、[SPSS Modeler ストリーム］オプションを選択し、**[詳細]** をクリックします。
- ［ファイル タイプの編集］ダイアログ ボックスで、**[SPSS Modeler で開く]** を選択し、**[編集]** をクリックします。
- **[アクションの実行に使用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 インターフェイスについて

IBM® SPSS® Modeler の使いやすいインターフェイスでは、データ マイニング処理の各ポイントで、特定ビジネスの専門知識が必要となります。予測、分類、セグメント化、関連性検出などのモデリング アルゴリズム

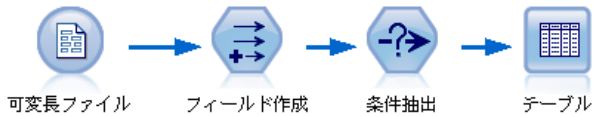
によって、強力かつ正確なモデルが作成されます。モデルの結果は簡単に展開でき、データベース、IBM® SPSS® Statistics、およびさまざまなアプリケーションで使用することができます。

SPSS Modeler での作業は、3 ステップのデータの処理で行われます。

- まず、データを SPSS Modeler に読み込みます。
- そして一連の操作によってデータを実行します。
- 最後に、データを宛て先に送信します。

各操作により、入力から各レコード単位でデータが流れて処理されて、最後に終点（モデルまたはデータ出力タイプ）に到達するため、この操作の流れは**データ ストリーム**と呼ばれます。

図 2-5
単純なストリーム



IBM SPSS Modeler ストリーム領域

ストリーム領域は、IBM® SPSS® Modeler ウィンドウ内で最も広い領域を占めています。ここで、データ ストリームの作成と操作を行います。

ストリームは、業務に関連するデータ操作のダイアグラムをインターフェイスのメイン領域に描画することによって作成します。各操作は**ノード**と呼ばれるアイコンで表されます。ノードは、データの流れと操作を表す**ストリーム**中に配置され、相互に接続されます。

SPSS Modeler では、同じストリーム領域内で、または別のストリーム領域を開いて、複数のストリームに関する作業を同時に行えます。セッションの間、ストリームは SPSS Modeler ウィンドウの右上にあるストリーム マネージャに格納されます。

ノード パレット

IBM® SPSS® Modeler のほとんどのデータおよびモデル作成ツールは、ストリーム領域下のウィンドウの下部にある［**ノード パレット**］にあります。

たとえば、[レコード設定] 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。

ストリーム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、またはアイコンをストリーム領域にドラッグ アンド ドロップします。次にそれらのノードを接続し、データの流れを表す**ストリーム**を作成していきます。

図 2-6
ノード パレットの [レコード設定] タブ

それぞれのパレット タブには、次のようなストリーム操作の各段階で使われる関連ノード群が用意されています。

- **入力**： ノードによって SPSS Modeler にデータが入力されます。
- **レコード設定**： 抽出、結合、および追加などの、データ **レコード**の操作を実行するノードです。
- **フィールド設定**： フィルタリング、新規フィールドの作成、およびフィールドの測定レベルの判断などの、データ **フィールド**に対する操作に用いられます。
- **グラフ**： モデル作成の前後に、データをグラフィカルに表示するために用いられます。グラフには、散布図、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。
- **モデリング。** ニューラル ネットワーク、ディシジョン ツリー、クラスタリング アルゴリズム、およびデータのシーケンス化などの、SPSS Modeler で利用できる優れたモデル作成 アルゴリズムを表すノードです。
- **データベース モデリング**： Microsoft SQL Server、IBM DB2、および Oracle データベースで使用できるモデル作成アルゴリズムを使用します。
- **出力** :SPSS Modeler で表示できるさまざまなデータ、グラフ、モデルの結果の出力を作成します。
- **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します。
- **SPSS Statistics** ： IBM® SPSS® Statistics プロシージャを実行するほか、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。

SPSS Modeler を使用していく過程で、このパレットの内容を自分に合わせてカスタマイズしていくことができます。 詳細は、 12 章 ノード パレットのカスタマイズ in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。

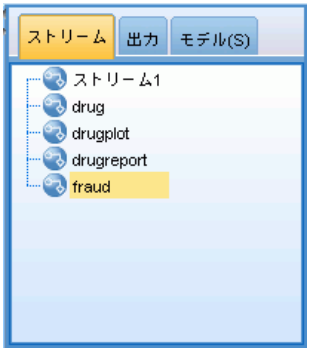
ノード パレットの下部にあるレポート パネルには、データ ストリームにデータを読み込む場合など、各種操作の進行状況に関するフィードバック情報が表示されます。ステータス パネルもノード パレットの下にあります。このウィンドウには、アプリケーションの現在の処理状況や、ユーザーへのフィードバックが必要な場合の指示などが表示されます。

IBM SPSS Modeler マネージャ

ウィンドウの右上がマネージャ パネルです。3 つのタブがあり、ストリーム、出力、モデルの管理を行います。

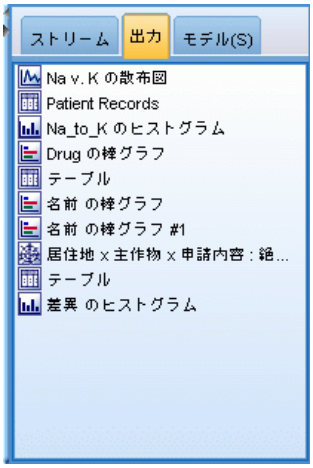
[ストリーム] タブでは、セッション中に作成されたストリームを開いたり、名前を変更したり、保存したり、削除することができます。

図 2-7
[ストリーム] タブ



[出力] タブには、IBM® SPSS® Modeler のストリーム操作で作成されたグラフおよびテーブルなどのさまざまなファイルが表示されます。ここに記載されているテーブル、グラフ、およびレポートを表示したり、名前を変更したり、または閉じることができます。

図 2-8
[出力] タブ



[モデル] タブは、最も強力なマネージャ タブです。このタブにはすべてのモデル **ナゲット**が表示されます。それらは現在のセッションの SPSS Modeler で生成されるモデルです。これらのモデルは、[モデル] タブから直接参照したり、領域中のストリームに追加することができます。

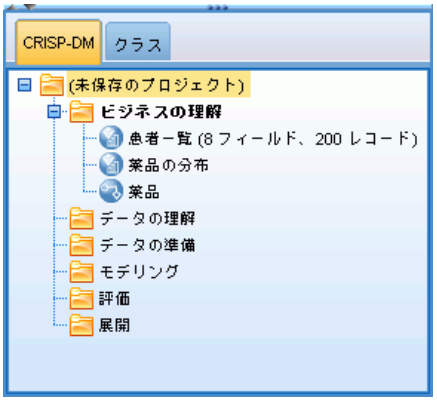
図 2-9
モデル ナゲットを含む [モデル] タブ

IBM SPSS Modeler プロジェクト

ウィンドウの右下には、データ マイニング 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）の作成と管理に使用する**プロジェクト** パネルがあります。IBM® SPSS® Modeler で作成したプロジェクトを表示するには、クラス ビューを使用する方法と、CRISP-DM ビューを使用する方法があります。

［CRISP-DM］タブでは、世界的に知られている方法論で実績のある CRISP-DM（Cross-Industry Standard Process for Data Mining）に基づいてプロジェクトを編成することができます。データ マイニングを熟知している方でも、初めてデータ マイニングを行う方でも、CRISP-DM ツールを使用することにより、より円滑にプロジェクトを編成し、最良の結‰Êを得ることができます。

図 2-10
CRISP-DM ビュー



［クラス］タブでは、SPSS Modeler で行った作業内容を、作成したオブジェクトの種類に応じて編成することができます。このビューは、データ、ストリーム、モデルなどを確認、整理するために役立ちます。

図 2-11
クラス ビュー

IBM SPSS Modeler ツールバー

IBM® SPSS® Modeler ウィンドウの上部には、アイコンが配置されたツールバーがあります。このツールバーは、役に立つさまざまな機能を提供しています。ツールバー ボタンとその機能を次に示します。


新規ストリームの作成


ストリームを開く


ストリームの保存


現在のストリームの印刷


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


貼り付けの選択


最後の操作を元に戻す


やり直し(R)


ノードの検索


ストリームのプロパティを編集

SQL 生成をプレビュー


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にだけ利用可能）


選択したノードをスーパーノードにカプセル化


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）


ストリーム中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 マークアップの非表示（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枝の表示で構成されています。

ストリーム コメントの詳細は、「ノードおよびストリームへのコメントおよび注釈の追加」（ p. ）を参照してください。

スコアリング枝表示の詳細は、「スコアリング枝」（ p. ）を参照してください。

モデルのリンクについては、『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
- アイコンのツールヒントを使用できるかどうか
- 大きいアイコンまたは小さいアイコンのどちらを使用するか

ツールバー表示をオンまたはオフにするには

▶ 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

ツールヒントまたはアイコンのサイズ設定を変更するには

▶ 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 > カスタマイズ

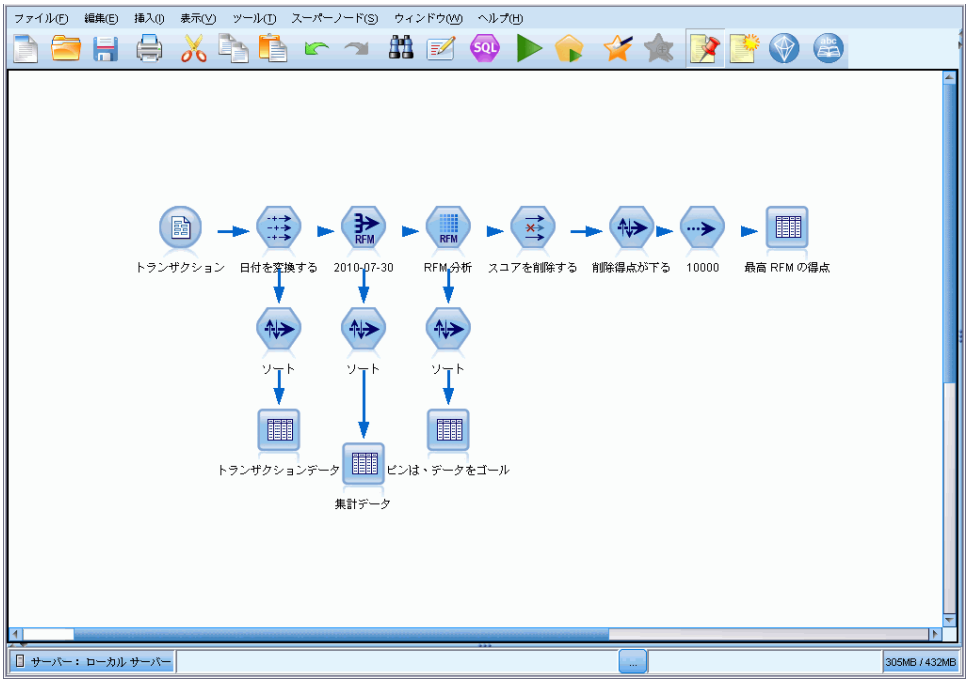
必要に応じて、[**ツールヒントを表示**] または [**大きいボタン**] をクリックします。

IBM SPSS Modeler ウィンドウのカスタマイズ

IBM® SPSS® Modeler インターフェイスのさまざまな構成要素の間にあるディバイダを使って、そのサイズを変更したり、ツールを閉じることができます。たとえば、大きいストリームを作成して作業を行う場合などに、各ディバイダにある小さい矢印をクリックしてノード パレット、マネージャ パネル、およびプロジェクト パネルを閉じることができます。これらのウィンドウを閉じることによってストリーム領域が拡がるため、大きいストリームや複数のストリームで作業を行うために十分な領域を確保することができます。

または、[表示] メニューで、[**ノード パレット**]、[**マネージャ**] 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 2-12
最大化されたストリーム キャンバス



ノード パレット、マネージャ パネル、およびプロジェクト パネルを閉じる代わりに、SPSS Modeler ウィンドウの横と下にある青いスクロール バーを動かして、ストリーム領域をスクロールすることもできます。

画面マークアップの表示を制御できます。画面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枝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

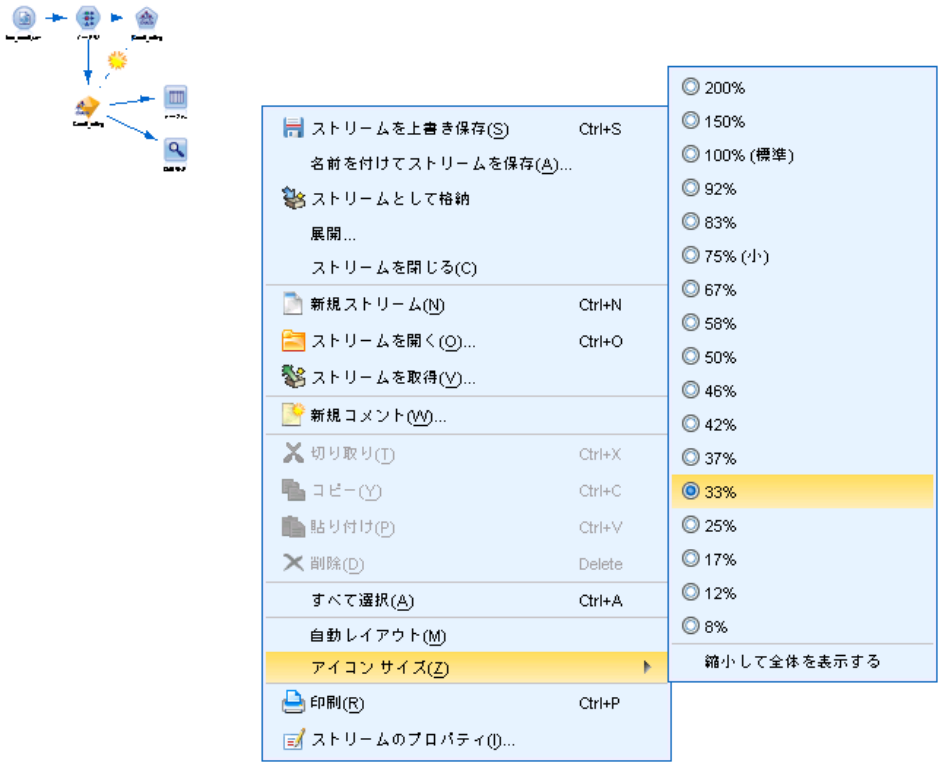
ストリームのアイコン サイズの変更

以下の方法で、ストリームのアイコンのサイズを変更できます。

- ストリーム プロパティの設定から
- ストリームのポップアップ メニューから
- キーボードの使用

ストリーム ビュー全体を標準アイコンサイズの 8% ～ 200% のいずれかのサイズに設定します。

図 2-13
アイコン サイズの変更



ストリーム全体のサイズ変更 (ストリーム プロパティから)

- メイン メニューから次の各項目を選択します。
  ツール > **ストリームのプロパティ** > **オプション** > **レイアウト**:
- [アイコンのサイズ] メニューからサイズを選択します。
- 結果を表示するには [**適用**] をクリックします。
- [OK] をクリックして、変更を保存します。

ストリーム全体のサイズ変更 (メニューから)

- 領域のストリームの背景を右クリックします。
- [**アイコン サイズ**] を選択し、該当するサイズを選択します。

ストリーム全体のサイズ変更 (キーボードを使用)

- Ctrl + [-] キーを押すと、ズームアウトします。

▶ Ctrl + [+] キーを押すと、ズームインします。

この機能は、複雑なストリームの全体のビューを表示するのに適しています。ストリームの印刷に必要なページ数を最小限にすることもできます。

IBM SPSS Modeler でのマウスの使用方法

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たままマウスを動かしてドラッグすれば、ノードを移動できます。
- **ダブルクリック**: マウスの左ボタンをダブルクリックすることにより、ストリーム領域にノードを配置したり、既存のノードを編集します。
- **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、ストリーム領域中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。中央ボタンがないマウスを使用している場合は、代わりに Alt キーを押しながらクリックしたり、ドラッグしてください。

ショートカット キーの使用

大部分の IBM® SPSS® Modeler ビジュアル プログラミング操作には、対応するショートカット キーが用意されています。たとえば、ノードをクリックしてキーボードの Del 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 キーを押したまま 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 と他のキー名で示されています。

Ctrl+X（切り取り）のように、標準の Windows 操作で使われているショートカット キーも数多くあります。SPSS Modeler では、後述するアプリケーション独自のショートカット キーのほかに、これらの標準のショートカット キーを利用することもできます。

注 : 一部、SPSS Modeler で使わ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 キーと重複している場合があります。これらの古いショートカット キーを使用する場合は、Alt キーも一緒に押しながら使用してください。たとえば、Ctrl+Alt+C キーを使用して、キャッシュをオンとオフに切り換えることができます。

テーブル 2-1
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	すべてを選択する
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	プリント
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 と切り替わる）
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	選択したストリーム領域上の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

テーブル 2-2
古いホットキーに対応するショートカット キー

ショートカット キー	関数
Ctrl+Alt+D	ノードの複製
Ctrl+Alt+L	ノードのロード
Ctrl+Alt+R	ノード名の変更
Ctrl+Alt+U	ユーザー入力ノードの生成
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
Del	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。

- ストリームのダイアグラム
- グラフ作成

- ［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）
- スクリプト（[ストリームのプロパティ]、[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）
- モデル（モデル ブラウザ、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）
- 注釈（出力の［注釈］タブを使って）

オブジェクトを印刷するには、次のようにします。

- プレビューを行わないで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。
- 印刷前に印刷の設定を行うには、［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行う場合は、［ファイル］メニューの **[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、選択されているプリンタのオプションを表示して、さまざまなオプションを設定するには、［ファイル］メニューの **[印刷]** を選択します。

IBM SPSS Modeler 自動化

高度なデータ マイニング作業は複雑で長期間になることもあるため、IBM® SPSS® Modeler にはさまざまな種類のコーディングや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、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強力な言語です。CLEM を使用すれば、経費と収入データから利益を算出するような簡単な操作から、Web ログ データを有益な情報を含む一連のフィールドやレコードに変換するような複雑な操作まで、さまざまなストリーム操作を行うことができます。詳細は、 7 章 CLEM について in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。
- **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じような操作を実行できます。CLEM のサブセットを使用して、ノードにオプションを設定し、フィールドを作成することができます。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。 詳細は、 2 章 スクリプトの概要 in IBM SPSS Modeler 15 スクリプトとオートメーション ガイド を参照してください。

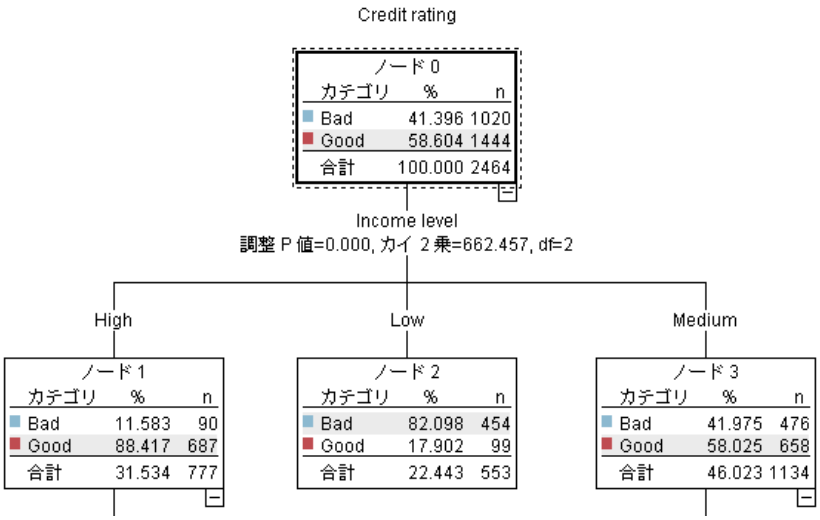
章
3

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。たとえば、金融機関はモデルを使用して、 過去の申請者に関して既に認識されている情報に基づき、融資申請者のリスクが低いか高いかを予測します。

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

図 3-1
簡単なディシジョン ツリー モデル



この例では、次のような一連のデシジョン ルールを使用して、レコードの分類（およびレスポンスの予測）を行う**デシジョン ツリー** モデルを使用します。

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

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

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

フィールド名	説明
Credit_rating	信用度:0=悪い、1=良い、9=欠損値
Age（年齢）	年齢
収入	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（信用度 = 良い）否か（信用度 = 悪い）ということを含めて、銀行から融資を受けている顧客に関する履歴情報のデータベースを管理します。この既存データを使用して、銀行は今後の融資申請者が債務不履行となる可能性がどれほど高いかを予測できるモデルを構築します。

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

この例では、streams サブフォルダの下の Demos フォルダ内にある modelingintro.str という名前のストリームを使用します。データ ファイルは、tree_credit.sav です。詳細は、 1 章 Demos フォルダ in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。

ここで、ストリームを詳しく見ていくことにしましょう。

- メイン メニューから次の各項目を選択します。
  File > ストリームを開く

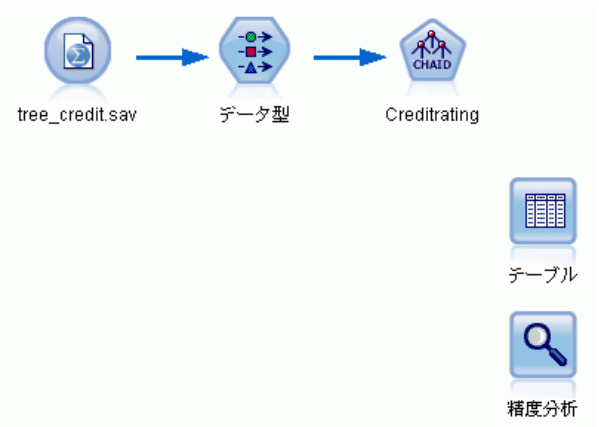
- ［開く］ダイアログ ボックスのツールバーの金のナゲット アイコンをクリックし、Demos フォルダを選択します。

- streams フォルダをダブルクリックします。

- modelingintro.str という名前のファイルをダブルクリックします。

ストリームの構築

図 3-2
モデル作成ストリーム



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

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

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

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

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

Statistics ファイル入力ノードは tree_credit.sav データ ファイルから SPSS Statistics 形式のデータを読み込みます。このデータ ファイルは Demos フォルダにあります(現在の IBM® SPSS® Modeler インストールの下のこのフォルダを参照するには、$CLEO_DEMOS という名前の特別な変

数が使用されます。これによって、現在のインストール フォルダやバージョンにかかわらず、パスが有効になります)。

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

データ型ノードが各フィールドの**測定レベル**を指定します。測定レベルは、フィールドのデータの種類を示すカテゴリです。入力データ ファイルは、3 つの異なる測定レベルを使用します。

連続型フィールド（[年齢] フィールドなど）には連続した数値が含まれるのに対し、**名義型**フィールド（[信用度] フィールド）には [悪い]、[良い]、または [クレジット履歴なし] などの複数の値があります。**順序型**フィー

ルド（[収入レベル] フィールドなど）は、特有の順序を持つ（この場合は [低]、[中] および [高]）複数の値を含むデータについて説明します。

図 3-4
データ型ノードによる対象フィールドおよび入力フィールドの設定

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

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

CHAID モデル作成ノードはモデルを生成します。

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

▶ ［作成オプション］タブをクリックします。

図 3-5
CHAID モデル作成ノードの［フィールド］タブ

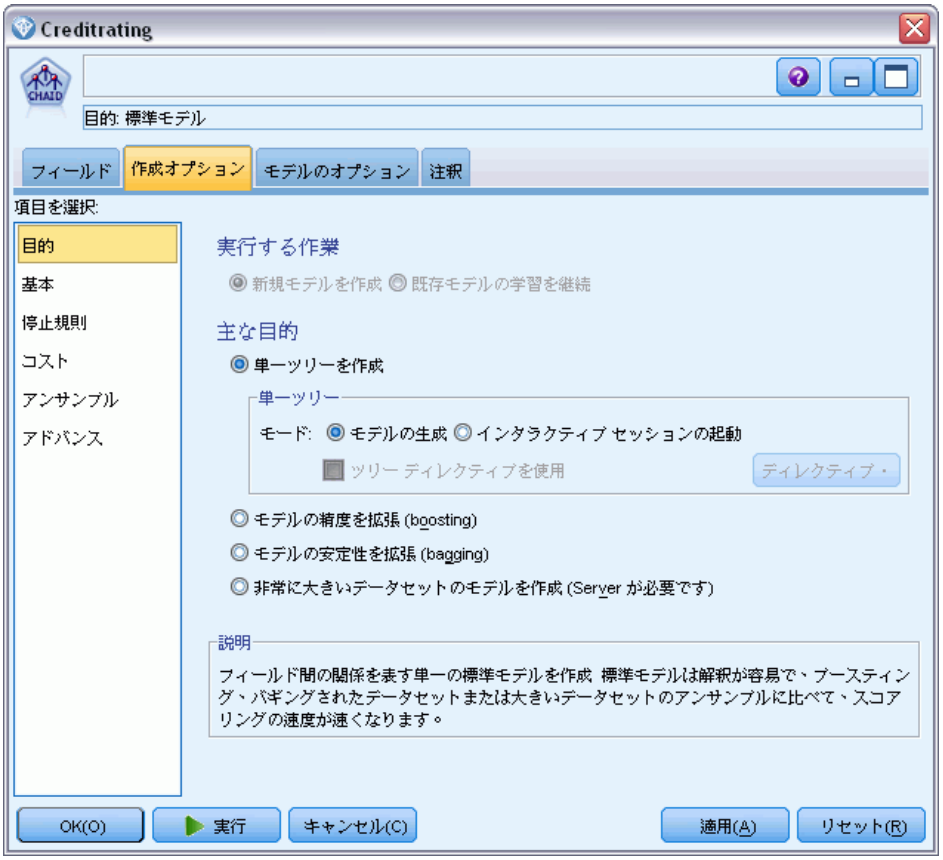
作成するモデルの種類を指定できるオプションがいくつかあります。

新規モデルが必要であるため、**［新規モデルの作成］**のデフォルトオプションを使用します。

また、拡張機能のない単一の標準ディシジョン ツリー モデルが必要であるため、デフォルトの目的オプション**［単一ツリーを作成］**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、モデルの微調整を行うことも可能ですが、この例では、デフォルトの設定 [**モデルの生成**] を使用して単純にモデルを生成します。

図 3-6
CHAID モデル作成ノードの [作成オプション] タブ

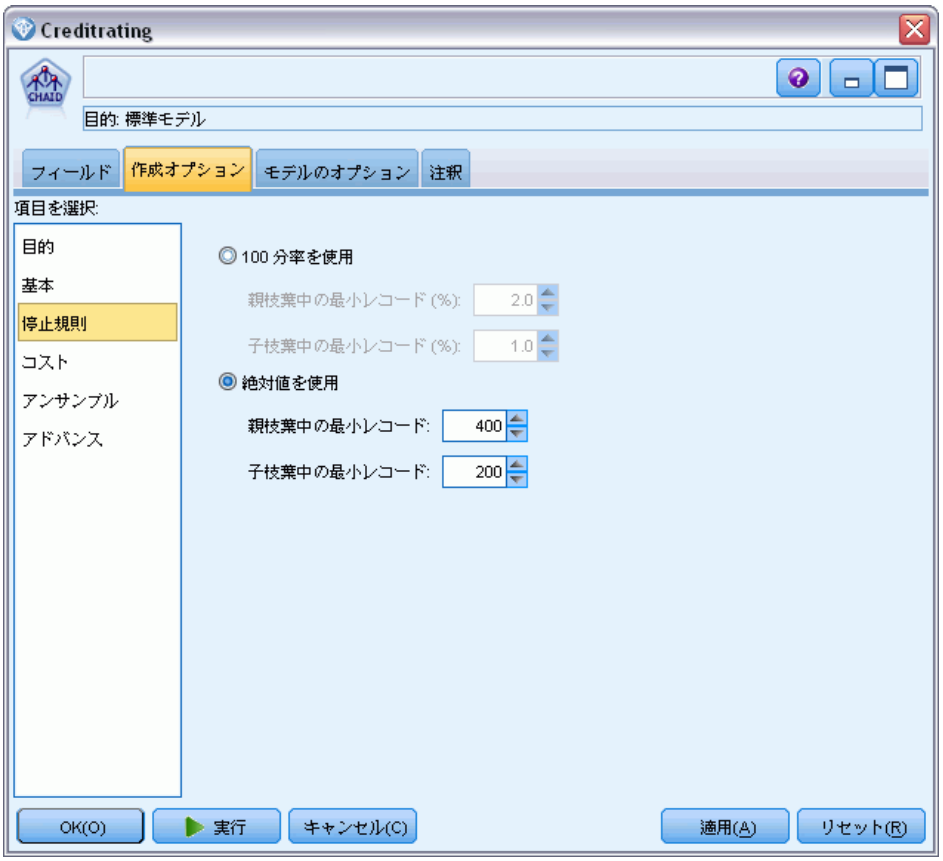


この例では、ツリーを単純にして、親ノードおよび子ノードのケースの最小数を大きくすることにより、ツリーの成長を制限します。

- ▶ [作成オプション] タブで、左側のナビゲータ ペインから [**停止規則**] を選択します。
- ▶ [**絶対値を使用**] オプションを選択します。
- ▶ [**親枝の最小レコード**] を 400 に設定します。

▶ [子枝の最小レコード] を 200 に設定します。

図 3-7
ディシジョン ツリー 構築の停止規則の設定



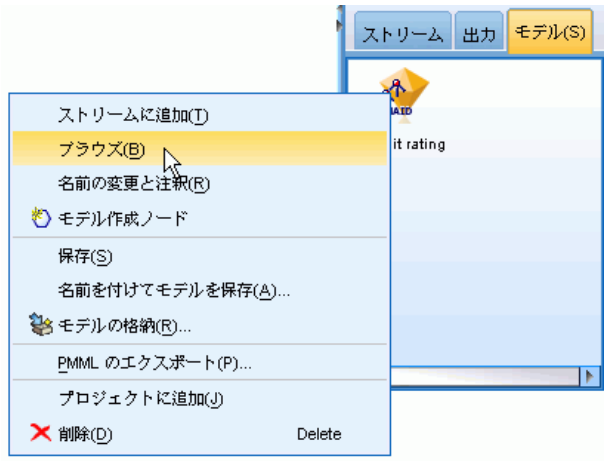
この例では他のすべてのデフォルト オプションを使用できるため、[実行] をクリックしてモデルを作成します(または、ノードを右クリックして、コンテキスト メニューから [実行] を選択するか、あるいは、ノードを選択して［ツール］メニューから [実行] を選択します)。

モデルの参照

実行が完了すると、モデル ナゲットがアプリケーション ウィンドウの右上角のモデル パレットに追加されます。また、モデルが作成されたモデル作成ノードへリンクした状態でストリーム領域内に配置されます。モデル

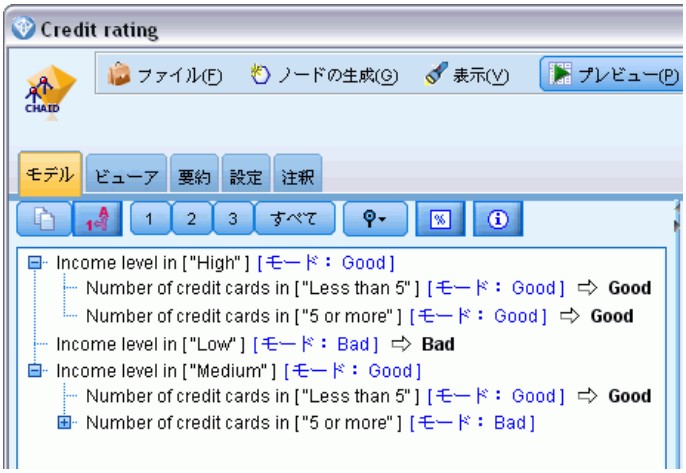
の詳細を表示するには、モデル ナゲットを右クリックして、モデル パレットの［**ブラウズ**］または領域の［**編集**］を選択します。

図 3-8
モデル パレット



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

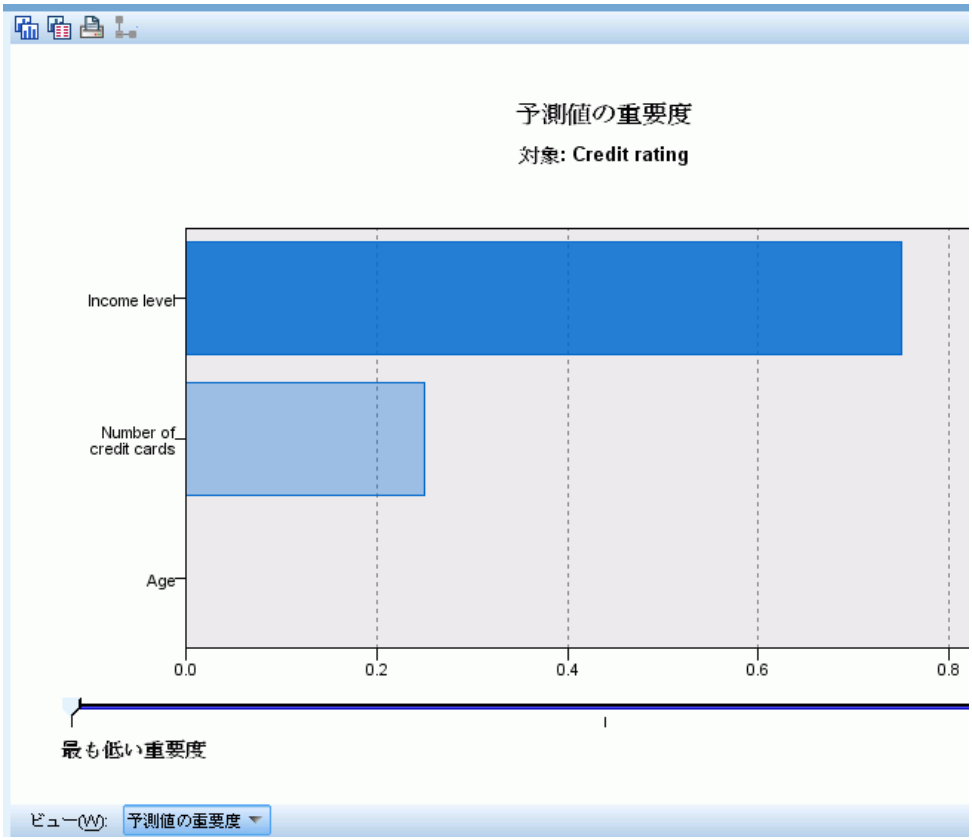
図 3-9
CHAID モデル ナゲット、ルール セット



各ディシジョン ツリー ターミナル ノード、あまり分割していないツリー ノードの場合、［良い］または［悪い］の予測が返されます。どちらの場合も、予測は**モード**、つまり、そのノード内に収まるレコードの最も一般的なレスポンスによって決定されます。

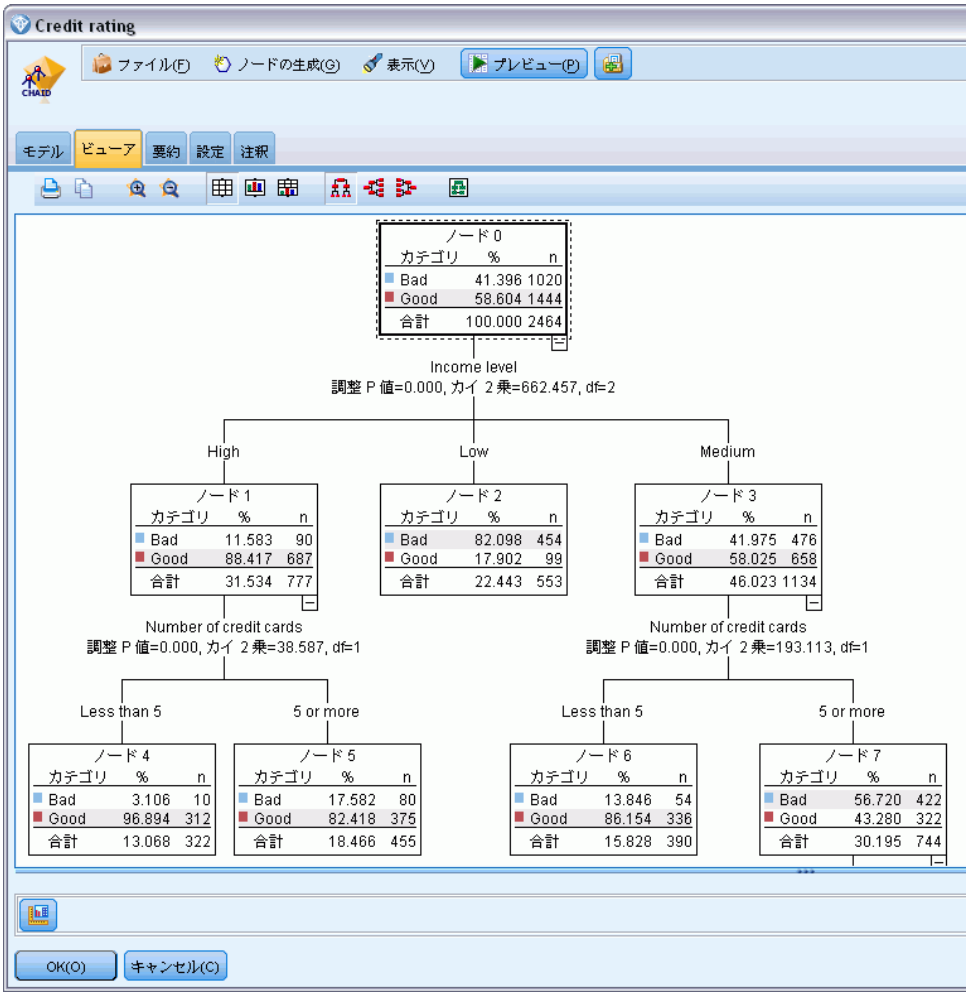
ルール セット右側の、[モデル] タブには予測値の重要度のグラフが表示されます。そのグラフには、モデル推定時の各予測値の相対的な重要度が表示されます。これから、[収入レベル] がこの場合最も有意であり、その他の唯一の有意な因子は [クレジット カード数] であることが分かります。

図 3-10
予測値の重要度グラフ



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

図 3-11
モデル ナゲットの［ビューア］タブ、ズーム アウトを選択



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

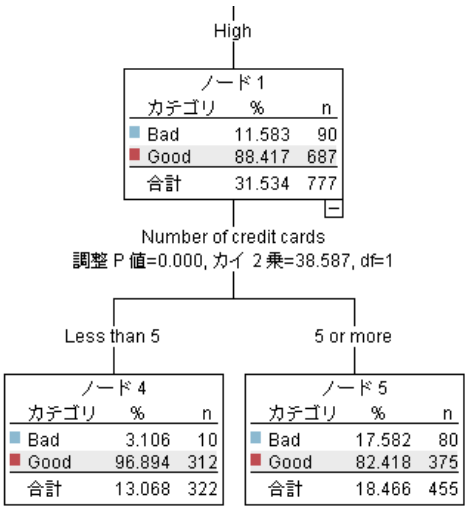
最初の分割は収入レベルを示すことが分かります。収入レベルが［低］カテゴリのレコードはノード 2 に割り当てられます。このカテゴリには高い割合の債務不履行者が含まれます。当然、このカテゴリの顧客に融資することは、高いリスクを有します。

ただし、このカテゴリの 16% の顧客は債務不履行となっておらず、予測が常に正しいとは限りません。すべての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、各レコードに「最も見込みの高い」レスポンスを予測することを可能にします。

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

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（ノード 4 および 5）に分類する方法について注意してください。高収入の顧客について、所有クレジット カード数が 5 枚未満の顧客にのみ融資した場合、成功比率が 89% から 97% まで伸びます。

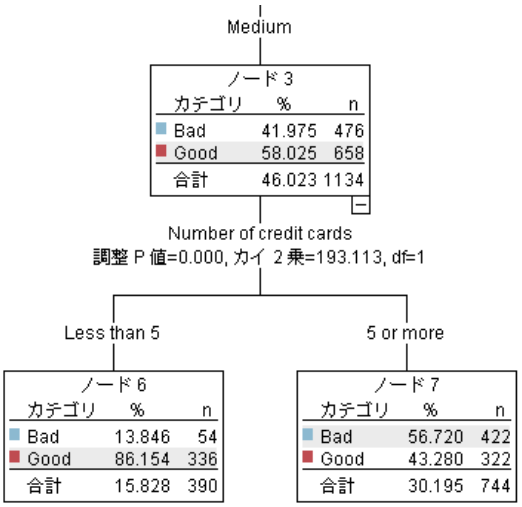
図 3-12
高収入の顧客のツリー ビュー



中程度の収入カテゴリ（ノード 3）の顧客についてはどうでしょう?これらの顧客は［良い］評価と［悪い］評価の間に分類されます。

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

図 3-13
中程度の収入の顧客のツリー ビュー



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

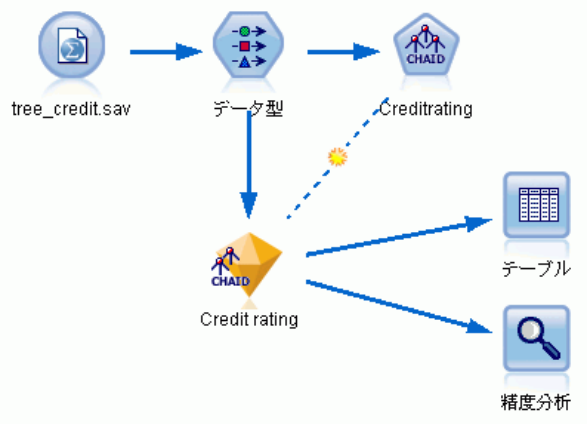
個々のレコードに予測値を割り当てるこのプロセスは、**スコアリング**として知られています。モデルを推定するために使用したのと同じレコードをスコアリングすることにより、モデルが学習データ（結果を知るためのデータ）に対してどれだけ正確に実行できるかを評価できます。その方法について説明します。

モデルの評価

モデルを参照すると、スコアリングが機能する方法を理解できます。ただし、それが「どれほど正確に」機能するかを評価するには、いくつかのレコードのスコアリングを行って、モデルによって予測されたレスポンスと実際の結果とを比較する必要があります。これで、モデルを推定する

のに使用されたのと同じレコードをスコアリングし、観測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 3-14
モデル評価を行うためのモデル ナゲットを出力ノードに接続



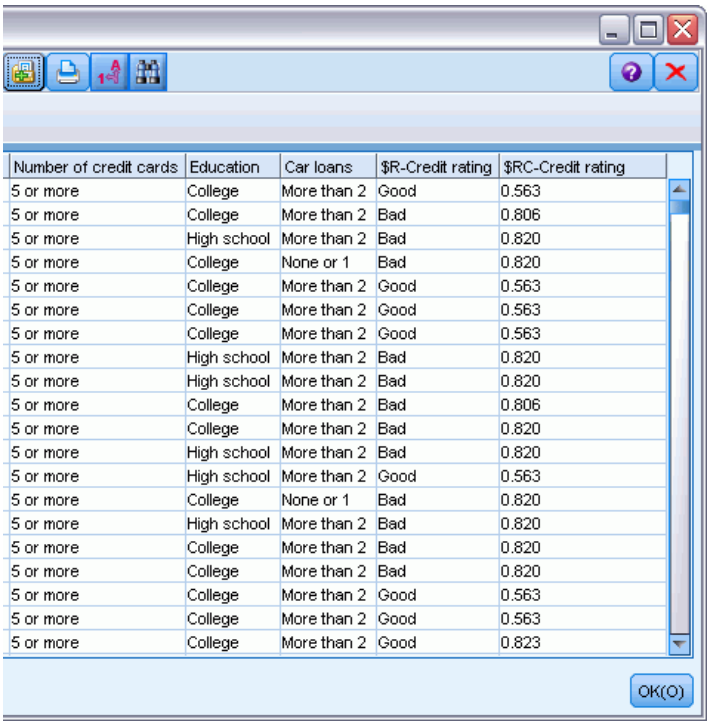
▶ スコアまたは予測値を確認するには、テーブル ノードをダブルクリックして**実行**します。

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

表記法により、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に基づいたもので、予測値には $R-、確信度値には $RC- の標準の接尾辞が付きます。それぞれのモデル タイプでそれぞれの接頭辞を

使用します。**確信度値**は予測値がどれだけ正確であるかに関するモデル独自の推定で、スケールは 0.0 ～ 1.0 です。

図 3-15
生成されたスコアおよび確信度値を示すテーブル



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Bad	0.806
5 or more	High school	More than 2	Bad	0.820
5 or more	College	None or 1	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.806
5 or more	College	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Good	0.563
5 or more	College	None or 1	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.823

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

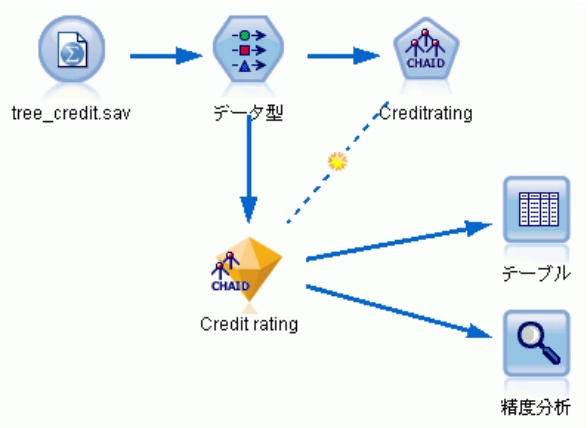
これを回避するには、すべてのノードが純粋に 100%、つまり、すべて良いまたは混在レスポンスのない 悪いになるまで、ツリーを小さい枝に分割し続けます。ただし、そのようなモデルは非常に複雑で、ほかのデータセットに対してうまく一般化できないことが考えられます。

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

▶ モデル ナゲットを分析ノードに接続します。

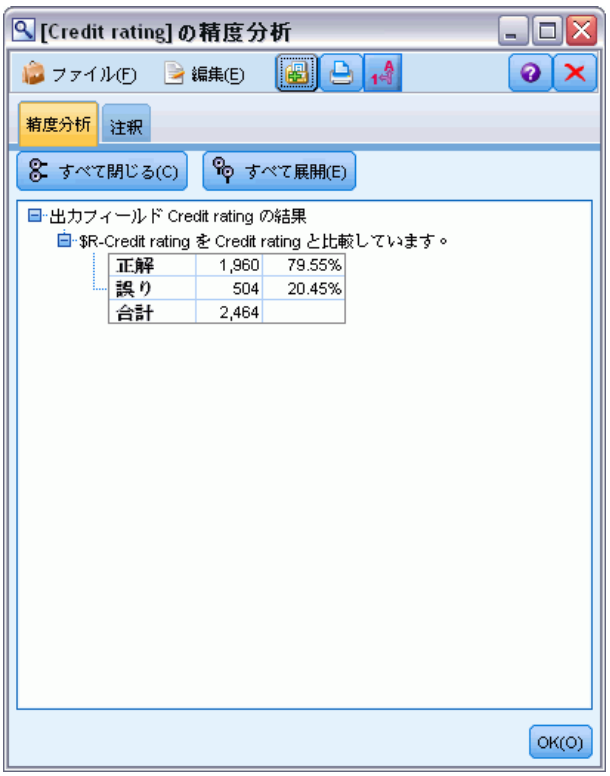
▶ 分析ノードをダブルクリックし、[実行] をクリックします。

図 3-16
分析ノードの接続



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

図 3-17
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、データをサンプルに分割し、学習および評価を行います。

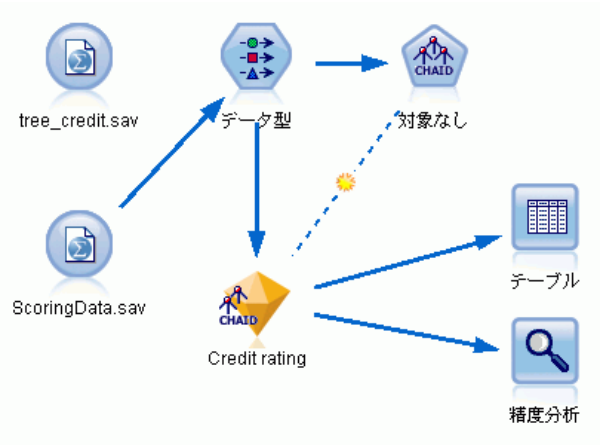
1 つのデータ区分サンプルをモデルの生成に使用し、別のデータ区分サンプルをテストに使用することにより、それが、いかにうまくほかのデータセットを一般化できるかについての良い目安を得ることができます。

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

レコードのスコアリング

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

図 3-18
スコアリング新しいデータの追加



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

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

モデル ナゲットを別のファイルに追加したり、モデルをこの形式をサポートするその他のアプリケーションで使用する PMML フォーマットにエクスポートしたり、IBM® SPSS® Collaboration and Deployment Services リポジトリに格納したりできます。これによって、モデルを全社的に展開し、スコアリングおよび管理できます。

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

要約

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

- モデル作成ノードは、結果がわかっているレコードを調べてモデルを推定し、モデル ナゲットを作成します。これはモデルの学習と呼ばれることもあります。
- モデル ナゲットは、レコードのスコアリングを行う予定のフィールドを含む任意のストリームに追加できます。すでに結果がわかっているレコード（既存の顧客など）をスコアリングすることによって、モデルがどれほどうまく実行されているかを評価できます。
- モデルが適度にうまく実行されていると満足したら、新しいデータ（見込み客など）のスコアリングを行って、その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。

章
4

フラグ型対象の自動化モデル作成

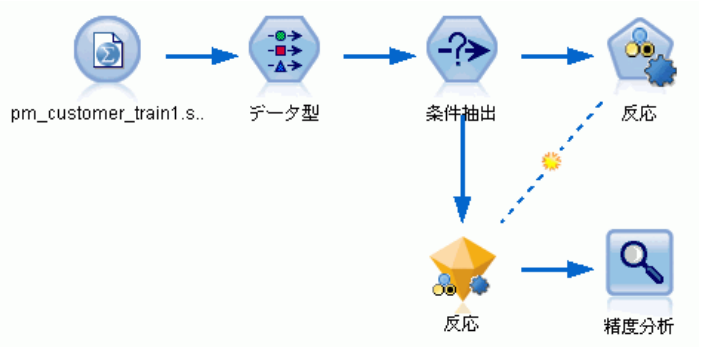
顧客のレスポンスのモデル作成（自動分類）

自動分類ノードでは、ある特定の顧客がローンの支払いを怠りがちになるかどうか、または特定のオファーに回答するかどうかなど、フラグ型（「はい」または「いいえ」）または名義型（セット型）対象のさまざまなモデルを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」）結果を検索します。比較的単純なストリームで、候補モデルを生成およびランク付けし、最善のモデルを選択して、単一の集計済みモデルに結合します。この方法により自動化の容易さと複数のモデルの結合の利点を組み合わせて、モデルから取得できる以上に精度の高い予測を実行できます。

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、さらに収益を上げることを望んでいる金融機関に基づいています。

この方法はでは、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、「5 章」（ p.63 ）を参照してください。

図 4-1
自動分類のサンプル ストリーム



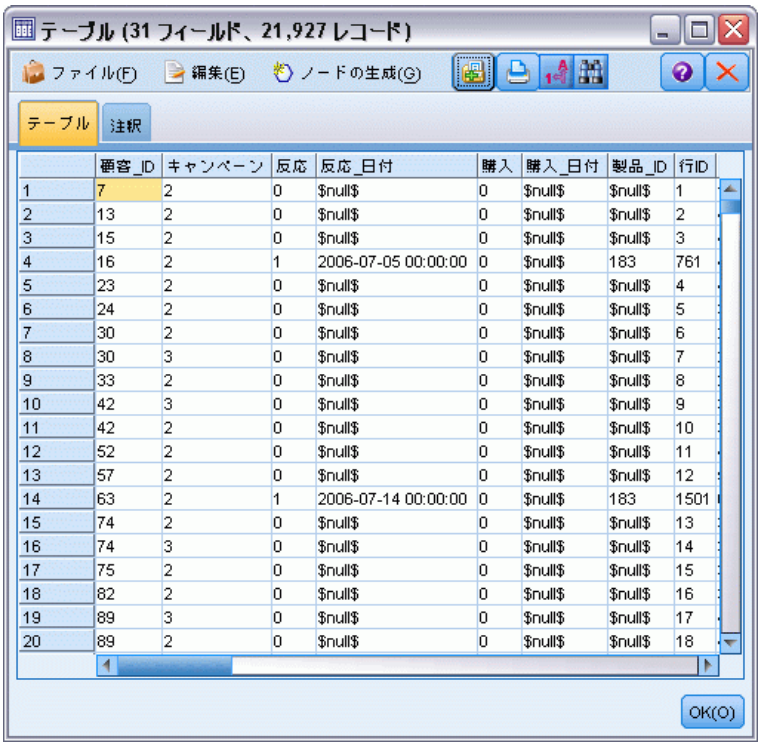
この例では、streams の Demo フォルダにインストールされているストリーム pm_binaryclassifier.str を使用します。使用されるデータ ファイルは、pm_customer_train1.sav です。 詳細は、 1 章 p.7 Demos フォルダ を参照してください。

履歴データ

ファイル pm_customer_train1.sav には、campaign フィールドの値が示すとおり、過去のキャンペーンの特定の顧客に作成したオファーを記録す移籍する履歴データがあります。レコードの最大数は、Premium account キャンペーンに残されます。

campaign フィールドの値は、データ内に整数としてコード化されます（たとえば 2 = Premium account）。後で、より重要な出力を作成するために使用できるよう、これらの値にラベルを定義します。

図 4-2
以前のプロモーションに関するデータ



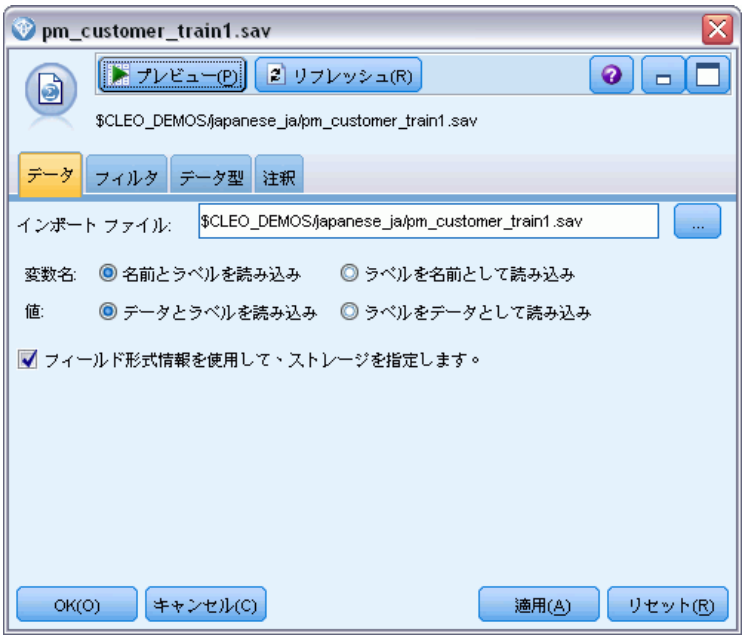
	顧客_ID	キャンペーン	反応	反応_日付	購入	購入_日付	製品_ID	行ID
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	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

また、ファイルには、オファーが受け入れられたかどうか（0 = いいえ、1 = はい）を示す［回答］フィールドが含まれます。これが予測したい **対象フィールド**または値です。 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が含まれています。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回答率を予測するモデルを構築または「学習」することができます。

ストリームの構築

- IBM® SPSS® Modeler インストール フォルダの Demos フォルダにある、pm_customer_train1.sav を示す Statistics ファイル入力ノードを追加します。(ファイル パスの $CLEO_DEMOS/ を、このフォルダを参照するショートカットとして指定することができます。表示されているとおり、パスには円記号ではなくスラッシュを使用する必要があります。)

図 4-3
データの読み込み



▶ 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 = **対象**）。このフィールドの尺度を［**フラグ型**］に設定します。

図 4-4
測定レベルおよび役割の設定

▶ 次のフィールドについては役割を［**なし**］に設定します。customer_id、campaign、response_date、purchase、purchase_date、product_id、Rowid および X_random。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。

▶ データ型ノードの［**値の読み込み**］ボタンをクリックし、値がインスタンス化されていることを確認します。

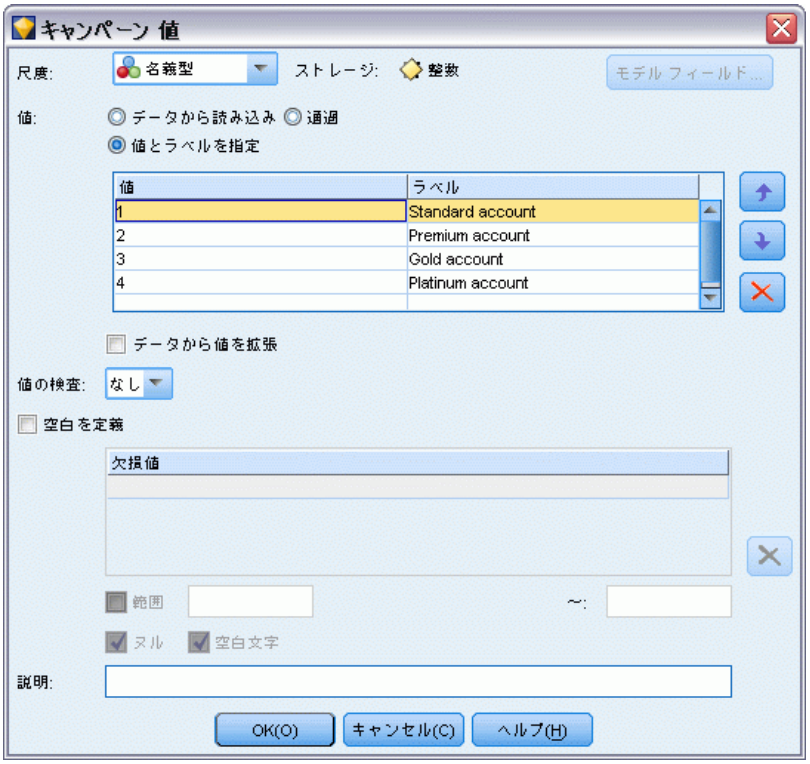
前述のとおり、入力データには、異なる種類の顧客アカウントを対象とした、4 つの異なるキャンペーンに関する情報が含まれています。これらのキャンペーンは、データ内で整数としてコード化されるため、各整数

が示すアカウント タイプを記憶しやすくなります。それぞれについてラベルを定義してください。

図 4-5
フィールドの値を指定

- [キャンペーン] フィールドの行で、[値] 列のエントリをクリックします。
- ドロップダウンリストから [指定] を選択します。

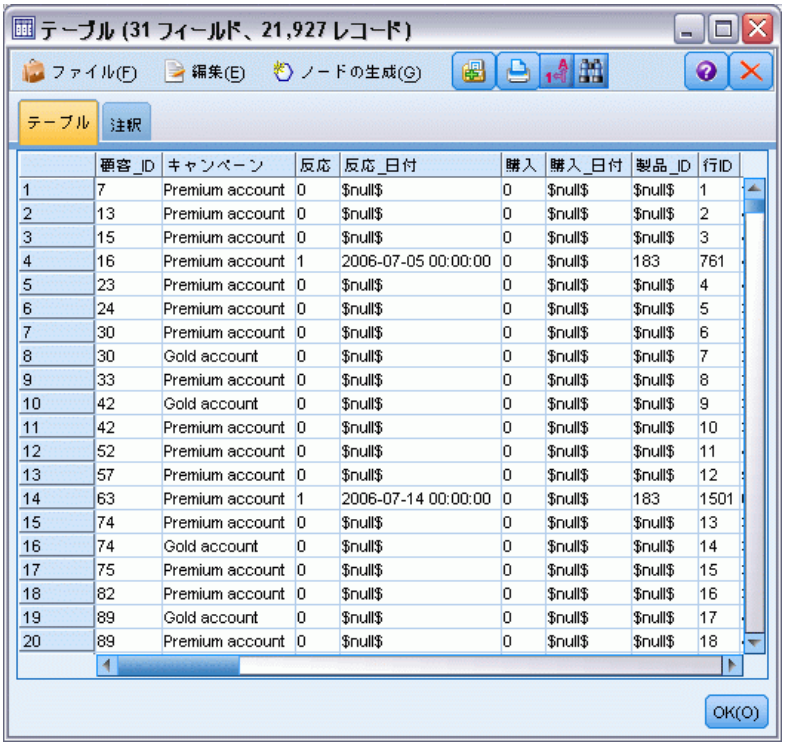
図 4-6
フィールド値のラベルの定義



- [ラベル] 列で、[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たラベルを入力します。
- [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が表示できるようになりました。

図 4-7
フィールド値ラベルの表示



	顧客_ID	キャンペーン	反応	反応_日付	購入	購入_日付	製品_ID	行ID
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	761
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 account	0	$null$	0	$null$	$null$	7
9	33	Premium account	0	$null$	0	$null$	$null$	8
10	42	Gold account	0	$null$	0	$null$	$null$	9
11	42	Premium account	0	$null$	0	$null$	$null$	10
12	52	Premium account	0	$null$	0	$null$	$null$	11
13	57	Premium account	0	$null$	0	$null$	$null$	12
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	Premium account	0	$null$	0	$null$	$null$	13
16	74	Gold account	0	$null$	0	$null$	$null$	14
17	75	Premium account	0	$null$	0	$null$	$null$	15
18	82	Premium account	0	$null$	0	$null$	$null$	16
19	89	Gold account	0	$null$	0	$null$	$null$	17
20	89	Premium account	0	$null$	0	$null$	$null$	18

- テーブルノードをこのデータ型ノードに接続します。
- テーブルノードを開いて、[**実行**] をクリックします。
- 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。
- [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ついての情報が含まれますが、分析は 1 度に 1 つのキャンペーンについて行われます。レコードの最大数は Premium account campaign に残るため（データ内で campaign=2 と

コード化)、条件抽出ノードを使用して、ストリームにこれらのレコードのみを選択することができます。

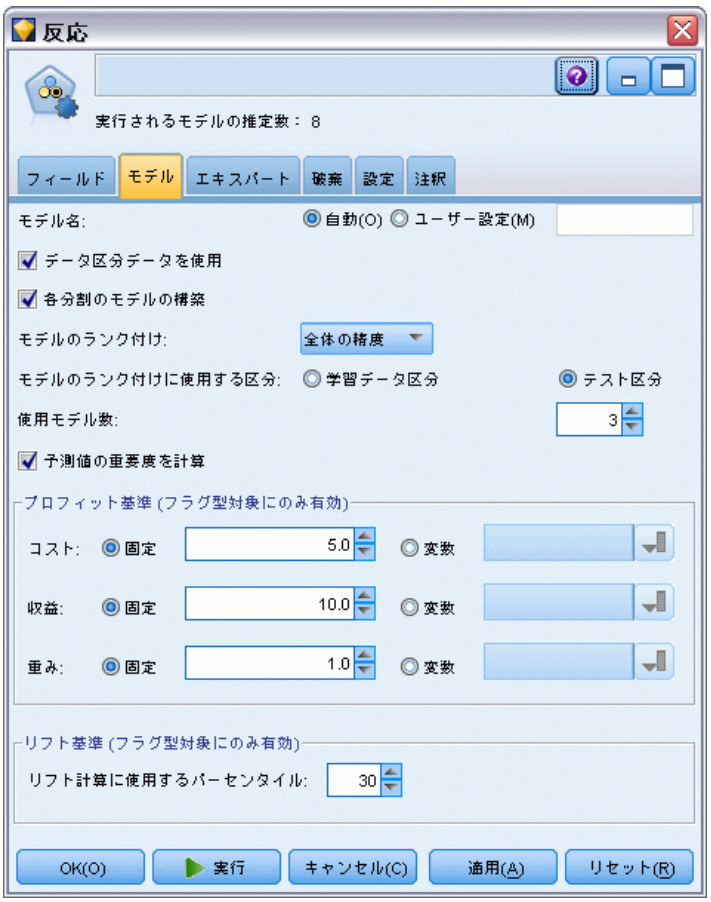
図 4-8
単一キャンペーンのレコードの選択

モデルの生成およびキャンペーン

- 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メトリックとして選択します。

▶ [使用するモデルの数] を 3 に設定します。 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 4-9
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。

▶ [判別分析] および [SVM] モデル タイプの選択を解除します。（これらのモデルはデータの学習に時間がかかるため、選択を解除して例の時間を短縮します。待機してもかまわない場合、選択したままにすることができます。）

［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 4-10
自動分類ノードの［エキスパート］タブ

▶ アンサンブル方法について、［設定］タブで、［**確信度-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。

単純な票決の場合、3 つのモデルのうち 2 つがはいと予測し、はいが 2 対 1 の票決で勝ちます。また、確信度-重み付き票決の場合、各予測の確信度または傾向値に基づいて、票決に重み付けすることができます。結合

された 2 つの［はい］の予測より高い確信度で 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 4-11
自動分類ノード:[設定] タブ

▶ **[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の［モデル］パレットに追加されます。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モデルの詳細が表示されます。(実際の状況では、多くのモデルが大規模なデータセットで作成される場合があるため、数時間かかる場合があります。)図 4-1 p. 49 を参照してください。

個々のモデルをより詳細に検証する場合、**[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、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、サムネイルをダブルクリックすると、フルサイズのグラフを生成できます。

図 4-12
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］タブで選択した尺度であるため、モデルは全体の精度に基づいてソートされます。C51 モデルはこの指標に基づいて最も正確にランク付けしますが、C&R Tree、CHAID モデルもほぼ正確です。

列の見出しをクリックして異なる列のソートを実行する、またはツールバーの [**ソート基準**] ドロップダウン リストの該当する指標を選択することができます。

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。

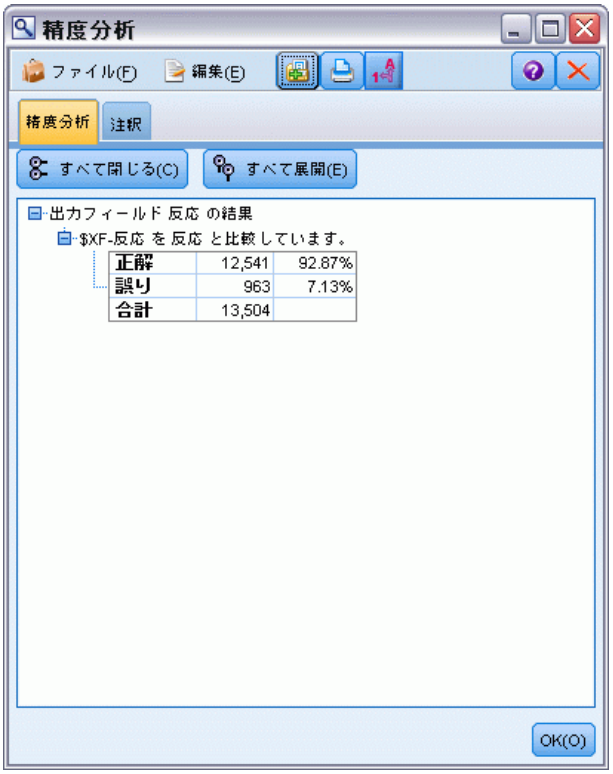
[**使用?**] 列で C51、C&R Tree、CHAID モデルを選択します。

モデル ナゲットの後、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、予測値は実際のレスポンス（元の［レスポンス］フィールド）に 92.82% のケースで一致します。

このケース（C51 の場合 92.86%）で 3 つの各モデルの中で最善のモデルほど正確でありませんが、違いは小さく、それほど重要ではありません。一般的に、学習データ以外のデータセットに適用される場合、アンサンブル モデルのパフォーマンスが良好です。

図 4-13
3 つのアンサンブル モデルの分析



要約

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、結合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。

- 全体の制度に基づいて、C51、C&R Tree、CHAID モデルは学習データで最も良く実行されました。
- アンサンブル モデルのパフォーマンスは各モデルの最善のモデルとほぼ同じくらい良好で、他のデータセットに適用された場合も良いパフォーマンスを示すと考えられます。目的がプロセスの自動化である場合、この方法を使用すると、モデルの仕様を深く掘り下げることなく最良の環境下で強力なモデルを取得することができます。

章 5

連続型対象の自動化モデル作成

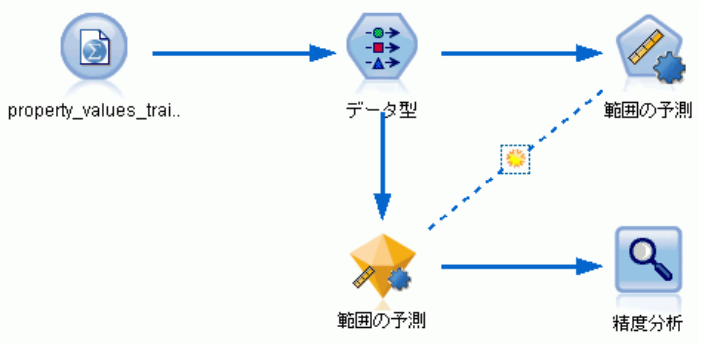
プロパティ値（自動数値）

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）のさまざまなモデルを自動的に作成し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。フラグ型または名義型の対象ではなく連続型ですが、このノードは自動分類ノードと同じ方法で動作します。

ノードを使用して、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 ナゲットに結合することもできます。この方法により自動化の容易さと複数のモデルの結合の利点を組み合わせて、モデルから取得できる以上に精度の高い予測を実行できます。

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

図 5-1
自動数値のサンプル ストリーム



この例では、streams の Demo フォルダ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。使用されるデータ ファイルは、property_values_train.sav です。詳細は、 1 章 p.7 Demos フォルダ 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ます。これが予測する**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	Label
property_id	プロパティ ID
隣接	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
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 フォルダにあります。このファイルにも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。課税対象価格がわ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、この値がまだわ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

- IBM® SPSS® Modeler インストール フォルダの Demos フォルダにある、property_values_train.sav を示す Statistics ファイル入力ノードを追加します。(ファイル パスの $CLEO_DEMOS/ を、このフォルダを参照する

ショートカットとして指定することができます。表示されているとおり、パスには円記号ではなくスラッシュを使用する必要があります。))

図 5-2
データの読み込み

▶ 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 = **対象**）。他のフィールドの役割は**入力**と設定されており、これらが予測フィールドとして使用されることを示しています。

図 5-3
対象フィールドの設定

▶ 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メトリックとして選択します。

▶ [使用するモデルの数] を 3 に設定します。 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 5-4
自動数値ノードの [モデル] タブ

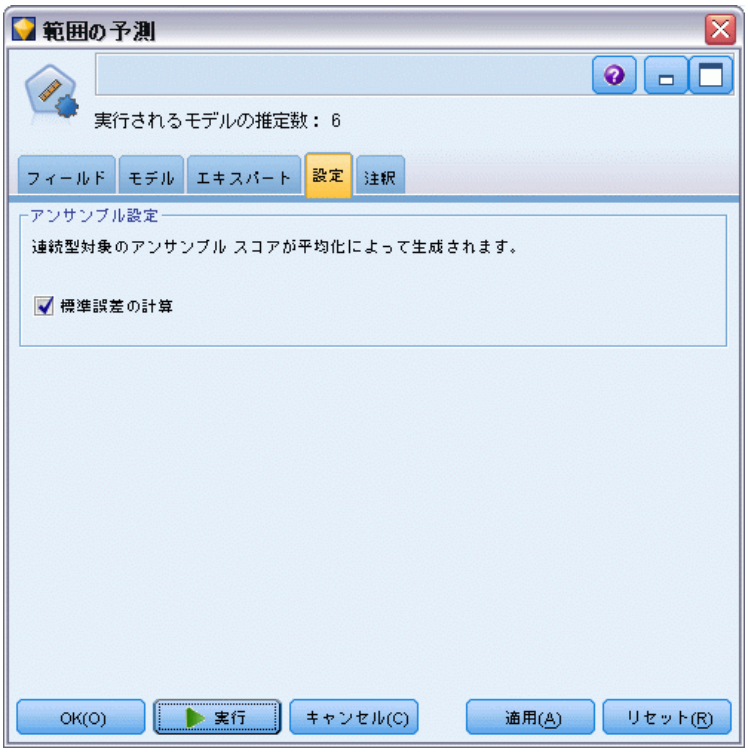
▶ [エキスパート] タブで、デフォルト設定をそのままにします。7 つのモデルの合計について、ノードは各アルゴリズムの 1 つのモデルを推定します。(また、これらの設定を変更して、各モデル タイプの複数のバリアントを比較することもできます。)

［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 5-5
自動数値ノードの［エキスパート］タブ

▶ ［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 5-6
自動数値ノードの［設定］タブ



モデルの比較

▶ ［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の［モデル］パレットに追加されます。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モデルの詳細が表示されます。（実際の状況では、多くのモデルが大規模なデータセットで推定されるため、数時間かかる場合があります。）図 5-1 p. 63 を参照してください。

個々のモデルをより詳細に検証する場合、[**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、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 5-7
自動数値の結果

デフォルトでは、自動数値ノードで選択した指標であるため、モデルは相関に基づいてソートされます。ランク付けのためには、強力な関係を示す 1 に近い値の相関の絶対値が使用されます。一般化線型モデルはこの指標に基づいて最も正確にランク付けしますが、他のモデルもほぼ正確です。一般化線型モデルにも、最低相対エラーがあります。

列の見出しをクリックして異なる列のソートを実行する、またはツールバーの [**ソート基準**] リストの該当する指標を選択することができます。

各モデルの観察値に対する予測値のプロットを表示し、それらの間の相関を迅速に視覚的に表示します。良好なモデルの場合、ポイントは対角線に沿ってクラスタリングします。この例のすべてのモデルに当てはまります。

[**グラフ**] 列で、サムネイルをダブルクリックすると、フルサイズのグラフを生成できます。

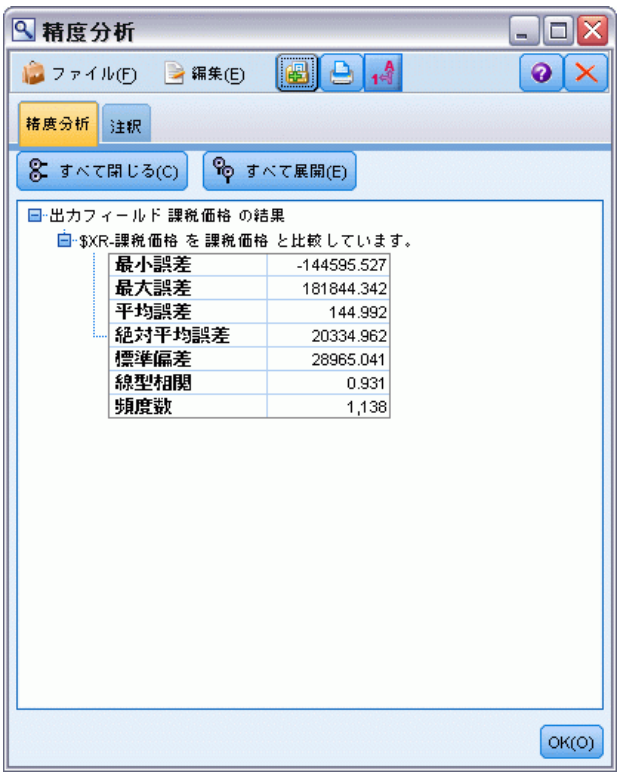
これらの結果に基づいて、3 つの最も正確なモデルをすべて使用するよう指定します。複数モデルの予測を組み合わせることにより、個々のモデルの制限を回避でき、全体の精度がより高くなります。

[**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、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 ノード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 つのモデルのうちのモデルより高い値です。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ります。

図 5-8
自動数値のサンプル ストリーム



要約

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、結合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。

- 全体の制度に基づいて、一般化線型、回帰、CHAID モデルは学習データで最も良く実行されました。
- 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 つのモデルより良好で、他のデータセットに適用された場合も良いパフォーマンスを示すと考えられます。目的がプロセスの自動化である場合、この方法を使用すると、モデルの仕様を深く掘り下げることなく最良の環境下で強力なモデルを取得することができます。

パート II:
データ準備の例

章
6

自動データ準備 (ADP)

分析用のデータ準備は、データ マイニング プロジェクトで最も重要なステップの 1 つで、これまではもっとも時間を要するステップの 1 つでした。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ある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取得、詳細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。または必要に応じて変更の作成および承認または拒否の前に変更をプレビューできます。

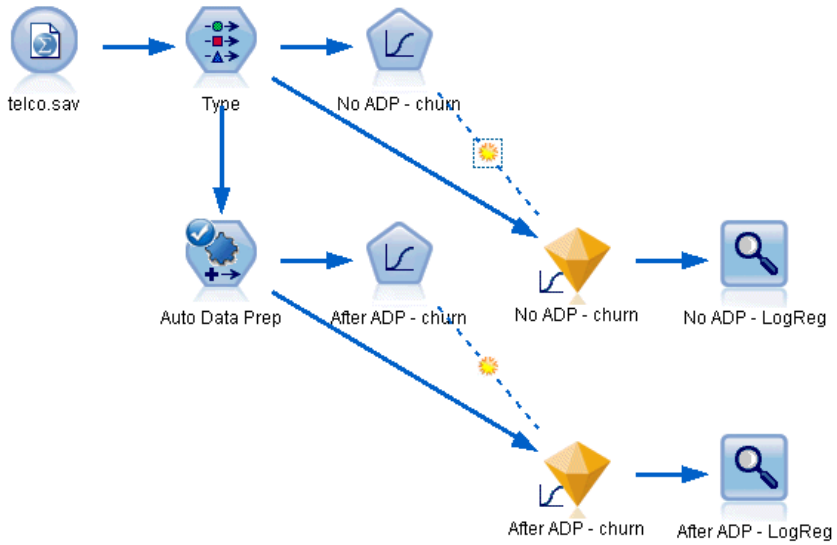
ADP ノードを使用すると、統計コンセプトの以前の情報がなくても、迅速かつ容易にデータ マイニングの準備できます。デフォルトの設定でノードを実行した場合、モデルはより迅速に作成およびスコアリングされます。

この例では、ADP_basic_demo.str というストリームを使用します。このストリームは telco.sav というデータ ファイルを参照し、モデル作成時にデフォルトの ADP ノード設定を使用した場合の精度の向上について示します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、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 ディレクトリにあります。

ストリームの構築

- ストリームを構築するには、IBM® SPSS® Modeler インストール ディレクトリの Demos ディレクトリにある telco.sav を示す Statistics ファイル入力ノードを追加します。

図 6-1
ストリームの構築



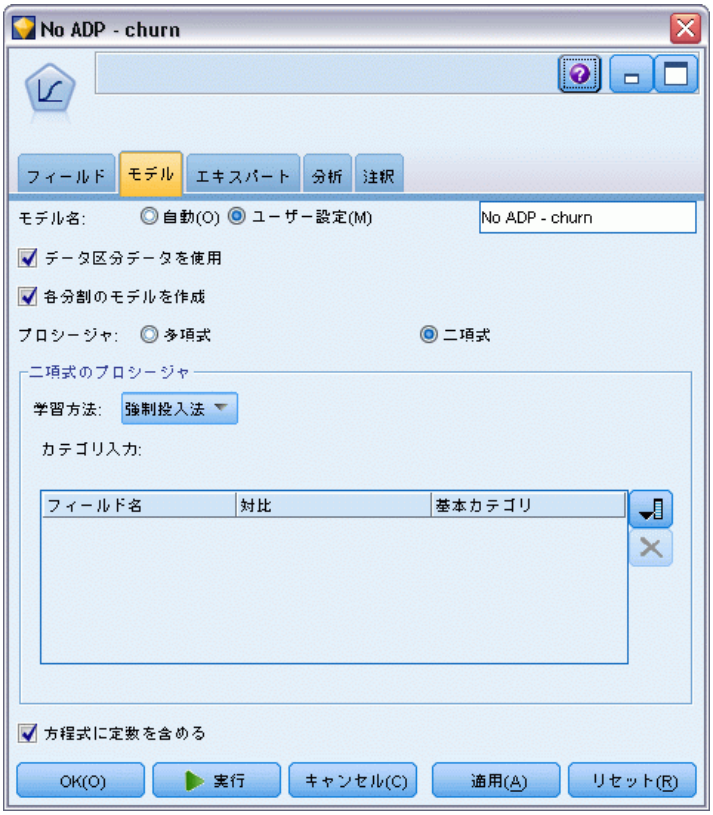
▶ データ型ノードを入力ノードに接続、解約の測定レベルを［**フラグ型**］に設定し、役割を［**対象**］に設定します。その他のフィールドの役割はすべて**入力**に設定します。

図 6-2
対象の選択

▶ ロジスティック ノードをデータ型ノードに接続します。

▶ ロジスティック ノードで、[モデル] タブをクリックし、[**二項検定**] 手続きを選択します。[モデル名] フィールドで、[**顧客**] を選択し、[**ADP なし - 解約**] と入力します。

図 6-3
モデル オプションの選択

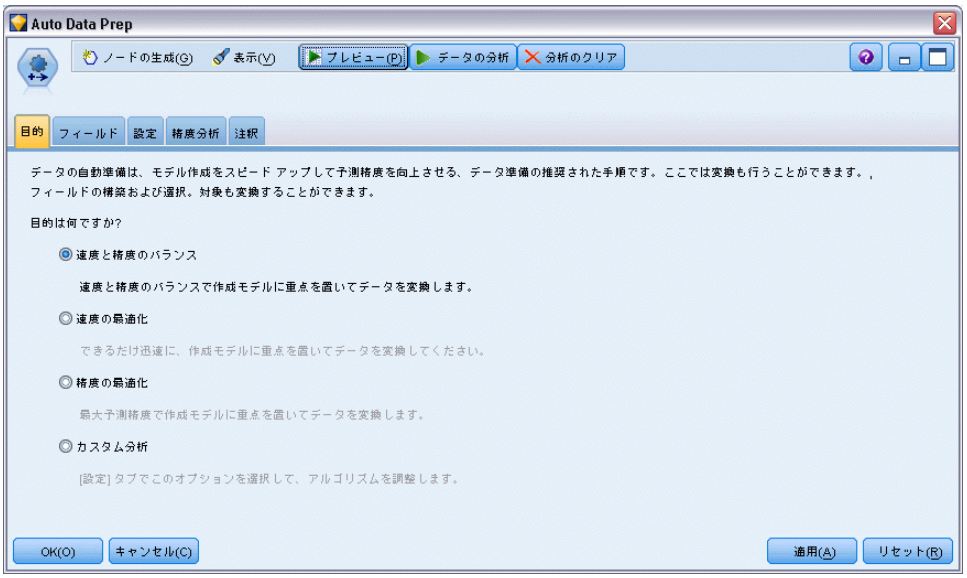


▶ ADP ノードをデータ型ノードに添付します。[目的] タブをデフォルトの設定のままにし、速度および精度のバランスを調整して、データを分析および準備します。

▶ [目的] タブのいちばん上で、[**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。

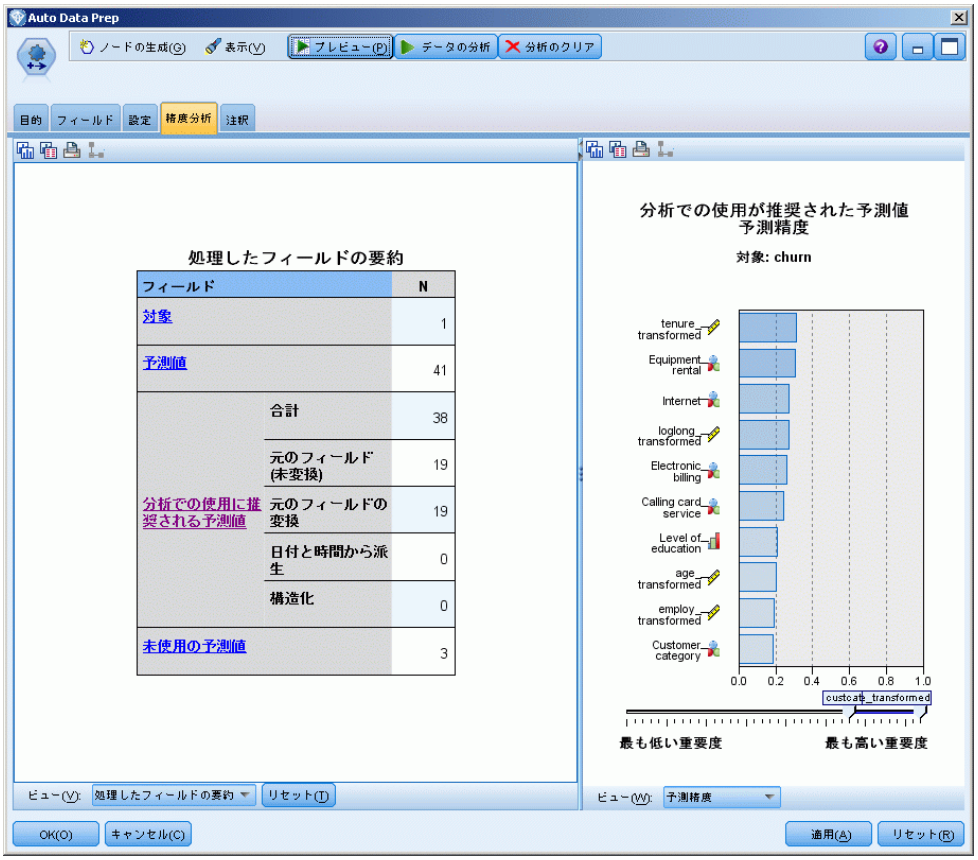
ADP ノードの他のオプションを使用すると、処理する精度により重きを置くか、速度により重きを置くかを指定したり、多くのデータ準備プロセスの手順を調整したりすることができます。

図 6-4
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 フィールドのうち、19 件が変換されて処理され、3 件が未使用のまま破棄されています。

図 6-5
データ処理の要約



▶ ロジスティック ノードを ADP ノードに接続します。

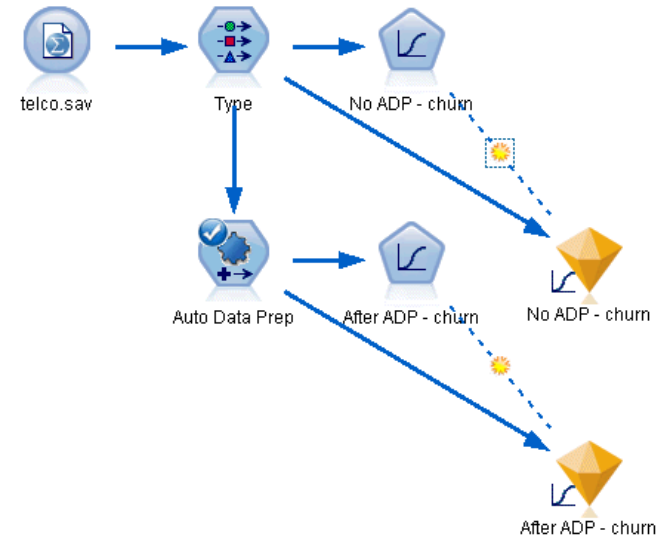
▶ ロジスティック ノードで、［モデル］タブをクリックし、[二項検定］手続きを選択します。［モデル名］フィールドで、[顧客］を選択し、[ADP 後 – 解約］と入力します。

図 6-6
モデル オプションの選択

モデルの精度の比較

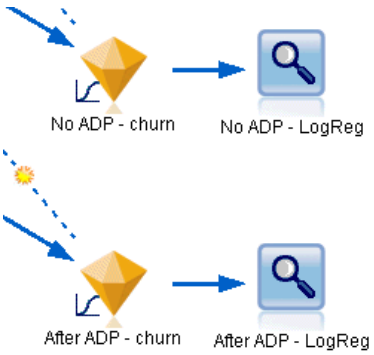
- 両方のロジスティック ノードを実行してモデル ナゲットを作成し、ストリームおよび右上のモデル パレットに追加されます。

図 6-7
モデル ナゲットの接続



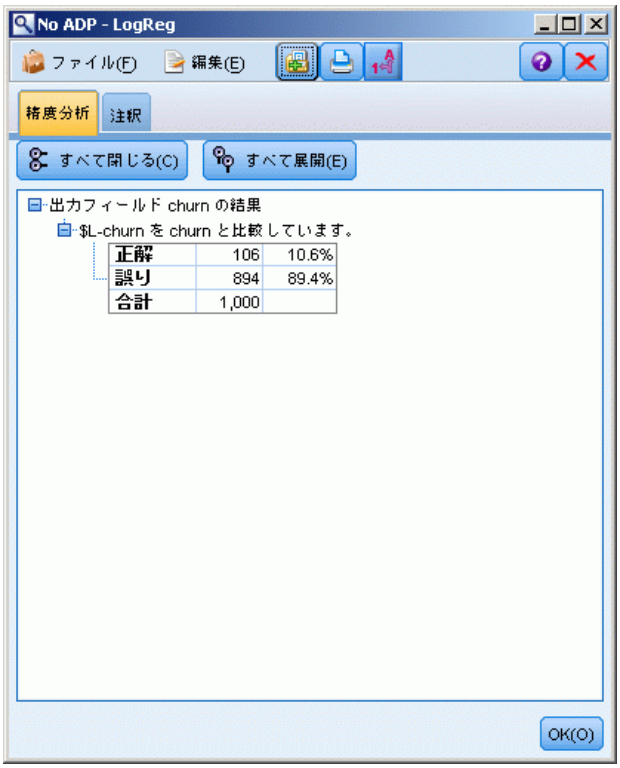
- 精度分析 ノードをモデル ナゲットに接続し、デフォルトの設定を使用して精度分析ノードを実行します。

図 6-8
精度分析ノードの接続



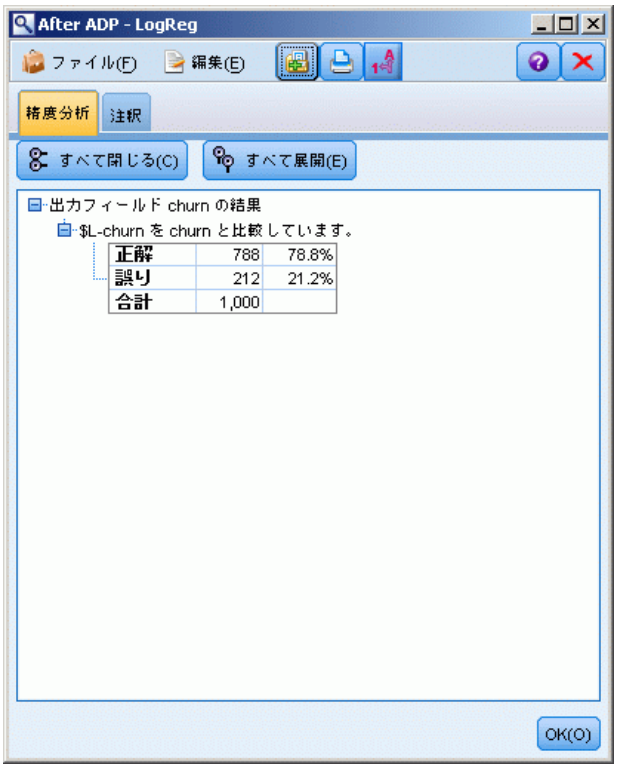
非 ADP 派生モデルの分析では、ロジスティック回帰ノードをデフォルト設定でデータを実行すると、モデルの精度がわずか 10.6% となることがわかります。

図 6-9
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、精度 78.8% のモデルを作成することがわかります。

図 6-10
ADP 派生モデルの結果



ただ ADP ノードを実行してデータの処理を調整することによって、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。

特定の論理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したい場合、モデル設定を直接操作したほうがいい場合もあります。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 を利用すると便利な場合があります。

IBM® SPSS® Modeler で使用されるモデル作成方法の数学的な基礎については、インストール ディスクの ¥Documentation ディレクトリにある『SPSS Modeler アルゴリズム ガイド』で説明されています。

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をどれだけうまく一般化しているかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。 詳細は、 4 章 データ区分ノード

in IBM SPSS Modeler 15 入力ノード、プロセス ノード、出力ノード を参照してください。

章 7

分析用のデータの準備（データ検査）

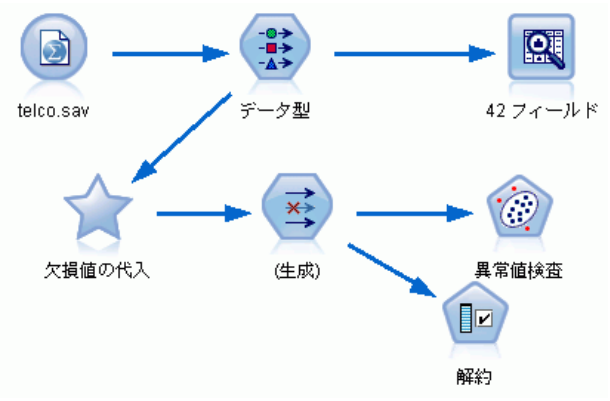
データ検査ノードは、IBM® SPSS® Modeler に取り込むデータを広範に検査するための手段を提供しています。データ検査レポートは、初期データ探索時に頻繁に使用され、各データ フィールドのヒストグラムや棒グラフのほかに、要約統計量を表示し、欠損値、外れ値、および極値の処理を指定できます。

この例では、telco_dataaudit.str という名前のストリームを使用します。このストリームは、telco.sav という名前のデータ ファイルを参照します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、Windows の [スタート] メニューの 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 ディレクトリにあります。

ストリームの構築

- ストリームを構築するには、IBM® SPSS® Modeler インストール ディレクトリの Demos ディレクトリにある telco.sav を示す Statistics ファイル入力ノードを追加します。

図 7-1
ストリームの構築



▶ データ型ノードを追加し、[解約] を対象フィールドとして指定します（役割 = **対象**）。それ以外のすべてのフィールドについては、役割を**入力**に設定して、それを唯一の対象にする必要があります。

図 7-2
対象の設定

▶ フィールドの測定レベルが適切に定義されていることを確認します。たとえば、値 0 や 1 を持つ多くのフィールドはフラグとして認識されます

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としてより正確に認識されます。

図 7-3
測定レベルの設定

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プロパティを変更するには、[値] 列のヘッダをクリックしてフィールドを列によってソートし、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、選択したフィールドの上で右クリックをすると、選択したフィールドの測定レベルまたは属性を変更することができます。

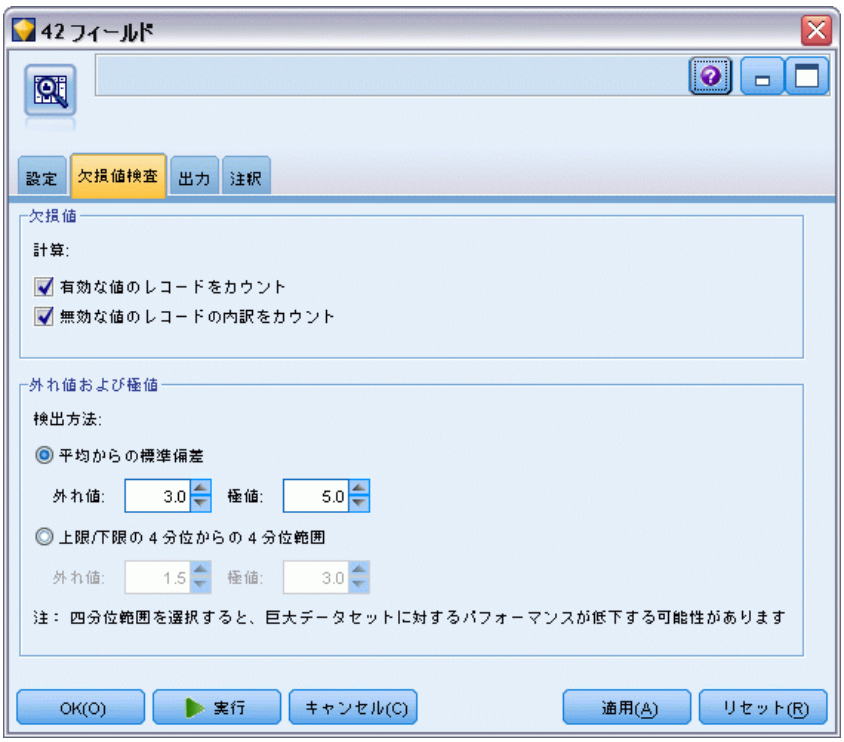
▶ データ検査ノードをストリームに接続します。[設定] タブでは、デフォルト設定はそのままにしてすべてのフィールドをレポートに含めます。[解

約］はデータ型ノードで定義された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。

図 7-4
データ検査ノード、[設定] タブ

［欠損値検査］タブでは、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 7-5
データ検査ノード、［欠損値検査］タブ



統計とグラフのブラウジング

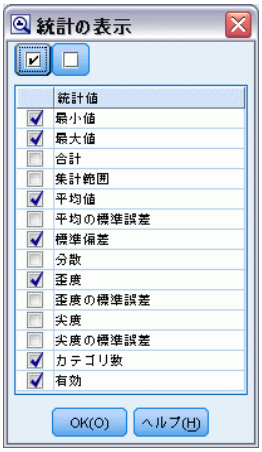
データ検査ブラウザが表示され、フィールドごとにサムネイル グラフや記述統計を示します。

図 7-6
データ検査ブラウザ

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（カテゴリ別フィールドの場合のみ）。

▶ 表示すべき統計を選択する場合は、ツールバーまたは［編集］メニューも使用できます。

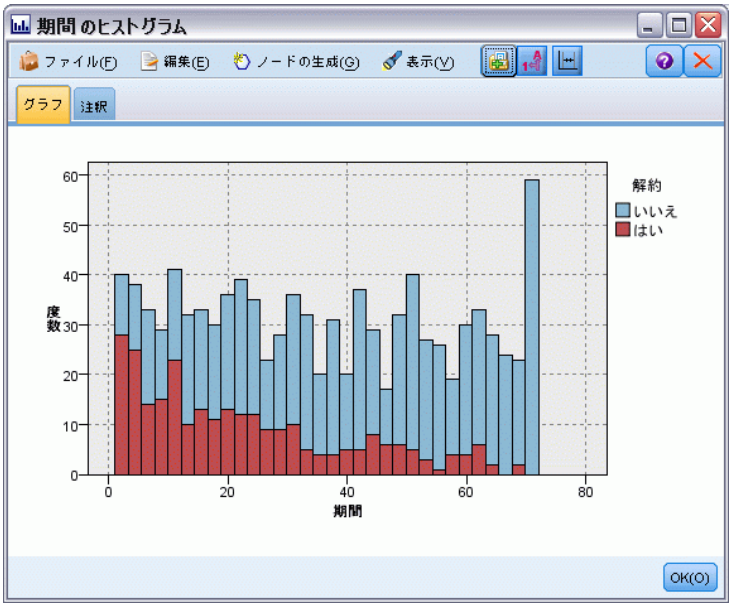
図 7-7
統計の表示



検査レポートのサムネイル 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。［解約］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 7-8
保有期間のヒストグラム



あるいは、1 つ以上のサムネイ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 キャンバス上に配置され、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 7-9
グラフ ノードの生成

外れ値および欠損値の処理

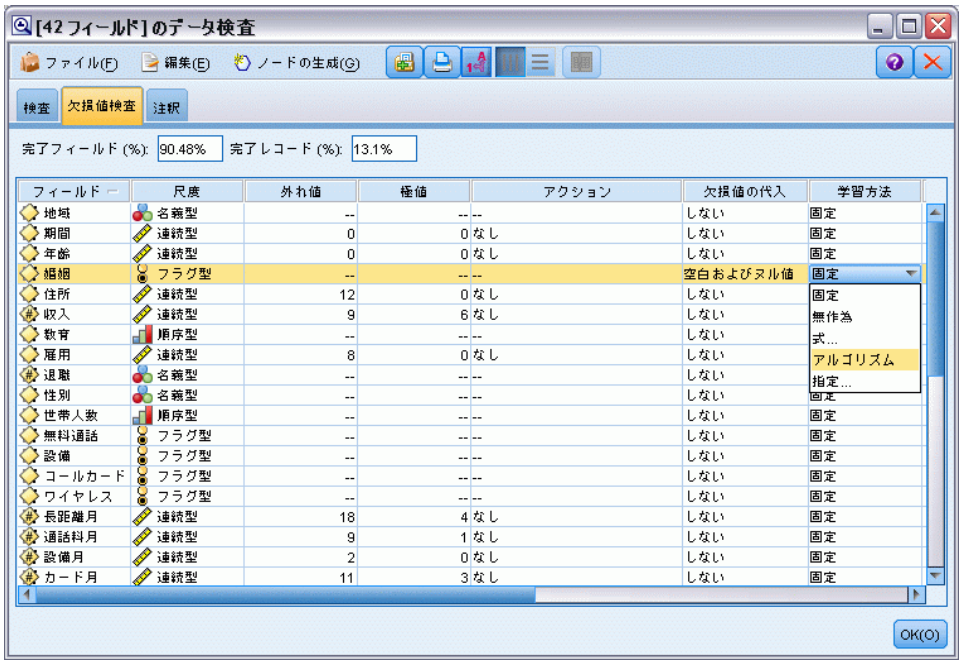
検査レポートの［欠損値検査］タブは、外れ値、極値、および欠損値に関する情報を示します。

図 7-10
データ検査ブラウザ、[欠損値検査] タブ

フィールド	尺度	外れ値	極値	アクション	欠損値の代入	学習方法
地域	名義型	--	--	--	しない	固定
期間	連続型	0	0	なし	しない	固定
年齢	連続型	0	0	なし	しない	固定
婚姻	フラグ型	--	--	--	しない	固定
住所	連続型	12	0	なし	しない	固定
収入	連続型	9	6	なし	しない	固定
教育	順序型	--	--	--	しない	固定
雇用	連続型	8	0	なし	しない	固定
退職	名義型	--	--	--	しない	固定
性別	名義型	--	--	--	しない	固定
世帯人数	順序型	--	--	--	しない	固定
無料通話	フラグ型	--	--	--	しない	固定
設備	フラグ型	--	--	--	しない	固定
コールカード	フラグ型	--	--	--	しない	固定
ワイヤレス	フラグ型	--	--	--	しない	固定
長距離月	連続型	18	4	なし	しない	固定
通話料月	連続型	9	1	なし	しない	固定
設備月	連続型	2	0	なし	しない	固定
カード月	連続型	11	3	なし	しない	固定

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。たとえば、1 つまたは複数のフィールドを選択したり、C&RT アルゴリズムなど、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。

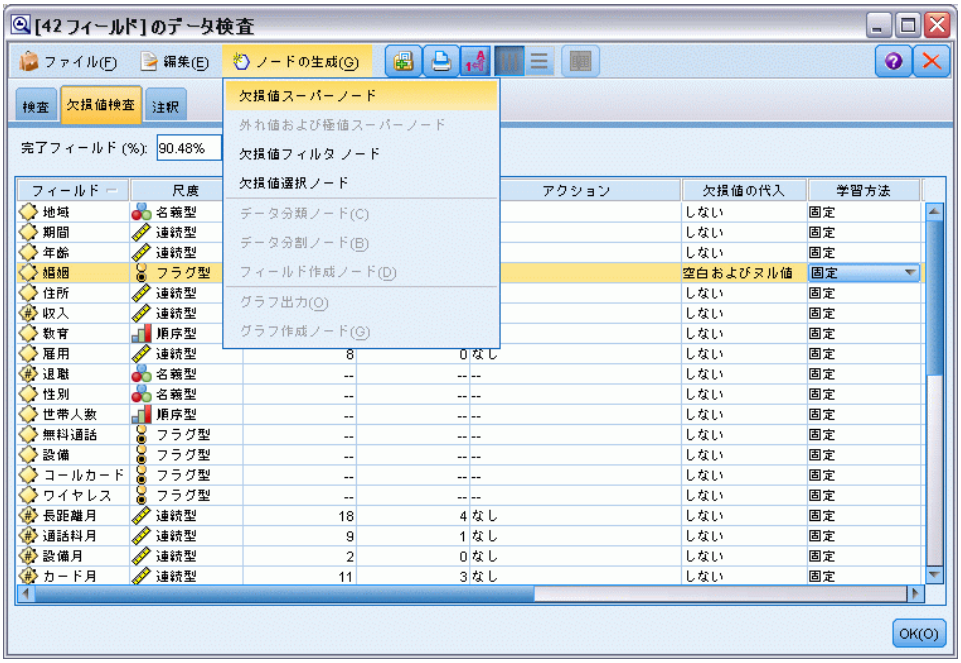
図 7-11
代入法の選択



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。

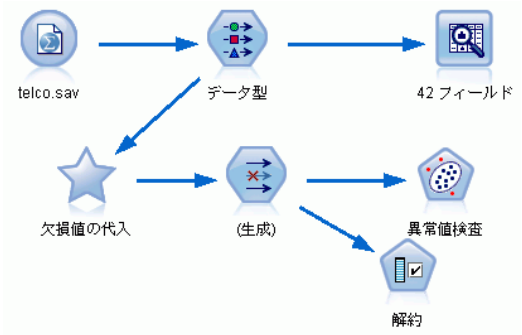
ノードの生成 > 欠損値スーパーノード

図 7-12
スーパーノードの生成



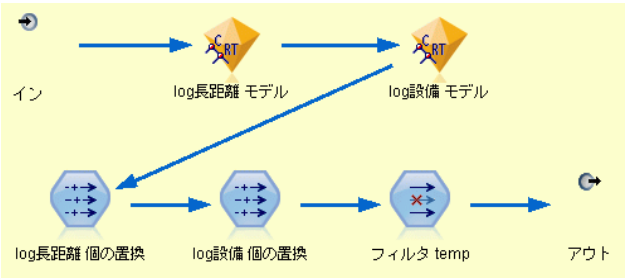
生成されたスーパーノードはストリーム領域に追加され、この場合、それをストリームに接続して変換を適用できます。

図 7-13
欠損値スーパーノードによるストリーム



スーパーノードは、要求された変換を実行する一連のノードを実際に含んでいます。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て［**ズームイン**］をクリックします。

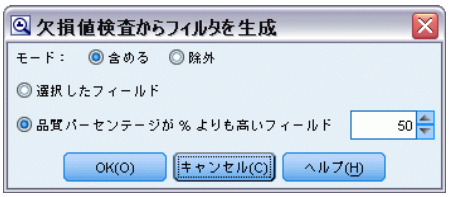
図 7-14
スーパーノードをズーム イン



たとえば、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、空白値やヌル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 モデルが作成されます。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。

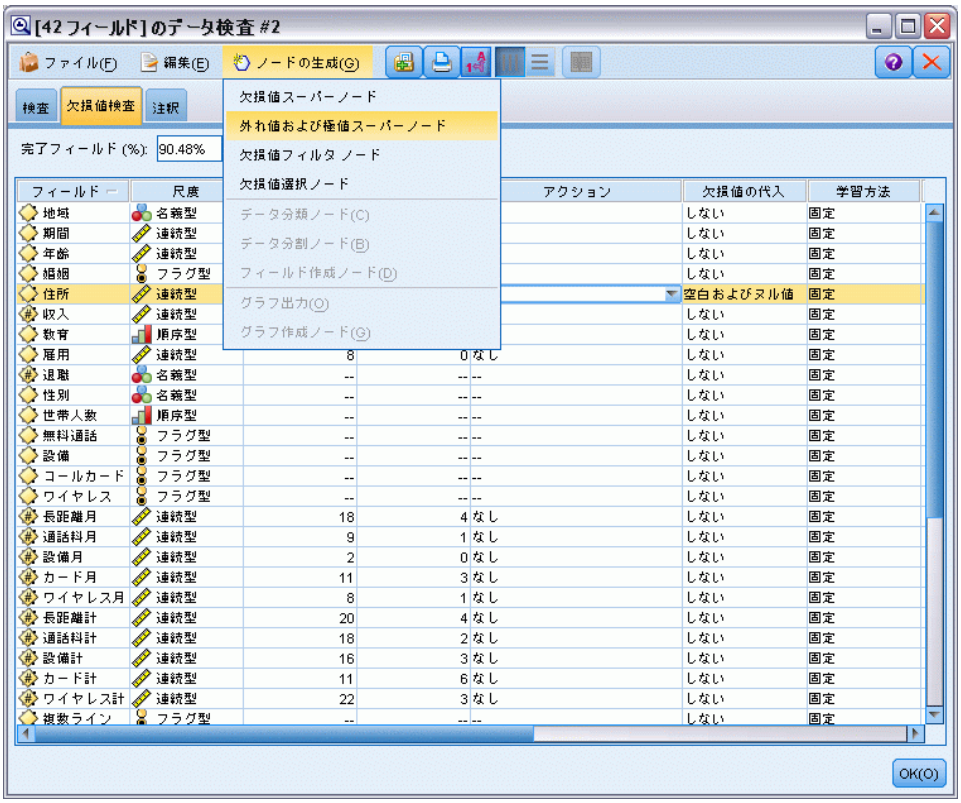
あるいは、条件抽出ノードまたはフィルタ ノードを生成して、欠損値を含むフィールドまたはレコードを削除できます。たとえば、指定された閾値未満の品質パーセンテージのフィールドをフィルタリングできます。

図 7-15
フィルタ ノードの生成



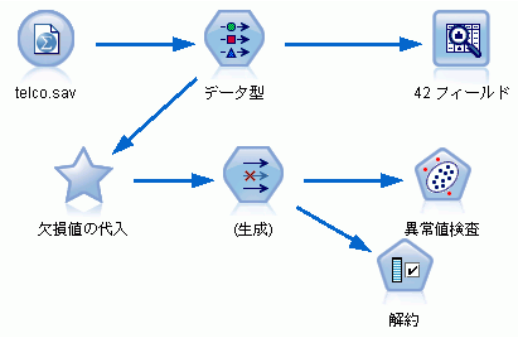
外れ値と極値は同じ方法で処理できます。希望の動作をフィールドごとに指定し（強制、破棄、無効のいずれか）、スーパーノードを生成して変換を適用します。

図 7-16
フィルタ ノードの生成



検査を終了してノードをストリームに追加したら、分析に取りかかることができます。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選別します。

図 7-17
欠損値スーパーノードによるストリーム



章 8

薬品による治療 (調査用グラフ /C5.0)

このセクションでは、医療研究者が研究データを整理する場合を考えてみましょう。ここでは、同じ病気に悩む患者に関するデータを収集しています。治療過程において、それぞれの患者に対して 5 種類の薬品の中のいずれかで効果がありました。そこで、今後同じ病気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。これは、DRUG1n という名前のデータ ファイルを参照します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、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 ディレクトリにあります。

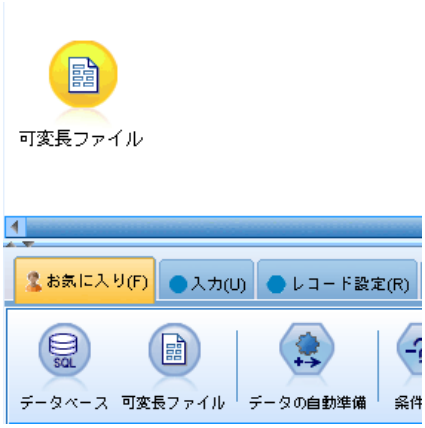
このデモでは、次のデータ フィールドを使用します。

データ フィールド	説明
Age (年齢)	(数値)
Sex	M (男性) または F (女性)
BP	血圧 :高、正常、または低
Cholesterol (コレステロール)	血中コレステロール :正常または高
Na	血液中のナトリウム濃度
K	血液中のカリウム濃度
Drug (薬品)	患者に効果があった処方薬

テキスト データの読み込み

可変長ノードを使用して、区切り記号付きのテキスト データを読み込むことができます。可変長ノードはパレットから追加します。[**入力**] タブをクリックして目的のノードを探すか、またはデフォルトで可変長ノードがある [**お気に入り**] タブを使用してください。次に、配置したノードをダブルクリックして、ダイアログ ボックスを表示します。

図 8-1
可変ファイル ノードの追加



[ファイル] ボックスの右にある [...] ボタンをクリックして、IBM® SPSS® Modeler がインストールされたディレクトリを参照します。Demos ディレクトリを開き、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] を選択します。ダイアログ ボックスに、ロードしたフィールドが表示されます。

図 8-2
[可変長ファイル] ダイアログ ボックス

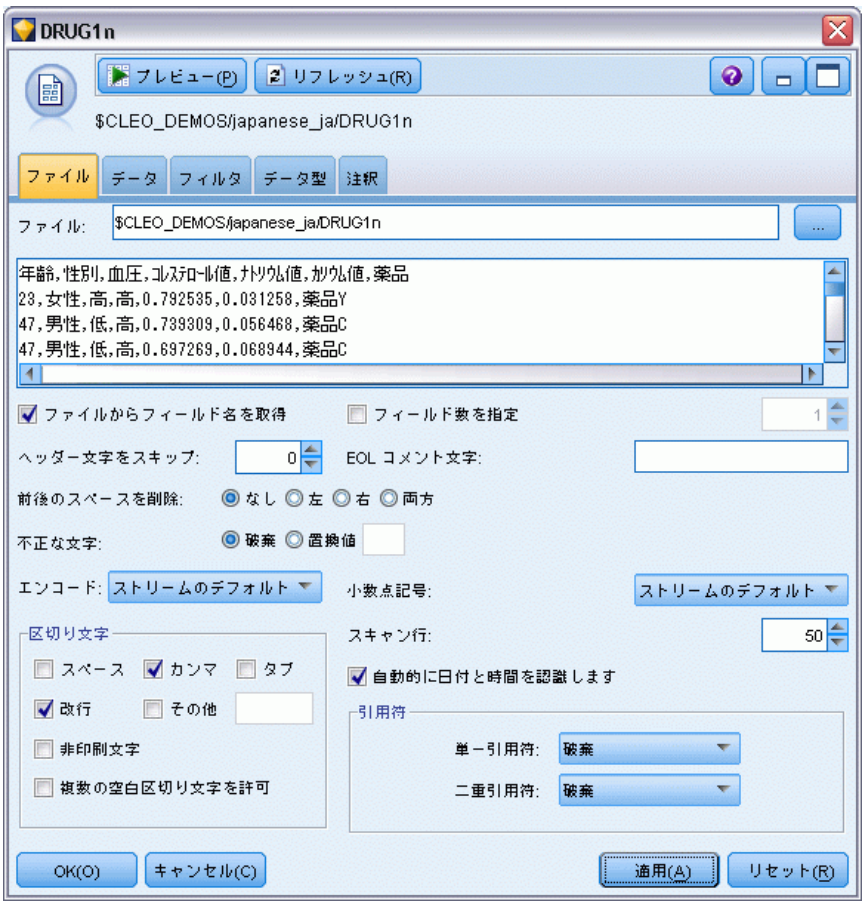


図 8-3
フィールドのストレージ タイプの変更

図 8-4
[データ型] タブの [値] オプションの選択

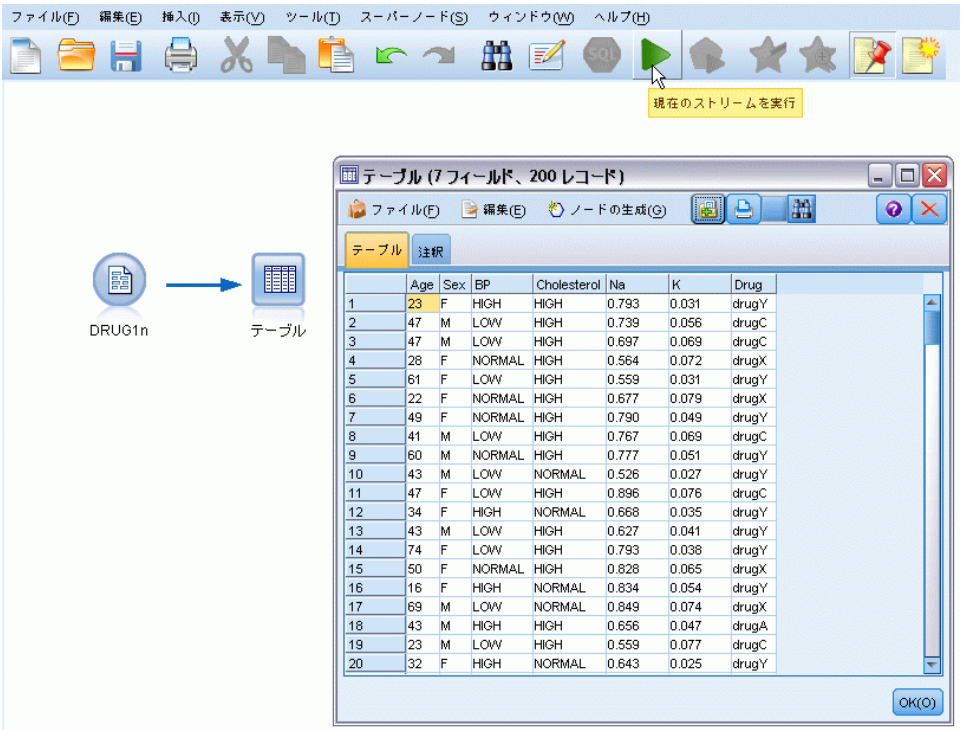
[**データ**] タブをクリックして、フィールドの**ストレージ**を変更します。ストレージは**尺度**とは異なり、またデータ フィールドの測定レベル（使用タイプ）とも異なることに注意してください。 [**データ型**] タブでは、データ中のフィールドのデータ型を確認することができます。また、[**値の読み込み**] を選択し、[値] 列の選択内容に基づいて、各フィールドの実際の値を表示することもできます。この過程は、**インスタンス化**として知られています。

テーブルの追加

データ ファイルをロードしたら、レコードの値を確認してみましょう。レコードの値を表示する方法の 1 つとして、テーブル ノードを含むストリームを作成することがあげられます。ストリームにテーブル ノードを配置するには、パレット中のアイコンをダブル クリックするか、またはアイコンをストリーム領域にドラッグ アンド ドロップしてください。

図 8-5
データ ソースに接続したテーブル ノード

図 8-6
ツールバーからストリームを実行



	Age	Sex	BP	Cholesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0.793	0.031	drugY
2	47	M	LOW	HIGH	0.739	0.056	drugC
3	47	M	LOW	HIGH	0.697	0.069	drugC
4	28	F	NORMAL	HIGH	0.564	0.072	drugX
5	61	F	LOW	HIGH	0.559	0.031	drugY
6	22	F	NORMAL	HIGH	0.677	0.079	drugX
7	49	F	NORMAL	HIGH	0.790	0.049	drugY
8	41	M	LOW	HIGH	0.767	0.069	drugC
9	60	M	NORMAL	HIGH	0.777	0.051	drugY
10	43	M	LOW	NORMAL	0.526	0.027	drugY
11	47	F	LOW	HIGH	0.896	0.076	drugC
12	34	F	HIGH	NORMAL	0.668	0.035	drugY
13	43	M	LOW	HIGH	0.627	0.041	drugY
14	74	F	LOW	HIGH	0.793	0.038	drugY
15	50	F	NORMAL	HIGH	0.828	0.065	drugX
16	16	F	HIGH	NORMAL	0.834	0.054	drugY
17	69	M	LOW	NORMAL	0.849	0.074	drugX
18	43	M	HIGH	HIGH	0.656	0.047	drugA
19	23	M	LOW	HIGH	0.559	0.077	drugC
20	32	F	HIGH	NORMAL	0.643	0.025	drugY

パレットでノードをダブル クリックすると、そのノードはストリーム領域中の選択されているノードと自動的に接続されます。または、ノードがまだ接続されていなければ、マウスの真ん中のボタンを使用して、入力ノードとテーブル ノードを接続できます。中央のボタンをシミュレートするには、Alt キーを押さえたままマウスを使用します。 テーブルを表示する場合は、ツールバーの緑の矢印ソタンをクリックしてストリームを実行します。または、テーブル ノードを右クリックして、[**実行**] を選択します。

棒グラフの作成

データ マイニング時、視覚的な要約を作成してデータを探索する場合に役立ちます。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が用意されています。たとえば、投薬を受けた患者の割合を薬品ごとに調べるには、棒グラフ ノードを使用します。

棒グラフ ノードをストリームに追加して入力ノードに接続し、次に棒グラフ ノードをダブルクリックして、表示用のオプションを編集します。

棒グラフ ノードを表示する対象フィールドとして、[薬品] を選択します。次に、ダイアログ ボックスから **[実行]** を選択します。

図 8-7
対象フィールドとして薬品を選択

結果のグラフは、データの「形状」を見るのに役立ちます。このグラフは、薬品 Y が最も効果があり、薬品 B と薬品 C が最も効果がなかったことを表しています。

図 8-8
薬品の種類と効果の分布

図 8-9
データ検査の結果

代わりに、データ検査ノードを追加してから実行することで、一度にすべてのフィールドの散布図と棒グラフを簡単に表示することもできます。データ検査ノードは、［出力］ タブから利用することができます。

散布図の作成

ここで、どのような因子が対象変数薬品に影響しているかを調べます。データからは、ナトリウムとカリウムの血中濃度が重要な因子となっていることがわかります。これらの値は両方とも数値なので、薬品カテゴリを折れ線グラフで色をオーバーレイして、ナトリウム対カリウムの散布図を作成することができます。

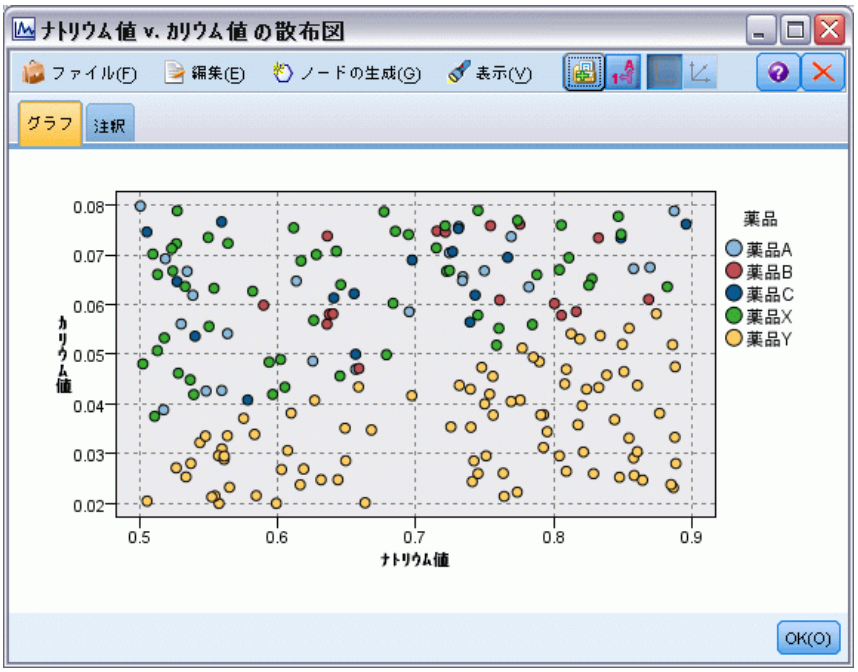
作業領域にプロット ノードを配置し、それを入力ノードに接続し、ダブルクリックしてノードを編集します。

［プロット］タブで、X フィールドに［Na］を、Y フィールドに［K］を、オーバーレイ フィールドに［薬品］を選択します。**[実行]** をクリックします。

図 8-10
散布図の作成

この散布図では、1 つの閾値が明確に現れています。この閾値より上なら適切な薬品は常に Y となり、下なら、適切な薬品が Y であることはありません。この閾値は、カリウム（K）に対するナトリウム（Na）の比率です。

図 8-11
薬品分布の散布図

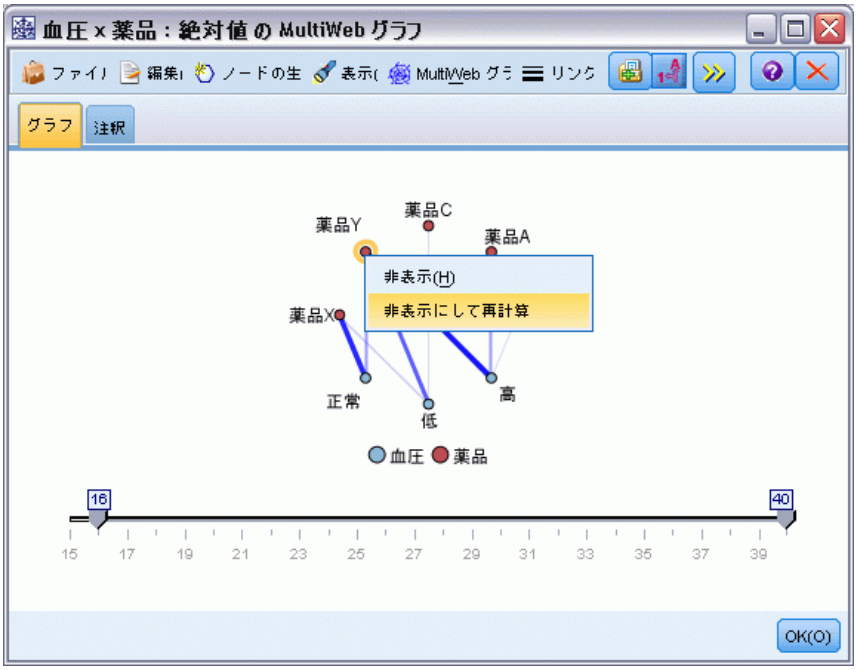


Web グラフの作成

データ フィールドの多くはカテゴリなので、異なるカテゴリ間の関連付けを行う Web グラフのプロットも試行できます。Web グラフ ノードを作業領域内の入力ノードに接続します。 Web グラフ ノードのダイアログ ソックスで［BP］（血圧）と［薬品］を選択します。**[実行]** をクリックします。

散布図から、薬品 Y が血圧の 3 つのレベルすべてと関連していることがわかります。これは何も意外なことではありません。薬品 Y が最適となる状況はすでにわかっています。他の薬品に焦点を移す場合は、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 のポイントを右クリックし、表示されるポップアップ メニューから **[非表示にして再計算]** を選択します。

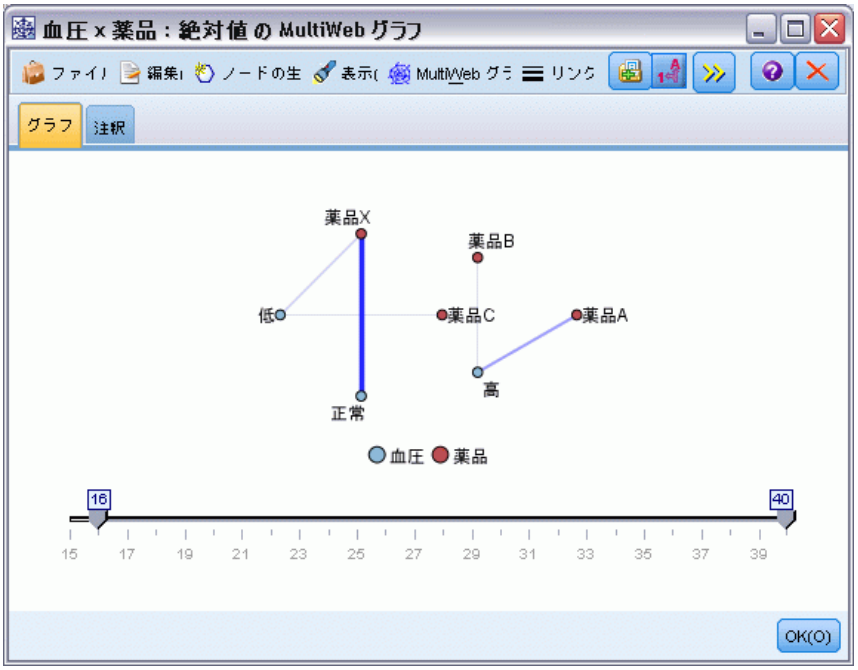
図 8-12
薬品の Web グラフと血圧



簡素化された散布図では、薬品 Y とそのすべてのリンクが表示されません。これで、薬品 A と薬品 B が高血圧に関係することがはっきりわかるようになりました。低血圧には、薬品 C と薬品 X だけが関係しています。また、正常な血圧には、薬品 X だけが関係しています。しかし現時点では、特定の患者に対して薬品 A と薬品 B、または薬品 C と薬品 X の

どちらを選択すればよいかを判断することはできません。このような場合に、モデル作成が役立ちます。

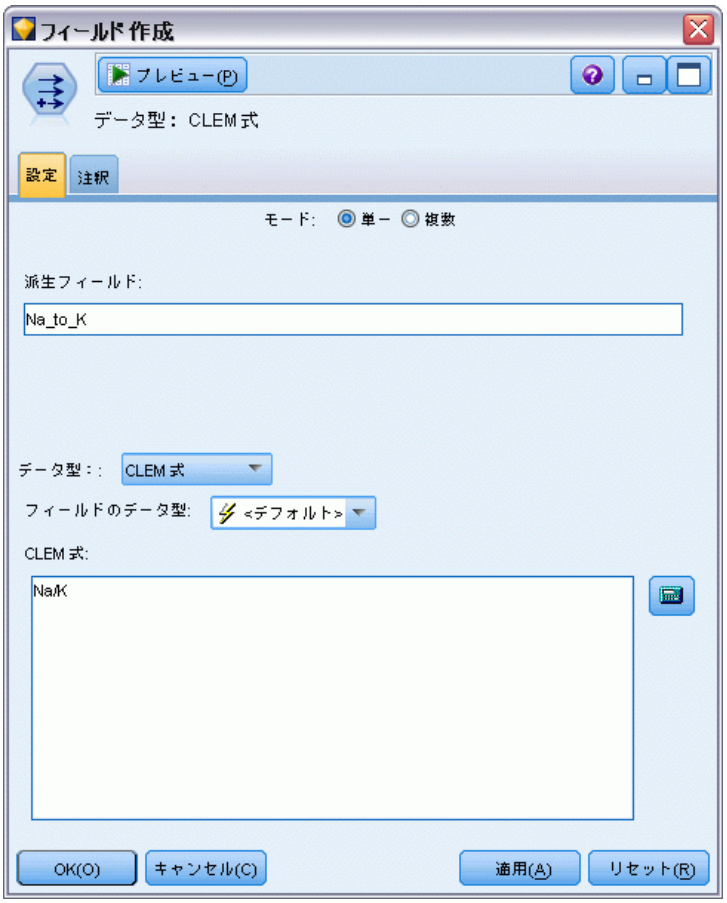
図 8-13
薬品 Y の Web グラフとその隠れたリンク



新規フィールドの作成

ナトリウムのカトリウムに対する比率で薬品 Y をどの時点で使用すべきかを予想できると思われるので、この比率の値を含むフィールドを各レコードに作成することができます。このフィールドは後に、5 種類の薬品のそれぞれをどの時点で使用すべきかを予測するモデルを構築する際に役に立ちます。 ストリーム レイアウトを単純化するには、まず DRUG1n 入力ノード以外のすべてのノードを削除します。フィールド作成ノード（[フィールド] 設定タブ）を DRUG1n に接続し、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 8-14
フィールド作成ノードの編集



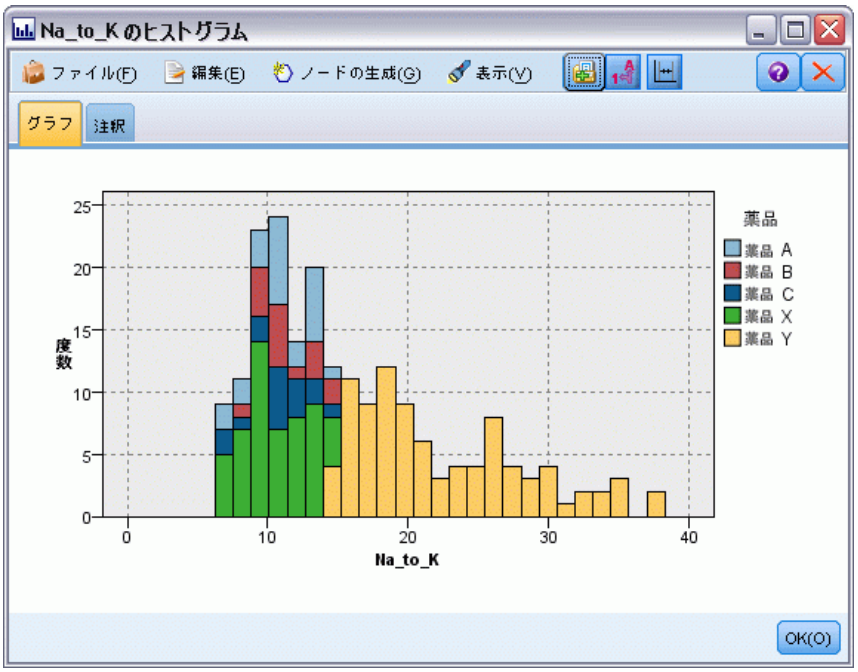
新規フィールドの名前に、「ナトリウム値/カリウム値」を指定します。新しいフィールドの値は、ナトリウム値をカリウム値で除算して得られるので、式に「**ナトリウム値/カリウム値**」と入力します。フィールドの右にあるアイコンをクリックして式を作成することもできます。 アイコンをクリックすると、Clem 式ビルダーが表示されます。これを利用すれば、すでに存在している関数、オペランド、およびフィールドと値を使って、対話的に式を作成することができます。

新規フィールドの棒グラフを確認するには、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 ノードのダイアログ ボックスで、プロットするフィールドとしてナトリウム値/カリウム値を、オーバーレイ フィールドとして薬品を指定します。

図 8-15
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、このようなグラフが表示されます。表示された内容によると、ナトリウム値/カリウム値の値が約 15 以上になったときに、薬品 Y を使用すればよいと判断できます。

図 8-16
ヒストグラム表示



モデルの構築

データの調査と操作により、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に思えます。ただし、これですべての関係を完全に説明することはできません。そこで、モデルを作成していくつかの結論を導き出してみることにします。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。

ここでは作成したフィールド［ナトリウム値/カリウム値］を使用するため、元のフィールド［ナトリウム値］と［カリウム値］がモデリング アルゴリズムで 2 度使用されないように、これらにフィルタを適用して除外できます。この作業は、フィルタ ノードで実行できます。

図 8-17
フィルタ ノードの編集

［フィルタ］タブで、［Na］および［K］の隣にある矢印をクリックします。フィールドが除外されることを表す赤い X が矢印上に表示されます。

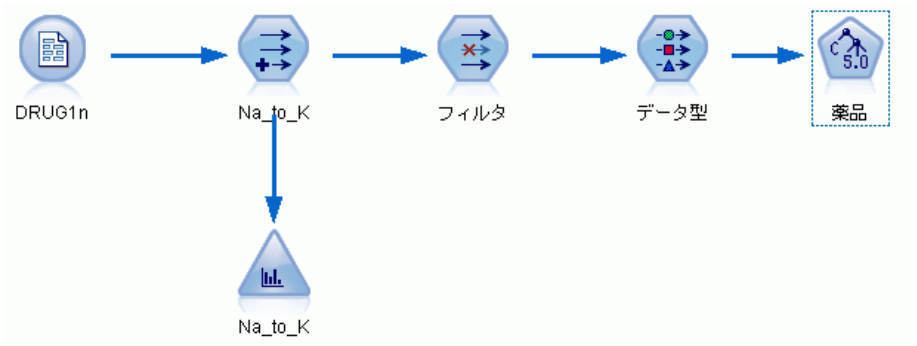
次に、データ型ノードをフィルタ ノードに接続します。データ型ノードでは、使用するフィールドのデータ型と、結果を推定するための使用方法を指定できます。

［データ型］タブで、［薬品］フィールドの役割を、［薬品］が予測対象のフィールドであることを示して、[対象] に設定します。他のフィールドの役割は、予測フィールドとして使用されるように、[入力] の設定のままにします。

図 8-18
データ型ノードの編集

モデルを評価するには、C5.0 ノードを作業領域に配置し、それを表示されているようにストリームの端に接続します。次に、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。

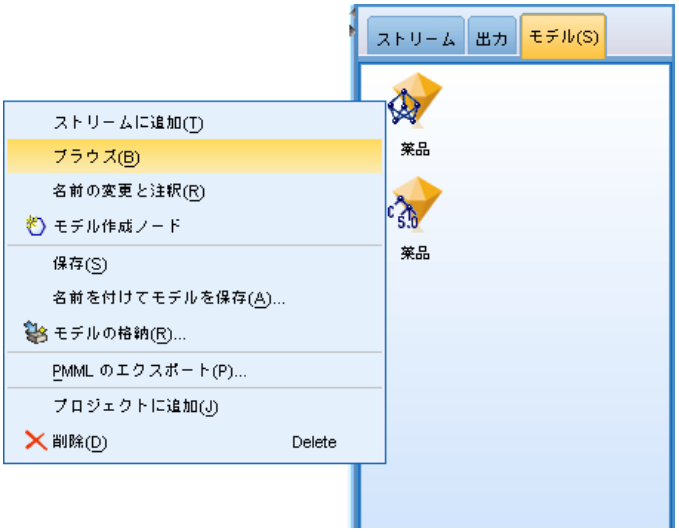
図 8-19
C5.0 ノードの追加



モデルの参照

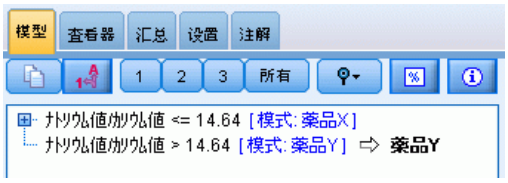
C5.0 ノードが実行されると、モデル ナゲットがストリーム、およびウィンドウの右上の［モデル］パレットに追加されます。モデルを参照するには、このアイコンのいずれかを右クリックし、コンテキスト メニューから［編集］または［参照］を選択します。

図 8-20
モデルの参照



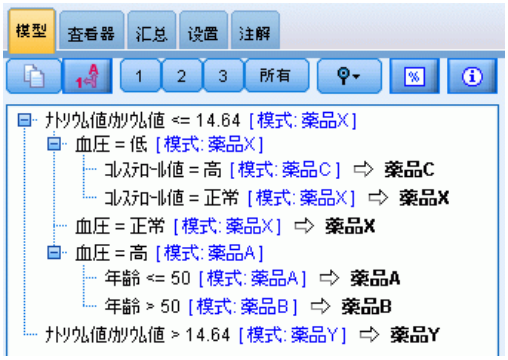
ルール ブラウザに、C5.0 ノードによって生成されたルールセットがディシジョン ツリー形式で表示されます。最初の状態では、ツリーは閉じられています。[すべて] ボタンをクリックすると、ツリーが展開されすべてのレベルが表示されます。

図 8-21
ルール ブラウザ



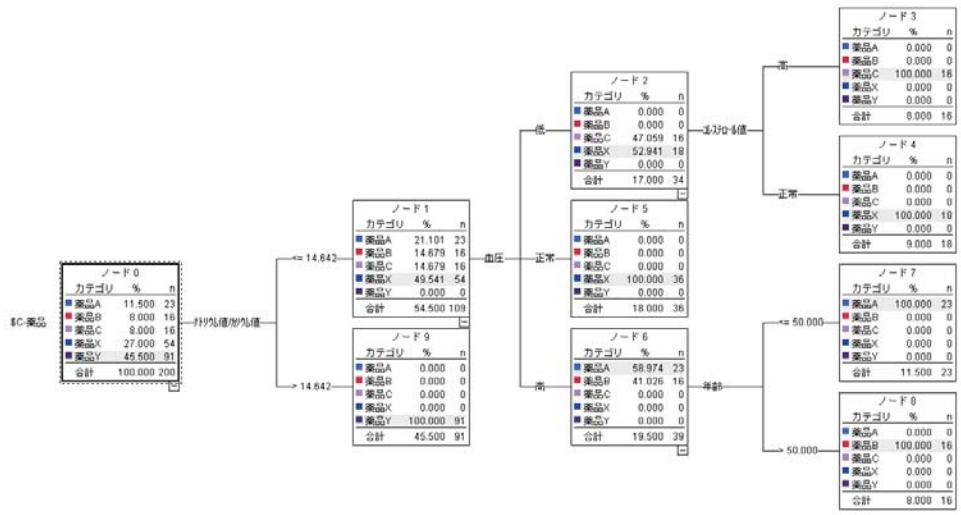
これで、パズルの欠けている断片が埋まりました。ナトリウム値対カリウム値の比率が 14.64 より小さく、高血圧の人の場合は、年齢によって選択する薬品が決まります。低血圧の人の場合は、コレステロールのレベルが最適の予測値と見なされます。

図 8-22
完全に開いたルール ブラウザ



[ビューア] タブをクリックすると、同じディシジョン ツリーをよりわかりやすいグラフィックで参照することができます。ここでは、血圧カテゴリごとのケース数や各ケースの比率などをより簡単に確認することができます。

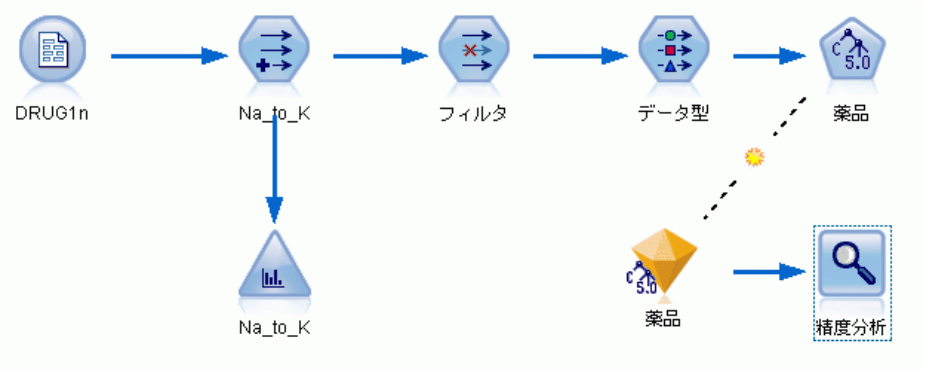
図 8-23
グラフィカル形式のディシジョン ツリー



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いて［実行］をクリックします。

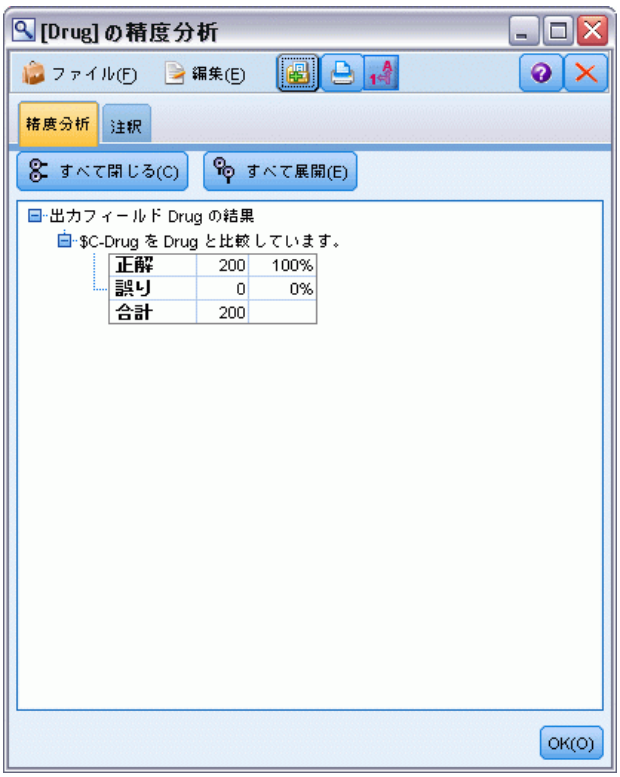
図 8-24
精度分析ノードの追加



精度分析ノードの出力には、このサンプル データセットにより、モデルが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が示されます。実際のデータセットでは 100% の精度はなかなか見られませ

んが、モデルが特定のアプリケーション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。

図 8-25
精度分析ノードの出力



章 9

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。膨大な数の予測フィールドから、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、そして選択します。最終的に、より高速で効果的な少数のモデル、予測フィールドを使用し、実行速度が速く、分かりやすいモデルが得られます。

この例で使用するデータは、架空の電話会社のデータ ウェアハウスのものであり、特別プロモーションに対するこの会社の 5,000 人の顧客からの応答に関する情報があります。このデータには、顧客の年齢、雇用、収入、電話利用状況の統計などの多くのフィールドが含まれています。3 つの「対象」フィールドは、顧客がこの 3 つのフィールドに反応したかどうかを示しています。この会社は、このデータを活用して、今後、類似のオファーに対してどの顧客が反応する見込みが最も高いかという予測を立てたいと考えています。

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。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、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 ディレクトリにあります。

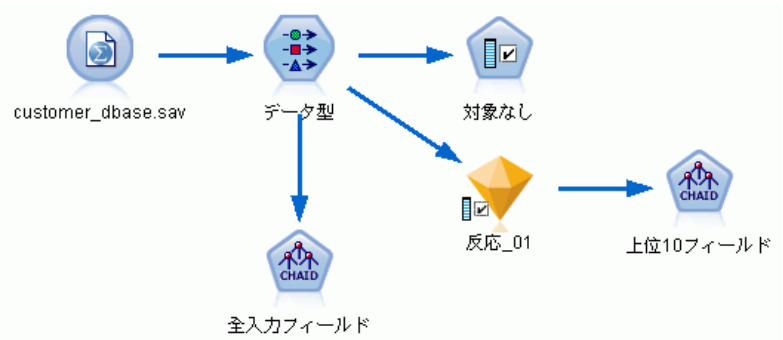
この例では、対象となるオファーの 1 つのみを取り上げます。どの顧客が最も良く販売促進活動に応答しやすいかを記述するモデルを作成するには、CHAID ツリー構築ノードを使用します。ここで次の 2 つのアプローチを比較します。

- 1 つはフィールド選択をしないアプローチです。データセットのすべての予測フィールドは、CHAID ツリーへの入力として使用されます。
- もう 1 つはフィールド選択をするアプローチです。フィールド選択ノードは、上位 10 の予測フィールドの選択に使用されます。その後で、これらの予測フィールドは 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と、フィールド選択が、いかに有効な結果を生成するかが分かります。

ストリームの構築

図 9-1
フィールド選択のサンプル ストリーム



- Statistics ファイル入力ノードを空白のストリーム キャンバスへ配置します。このノードをサンプル データ ファイル customer_dbase.sav に設定します。このファイルは IBM® SPSS® Modeler をインストールしたディレクトリの下にある Demos ディレクトリにあります（または、streams ディレクトリ内にあるサンプル ストリーム ファイル featureselection.str を開きます）。
- データ型ノードの追加[データ型]タブで、一番下までスクロールして、[response_01] の役割を［対象］に変更します。リスト上部の顧客 ID (custid) とそのほかの回答フィールド（response_02 と response_03）に対しては、役割を［なし］に変更します。その他のすべてのフィールドに対しては、役割を［入力］に設定したままにし、[**値の読み込み**] ボタンをクリックして、[OK] をクリックします。
- フィールド選択モデリング ノードをストリームに追加します。このノードで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。
- 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。

► ストリームまたは［モデル］パレットのモデル ナゲットを右クリックし、［編集］または［ブラウズ］を選択して結果を調べます。

図 9-2
フィールド選択モデル ナゲットのモデル タブ

上位のパネルは、それらのフィールドが予測に有用なことを示しています。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。下位のパネルは、どのフィールドが解析からスクリーニング(選別)されたか、またその理由を示しています。上位のパネルにあるフィールドを検証して、これ以降のモデリング 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。

- 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、これらの予測フィールドをさらに絞り込みたいと思います。

- 最初の列でチェック マークを使用して上位 10 の予測値のみを選択し、不要な予測値を非選択にします(行 11 のチェック マークをクリックし、Shift キーを押したまま行 34 のチェック マークをオンにします)。モデル ナゲットを閉じます。

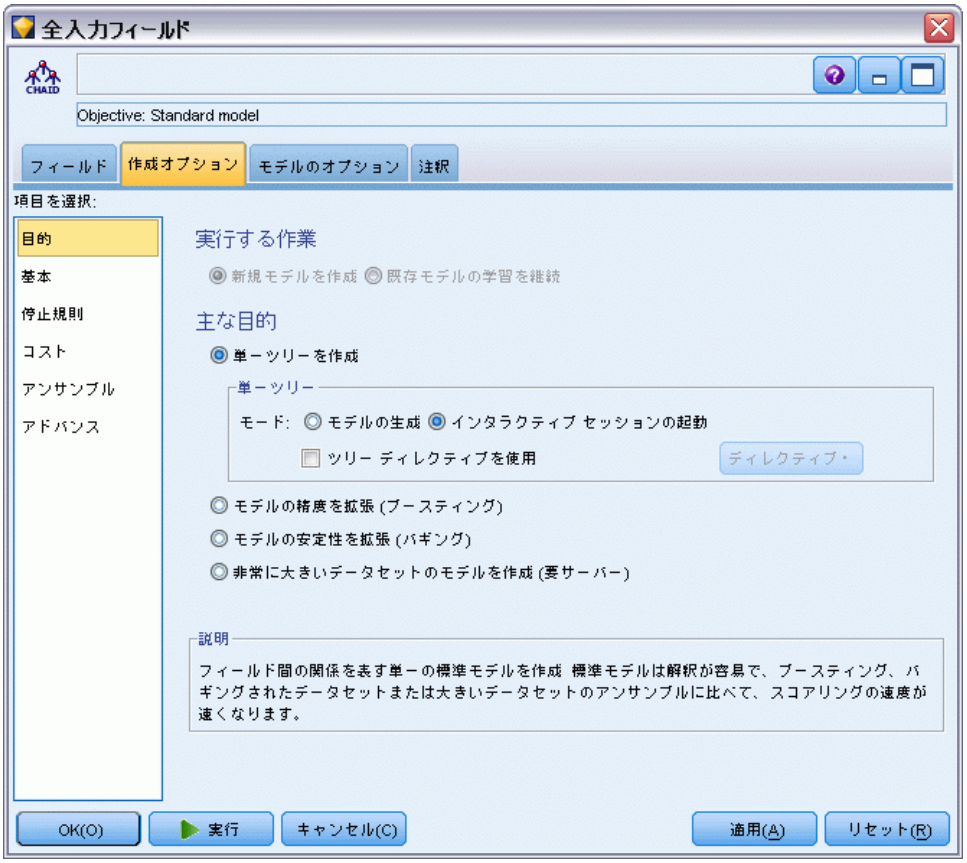
- フィールド選択なしで結果を比較するには、2 つのCHAID モデリングノードをストリームに追加する必要があります。フィールド選択を行うモデルと、フィールド選択を行わないモデルを生成します。

- 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。

- 各 CHAID ノードを開き、[作成オプション] を開き、[目的] パネルで **[新規モデルの作成]**、**[単一ツリーを作成]** および **[インタラクティブ セッションの起動]** を選択します。

［基本］パネルで、[ツリーの最大の深さ] を 5 に設定します。

図 9-3
すべての予測値フィールドに対する CHAID モデリング ノードの [目的] 設定

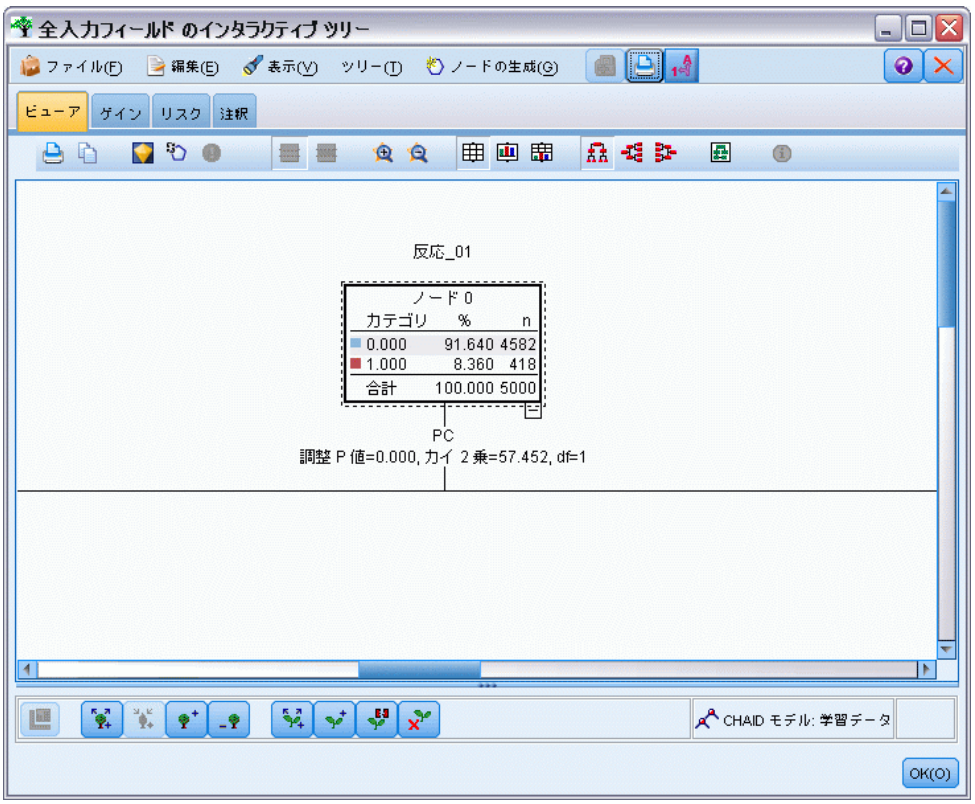


モデルの構築

- データセット内のすべての予測フィールドを使用する CHAIDノード（データ型ノードに接続した方）を実行します。実行時に必要な時間に注目します。結果ウィンドウがテーブルを表示します。

▶ メニューから [Tree > Grow Tree] を選択し、展開したツリーを成長させて表示します。

図 9-4
ツリー ビルダーにおけるツリーの成長



▶ 同じ手順を、10 個のみの予測フィールドを使用する CHAID ノードに対して実行します。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実行速度が速かったはずです。データセットがかなり小さいので、おそらく実行時間の差は数秒のはずです。しかし、もっと大きな実際のデータセットに対しては、この差は非常に大きな差異–つまり数分ないし数時間にも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。2 番目の方法を採用する前に、それが効果的であるかどうか、またすべての予測フィールドを使用するモデルと比較する方法を調べる必要があります。

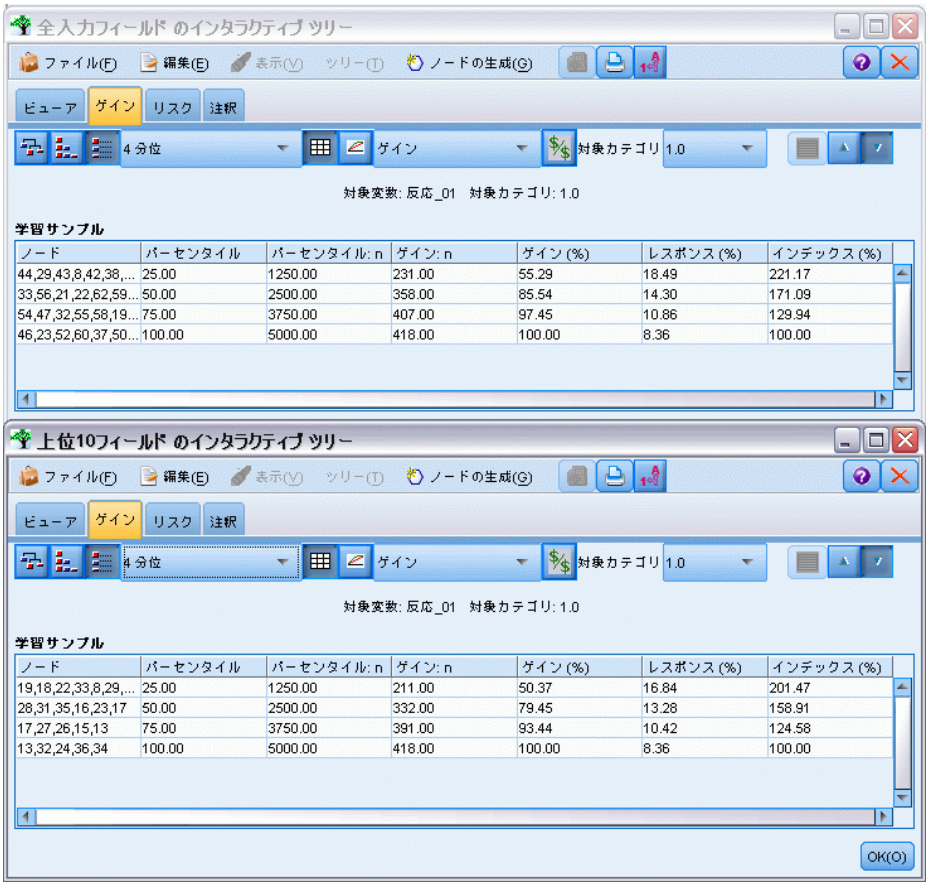
結果の比較

2 つの結果を比較するには、効果の指標（インデックス）が必要です。ツリー ビルダーの［ゲイン］タブでこの指標（インデックス）を使用することができます。［**リフト**］を見て、データセット内のすべてのレコードを比較して、ノード内のレコードがどの程度対象カテゴリに含まれる可能性が高いかを調べます。たとえば、148 % のリフト値は、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 1.48 倍も対象カテゴリに含まれやすいことを示しています。リフトは、［ゲイン］タブのインデックス列に表示されます。

- 完全な予測フィールドについては、［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］を［1.0］に変更します。［4 分位］ツールバー ボタンを最初にクリックして、表示を 4 分位に切り替えます。このボタンの右にあるドロップダウン リストから**[4 分位]**を選択します。

- ツリー ビルダーで、10 の予測フィールドすべてにこの手順を実行すると、次の図のような 2 つの似ているゲイン テーブルが表示され、比較することができます。

図 9-5
2 つの CHAID モデルのゲイン グラフ



各ゲイン テーブルは、そのツリーのターミナル ノードを 4 分位にグループ化しています。2 つのモデルの効果を比較するには、各テーブルの上位の 4 分位のリフト（インデックス値）を調べます。

すべての予測フィールドが含まれた場合、モデルは 221% のリフトを示します。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 倍反応する見込みがあります。これらの特徴の内容を確認するには、クリックして最上位の行を選択します。次に、［ビューア］タブに切り替えます。ここでは該当するノードが黒のハイライトで表示されます。ハイライト表示されているターミナル モードまでツリーを下にたどり、予測フィールドがどの程度分割されているかを確認します。上位の 4 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 モデルに変換された場合、10 件の顧客のプロファイルを管理するのは困難です。

フィールド選択によって識別されたわずか 10 の予測フィールドが含まれるだけで、リフトはおよそ 194% となります。すべての予測フィールドを使用するモデルに比べ、このモデルが必ずしも際立って優れているわけ

ではありませんが、有用なのは間違いありません。ここで上位 4 分位に含まれるノードはわずか 4 つのみであり、非常にシンプルになっています。したがって、フィールド選択モデルは、すべての予測フィールドを持つモデルよりも望ましいです。

要約

フィールド選択の利点を見直してみましょう。予測フィールドが少ないほどコストが少なくなります。つまり、データ収集量、プロセス、モデルへのデータ入力が少なくなります。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としても、予測フィールドが少数なのでモデル構築が大幅に速くなりました。より大きな実際のデータセットの場合は、この時間短縮は非常に大きなものになります。

より少ない予測フィールドを使用すると、スコアリングが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。予想値の数が大きいほど、モデルをオーバーフィットする危険があります。このシンプルなモデルは他のデータセットへよりうまく一般化する可能性があります（ただし確認のために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ができるため、最も重要な予測フィールドをツリーが識別できるようになりました。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、また、ツリーの深度と複雑性をコントロールするためにレベル1ごとにツリーを成長させることもできます。しかし、フィールド選択ノードは処理が高速で使い方は簡単です。すべての予測フィールドを1回の速い操作でランク付けできるため、最も重要なフィールドを敏速に識別することができます。また、含める予測フィールドの数を変更することもできます。上位 10 件の代わりに、上位 15 件または 20 件の予測フィールドを使用してこの例をもう一度実行し、最適モデルを決定するために結果を比較することも簡単にできます。

章
10

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（データ分類）

二項ロジスティック回帰モデルおよび自動分類モデルを含む 2 値の分類モデルの場合、文字列フィールドは最大 8 文字に制限されています。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。

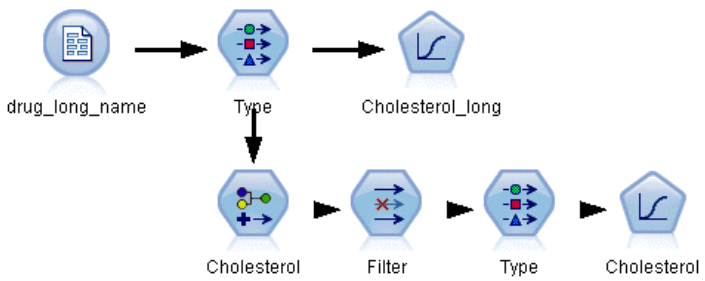
この例では、reclassify_strings.str という名前のストリームを使用します。これは、drug_long_name という名前のデータ ファイルを参照します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、Windows の [スタート] メニューの IBM® SPSS® Modeler プログラム グループからアクセスできます。reclassify_strings.str ファイルは、streams ディレクトリにあります。

この例では、ストリームの一部に焦点を当て、長すぎる文字列で生成されるエラーの種類を示し、データ分類ノードを使用して文字列を適切な長さに変更する方法を説明します。この例では二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二項ロジスティック回帰モデルを生成した場合も同様に適しています。

データの分類

- 変数ファイル入力ノードを使用して、Demos フォルダのthe dataset drug_long_name というデータセットに接続します。

図 10-1
二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



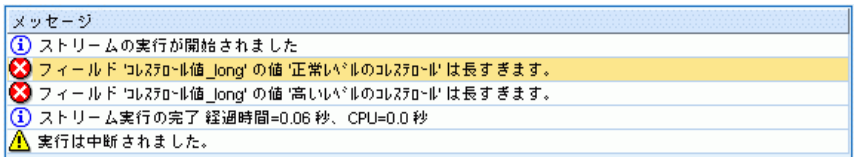
- データ型ノードを入力ノードに追加して、対象として Cholesterol_long を選択します。
- ロジスティック回帰ノードをデータ型ノードに追加します。
- ロジスティック回帰ノードで、[モデル] タブをクリックし、[二項検定] 手続きを選択します。

図 10-2
「Cholesterol_long」フィールドの長い文字列の詳細

- 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。

  こうしたエラー メッセージが表示された場合、この例の後半で説明されている手順に従って、データを変更します。

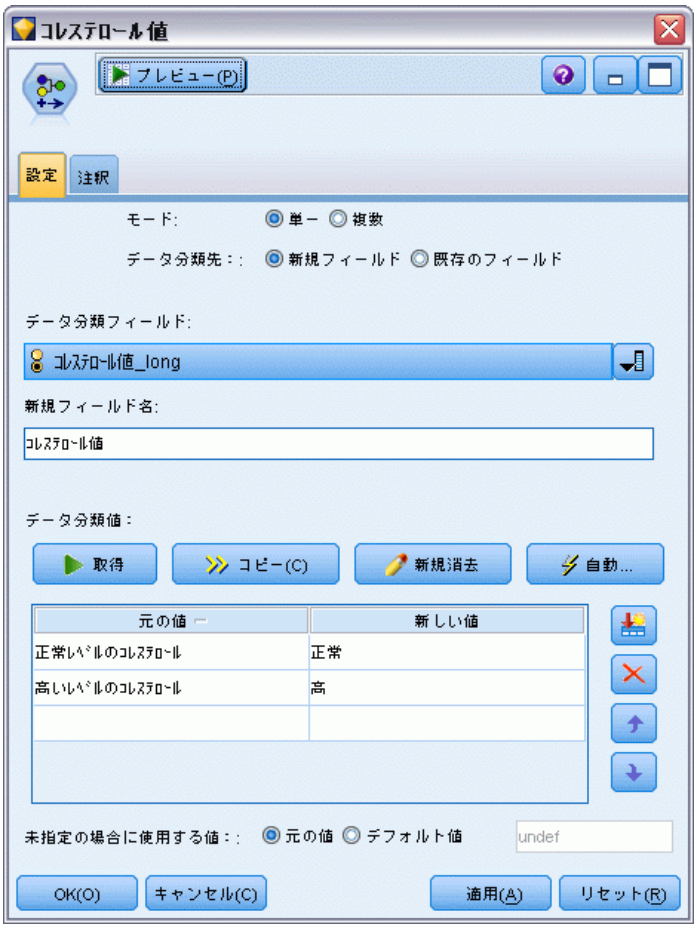
図 10-3
二項ロジスティック回帰ノード実行時に表示されたエラー メッセージ



- データ分類ノードをデータ型ノードに追加します。
- データ分類フィールドで、[Cholesterol_long] を選択します。

- 新しいフィールド名として、「Cholesterol」と入力します。
- [取得] ボタンをクリックして、Cholesterol_long の値を元の値の列に追加します。
- 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 10-4
長い文字列のデータ分類



- フィルタ ノードをデータ分類ノードに追加します。

▶ ［フィルタ］列で、[Cholesterol_long] をクリックして削除します。

図 10-5
データの「Cholesterol_long」フィールドの削除

▶ データ型ノードをフィルタ ノードに追加して、対象として [Cholesterol] を選択します。

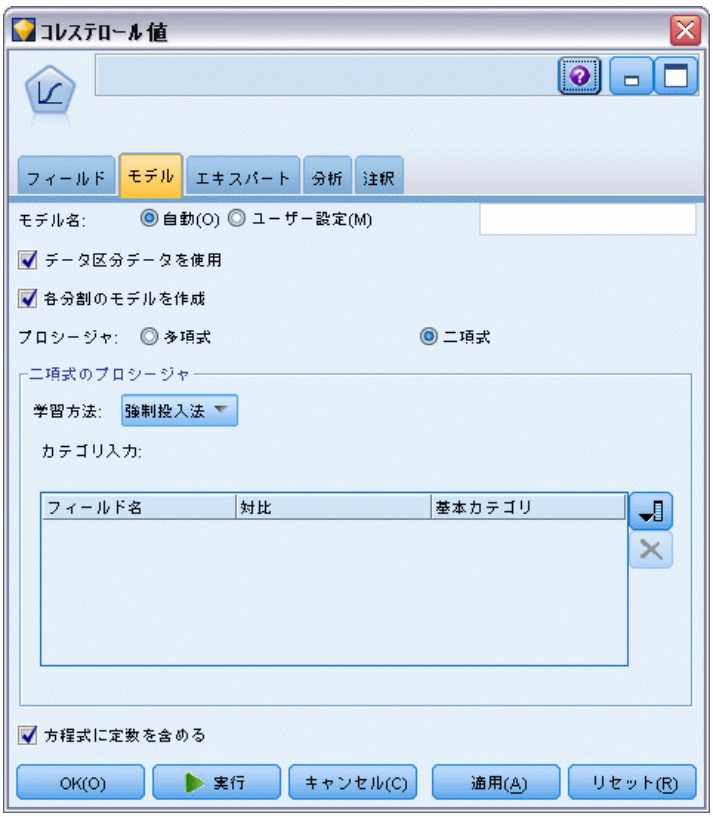
図 10-6
「Cholesterol」フィールドの短い文字列の詳細

▶ ロジスティック ノードをデータ型ノードに追加します。

▶ ロジスティック ノードで、［モデル］タブをクリックし、[**二項検定**] 手続きを選択します。

▶ 二項ロジスティック 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 10-7
手順に二項検定を選択



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なこの種類のストリームの詳細については、次の例を参照してください。

- 自動分類ノード。詳細は、 4 章 p. 49 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。
- 二項ロジスティック回帰。詳細は、 13 章 p. 179 電気通信会社の顧客の解約（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 の使用方法の詳細は、ユーザー ガイド、ノード リファレンス、アルゴリズム ガイドなど、インストール ディスクの Documentation ディレクトリで使用できます。

パート III:
モデル作成の例

章
11

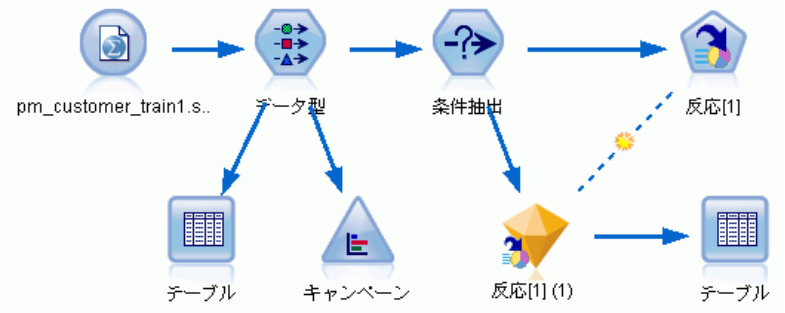
顧客レスポンスのモデル作成（ディシジョン リスト）

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顧客関係管理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、今後さらに収益を上げることを望んでいる 金融機関に基づいています。特に、この例では、ディシジョン リスト モデルを使用して、以前の販売促進を基に顧客が最も好意的な反応を示す特徴を識別し、その結果に基づいてメーリング リストを生成します。

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異なる手法には、自動分類ノードを代わりに使用することができます。

図 11-1
ディシジョン リストのサンプル ストリーム

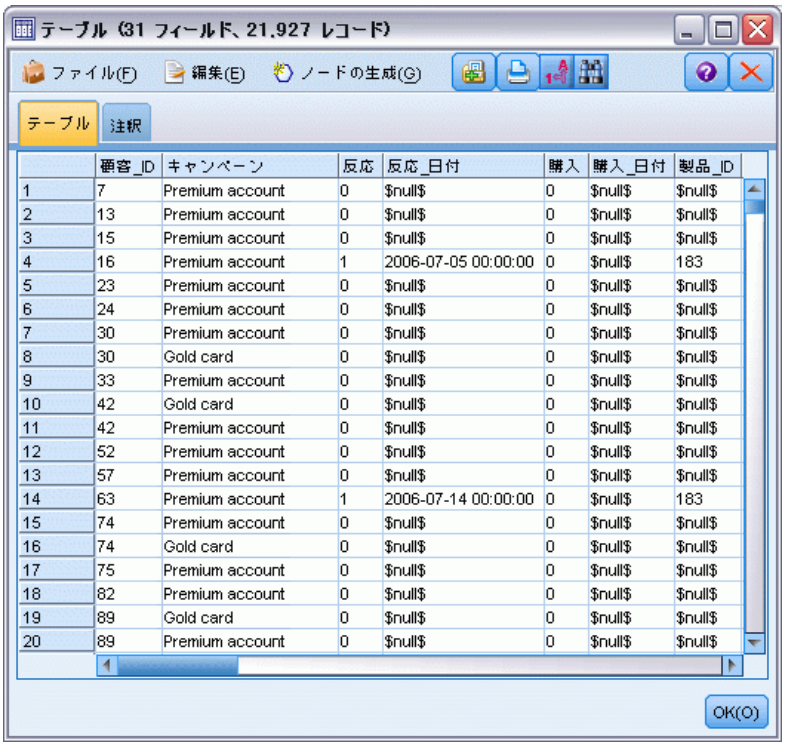


この例では、ストリーム pm_decisionlist.str を使用し、データファイル pm_customer_train1.sav を参照します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、Windows の［スタート］メニューの IBM® SPSS® Modeler プログラム グループからアクセスできます。pm_decisionlist.str ファイルは、streams ディレクトリにあります。

履歴データ

ファイル pm_customer_train1.sav には、campaign フィールドの値が示すとおり、過去のキャンペーンの特定の顧客に作成したオファーを記録す移籍する履歴データがあります。レコードの最大数は、Premium account キャンペーンに残されます。

図 11-2
以前のプロモーションに関するデータ



	顧客_ID	キャンペーン	反応	反応_日付	購入	購入_日付	製品_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 card	0	$null$	0	$null$	$null$
9	33	Premium account	0	$null$	0	$null$	$null$
10	42	Gold card	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 card	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 card	0	$null$	0	$null$	$null$
20	89	Premium account	0	$null$	0	$null$	$null$

campaign フィールドの値は、データ内に整数としてコード化され、データ型ノードに定義されたラベルが付けられます（たとえば 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。

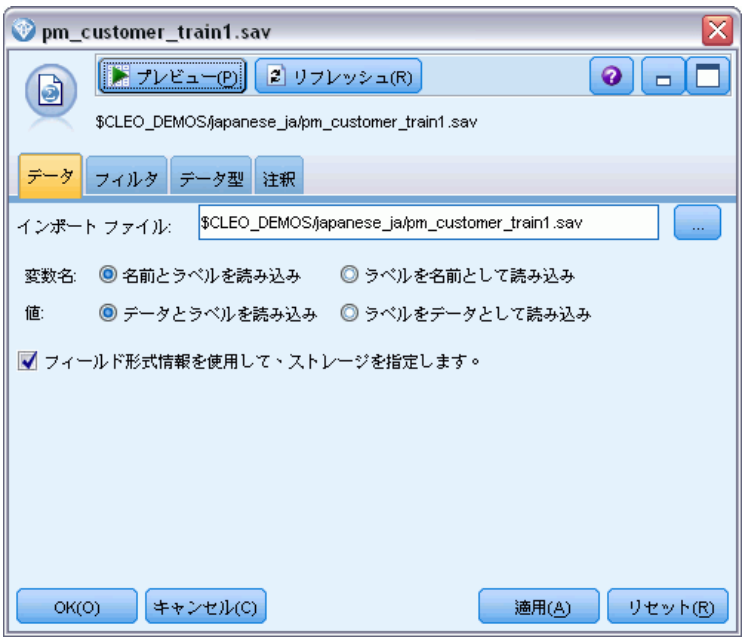
ファイルには、特定の特徴に基づいた異なるグループの回答率を予想するモデルを作成または「学習」するために使用することができる、各顧客についての人口統計および財務データを含む多くのフィールドが存在します。

ストリームの構築

- IBM® SPSS® Modeler インストール フォルダの Demos フォルダにある、pm_customer_train1.sav を示す Statistics ファイル ノードを追加しま

す。(ファイル パスの $CLEO_DEMOS/ を、このフォルダを参照するショートカットとして指定することができます。)

図 11-3
データの読み込み



▶ 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 = **対象**）。このフィールドの測定レベルを［**フラグ型**］に設定します。

図 11-4
測定レベルおよび役割の設定

▶ 次のフィールドについては役割を［**なし**］に設定します。customer_id、campaign、response_date、purchase、purchase_date、product_id、Rowid および X_random。これらのフィールドはすべてデータ内で使用しますが、実際のモデル作成には使用されません。

▶ データ型ノードの［**値の読み込み**］ボタンをクリックし、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ついての情報が含まれますが、分析は 1 度に 1 つのキャンペーンについて行われます。レコードの最大数は Premium campaign に残るため（データ内で campaign = 2 とコード

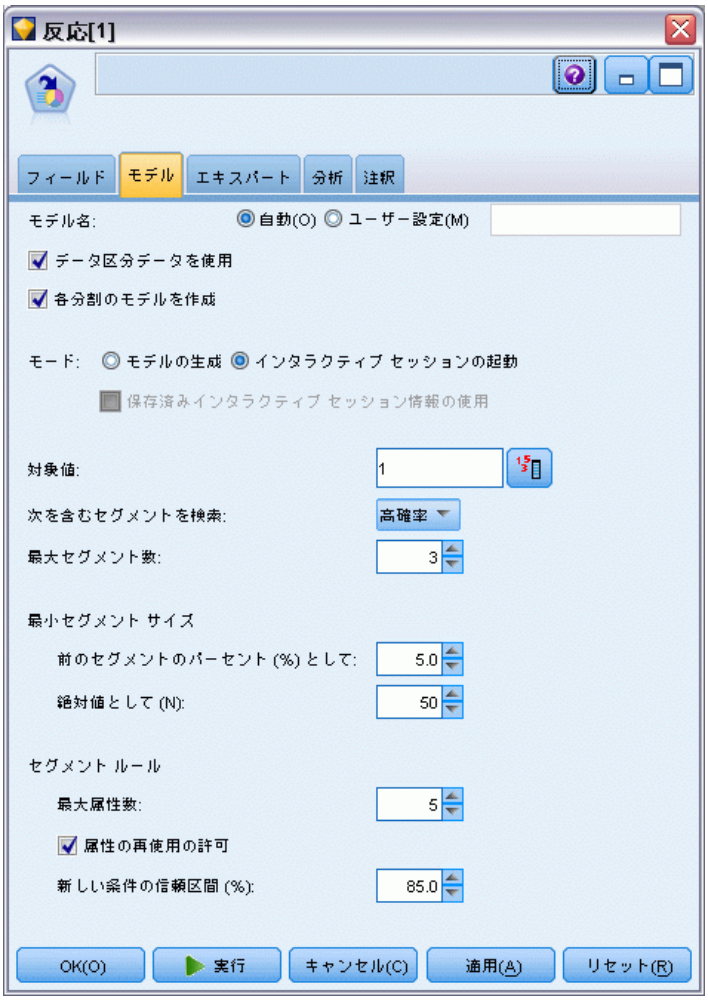
化)、条件抽出ノードを使用して、ストリームにこれらのレコードのみを選択することができます。

図 11-5
単一キャンペーンのレコードの選択

モデルの作成

▶ 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 11-6
ディシジョン リスト ノードの [モデル] タブ



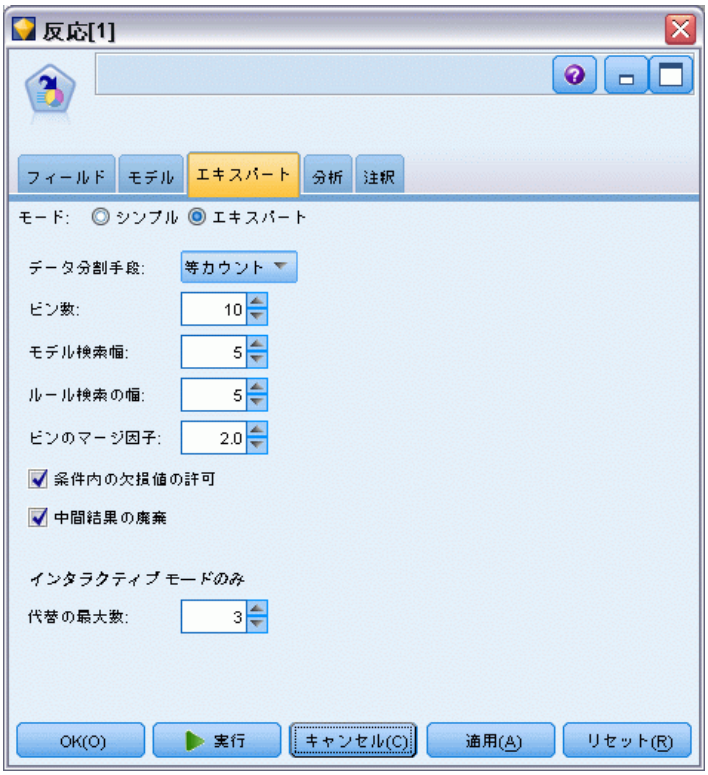
▶ [**インタラクティブ セッションの起動**] を選択します。

▶ この例でモデルを簡単にするため、セグメントの最大値を 3 に設定します。

▶ 新しい条件の信頼区間を 85% に変更します。

▶ ［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。

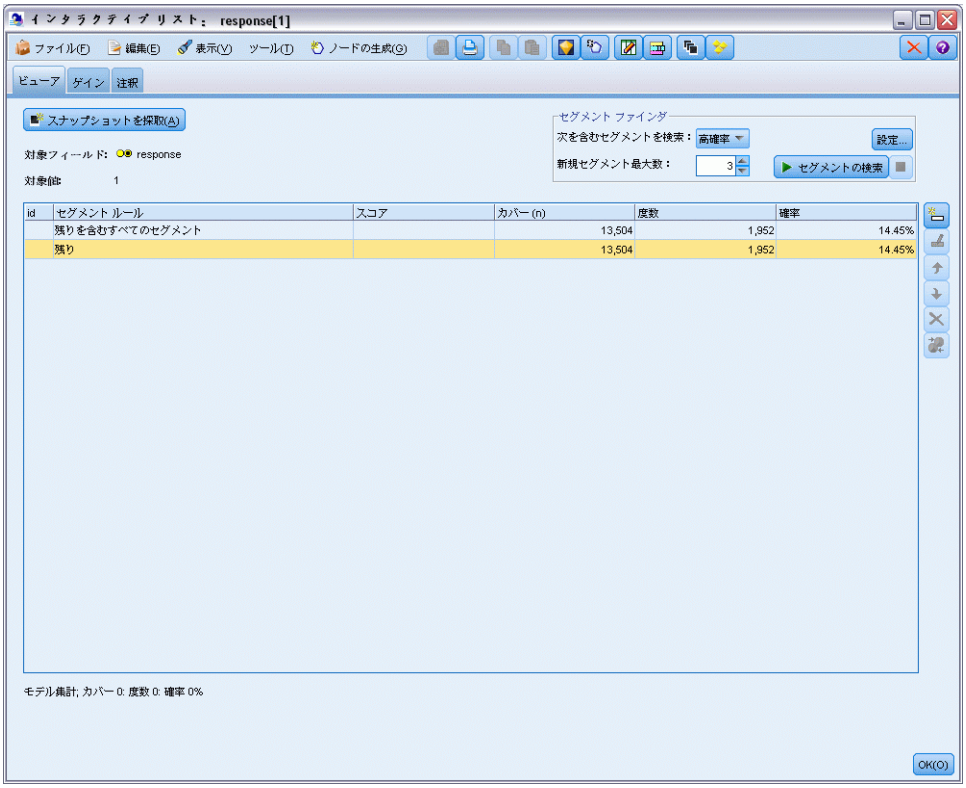
図 11-7
ディシジョン リスト ノードの [エクスパート] タブ



▶ [**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。

▶ [**実行**] をクリックし、Interactive List Viewer を表示します。

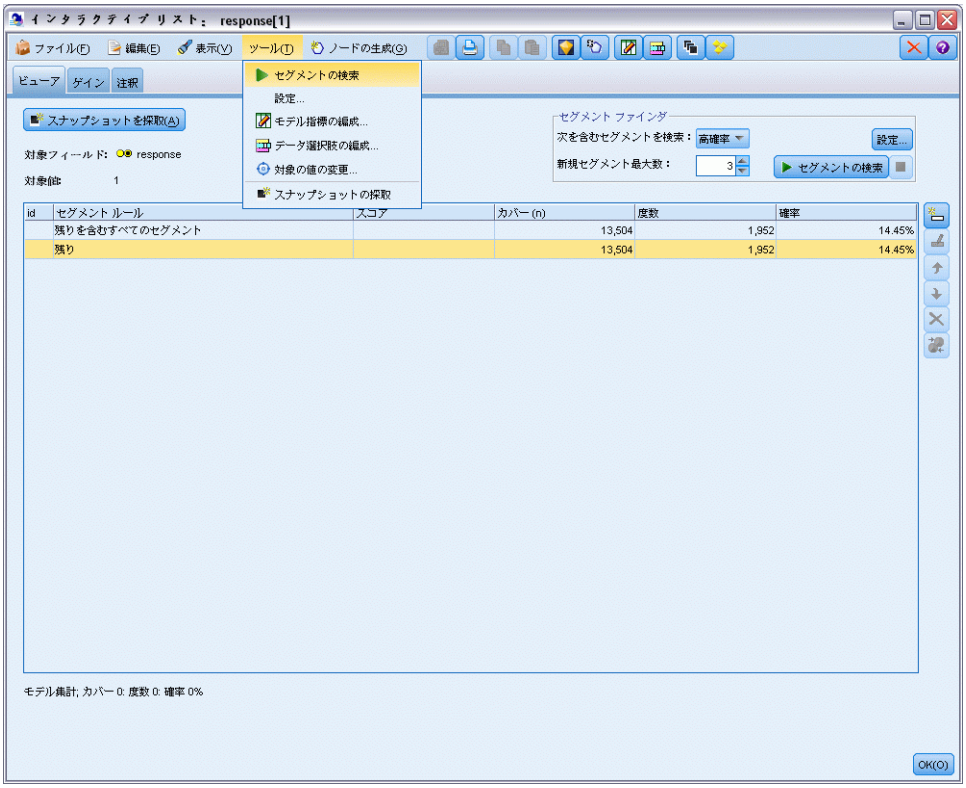
図 11-8
Interactive List Viewer



セグメントがまだ定義されていないために、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 14.45% の 1,952 がはいと回答しています。この割合を向上させるには、有望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。

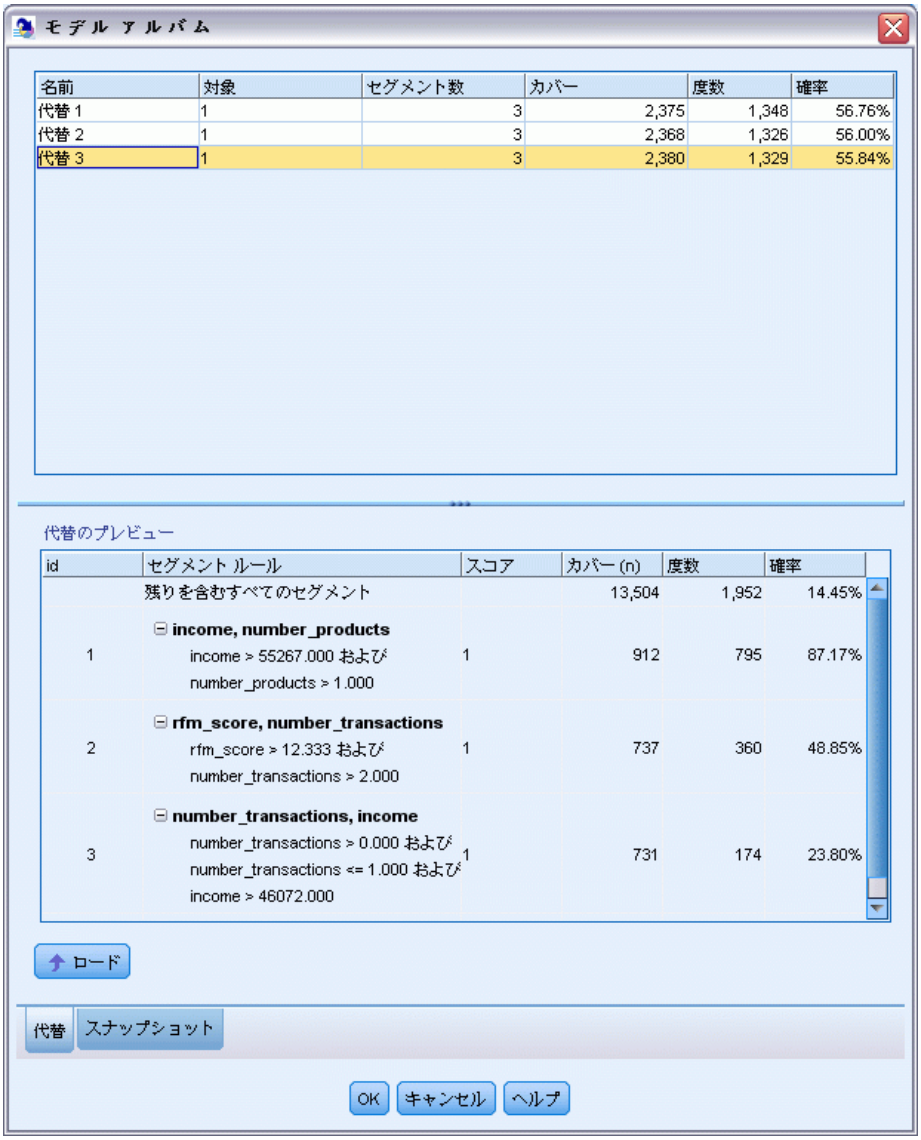
▶ Interactive List Viewer で、メニューから次の項目を選択します。
ツール > [セグメントの検索]

図 11-9
Interactive List Viewer



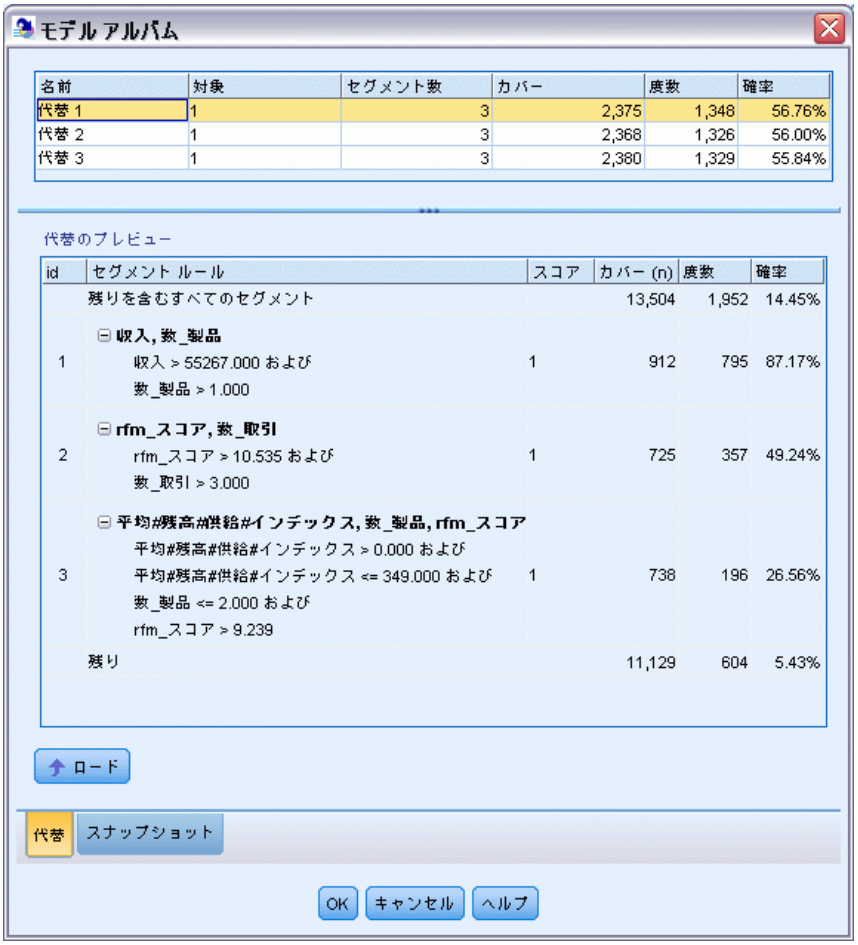
こうすると、ディシジョン リスト ノードで指定した設定に基づいて、デフォルトのマイニング タスクが実行されます。完了したタスクは 3 つの代替モデルを戻し、それらのモデルは、［モデル アルバム］ダイアログ ボックスの［代替］タブに一覧されます。

図 11-10
使用可能な代替モデル



▶ リストから最初の代替モデルを選択します。その詳細は、代替プレビュー領域に表示されます。

図 11-11
選択された代替モデル



代替プレビュー領域を使用して、作業モデルを変更することなくあらゆる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調査することができます。

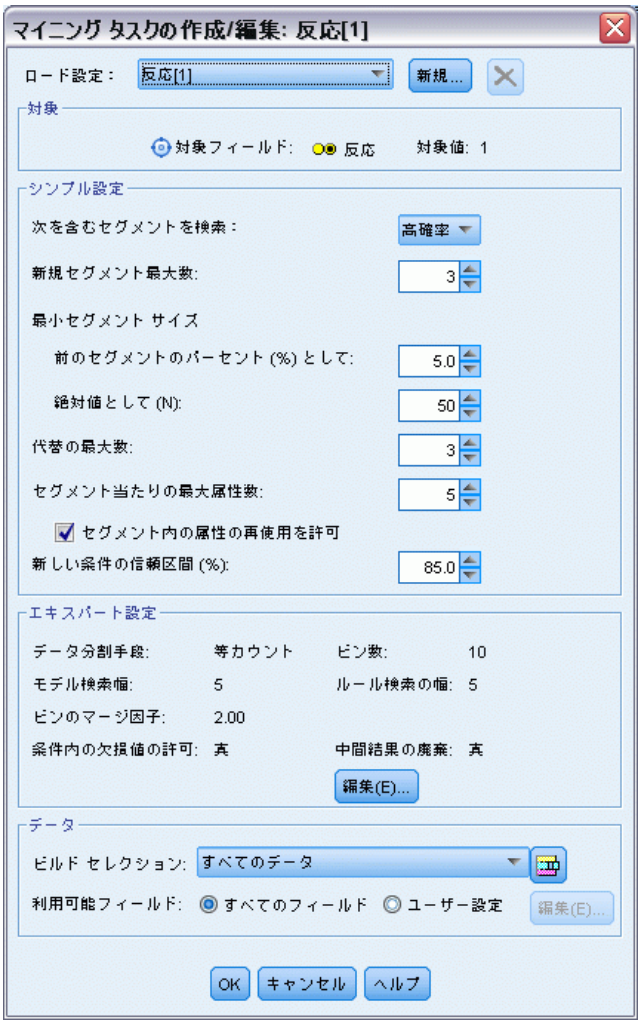
注 ： モデルをより詳細に表示するには、ここで示すとおりダイアログ内の代替プレビュー領域を最大化します。領域の境界線をドラッグすると最大化します。

収入、月ごとの取引数、RFM スコアなどの予測フィールド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、全体サンプルの小さ

な部分をカバーするに過ぎないために、数百件のヒットがある 11,000 件以上のレコードを残りの部分として残してしまいます。成績の悪いもの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。

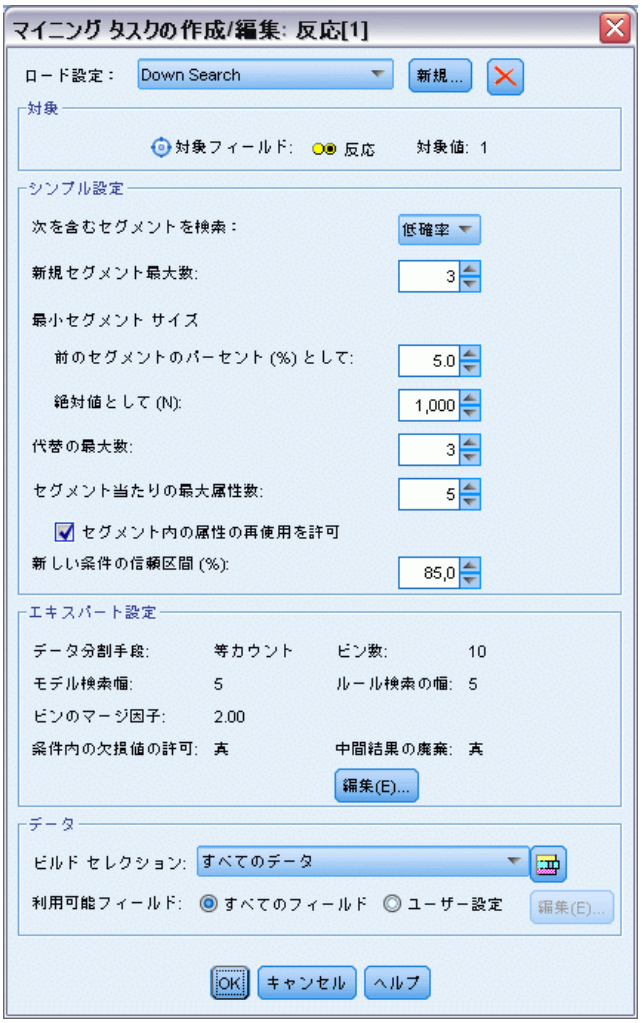
▶ 異なるモデル作成の方法を試すには、メニューから次の項目を選択します。
ツール > 設定

図 11-12
[マイニング タスクの作成/編集] ダイアログ ボックス



▶ [新規] ボタン（右上）をクリックしてマイニング タスクを作成し、[新規設定] ダイアログでタスク名として［下方検索］を指定します。

図 11-13
[マイニング タスクの作成/編集] ダイアログ ボックス



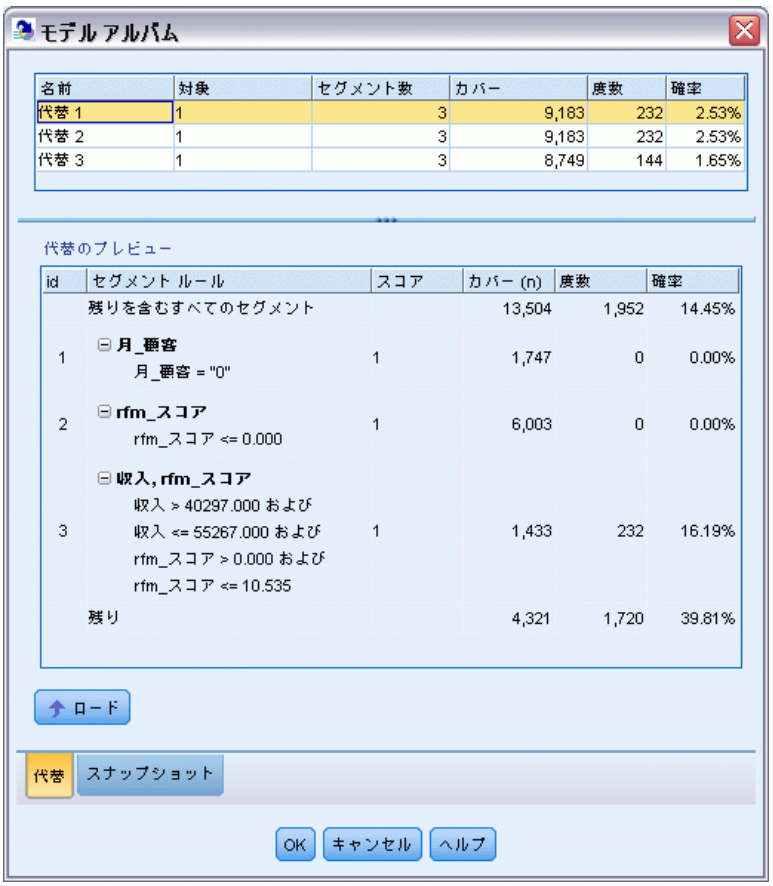
- タスクの検索方向を [低い確率] に変更します。こうすると、アルゴリズムは、最高ではなく「最小」レスポンスのセグメントを検索します。
- 最小セグメント サイズを 1,000 まで拡大します。[OK] をクリックし、Interactive List Viewer に戻ります。

▶ Interactive List Viewer で、[セグメント ファインダ] 領域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 11-14
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戻され、それらは [モデル アルバム] ダイアログ ボックスの [代替] タブに追加され、以前の結果と同じ方法でプレビューできます。

図 11-15
[下方向検索] モデルの結果

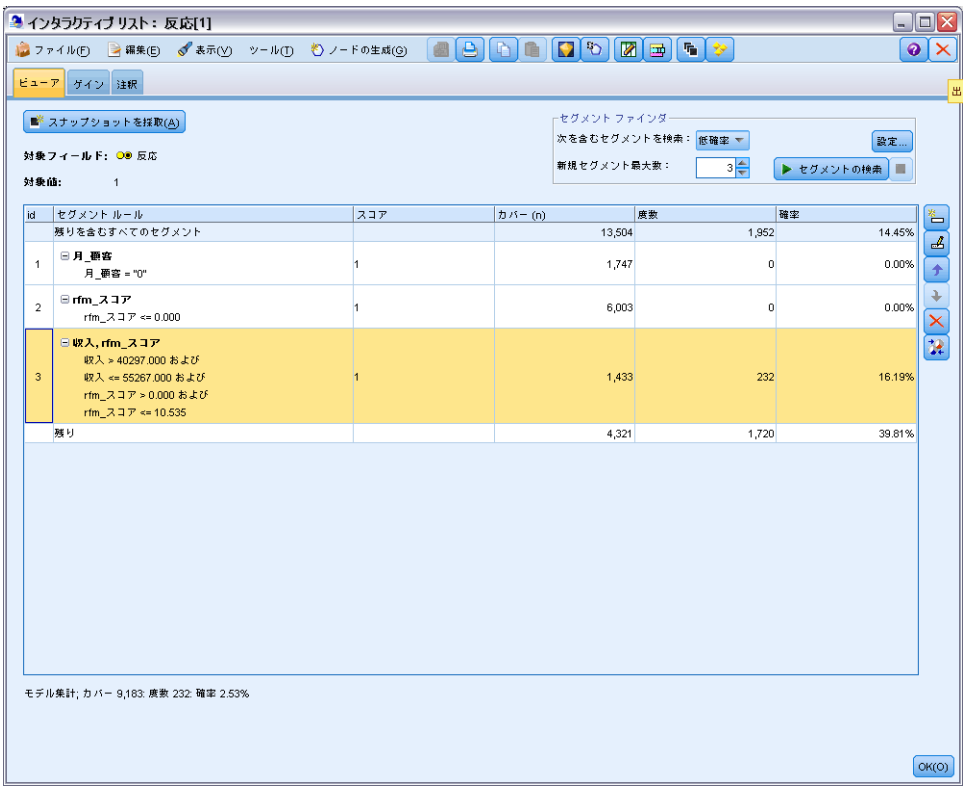


今回は、各モデルで、高いレスポンスではなく低いレスポンス確率のセグメントが特定されました。最初の代替モデルを調べると、これらのセグメントを単純に除外するだけで、残りの部分のヒット率が 39.81% に増加します。これは、以前に調べたモデルより低いものですが、カバー率はより高くなっています（つまり合計ヒットがより高い）。

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、その後高い確率の検索を行って 2 つのアプローチを結合させることで、この結果を改善させることができます。

▶ [**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 11-16
セグメントの除外



▶ 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、セグメント間でヒットがゼロのレコードをおよそ 8,000 取得し、予測されるオファーからそれらのセグメントを除外します。（除外されたセグメントは、それを示すためにヌルと得点がつけられます。）

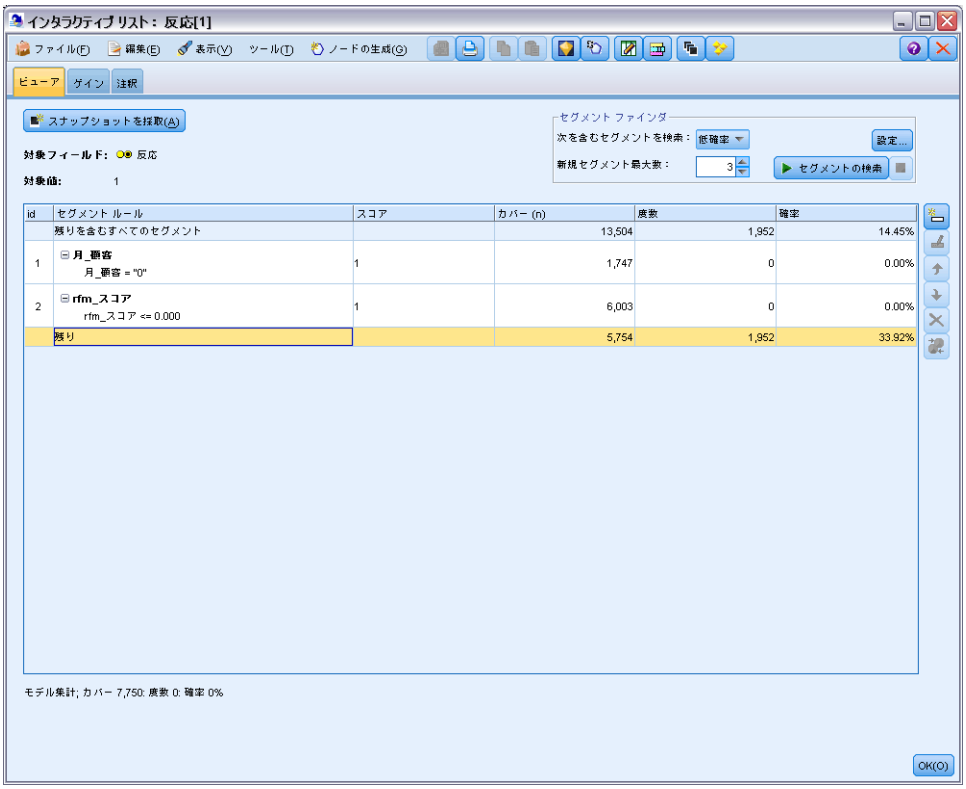
▶ 3 番目のセグメントを右クリックして、[**セグメントの削除**] を選択します。16.19% という、このセグメントのヒット率は、14.45% の基準率と差がないために、これを留めておくことを正当化するに足る情報は追加されません。

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、単にそれに得点をつける方法が変更されますが、セグメントの削除では、モデルから完全に削除します。

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ます。

▶ 次のマイニング タスクが残りに部分だけに適用されるように、テーブルの残りの部分の行をクリックして選択します。

図 11-17
セグメントの選択



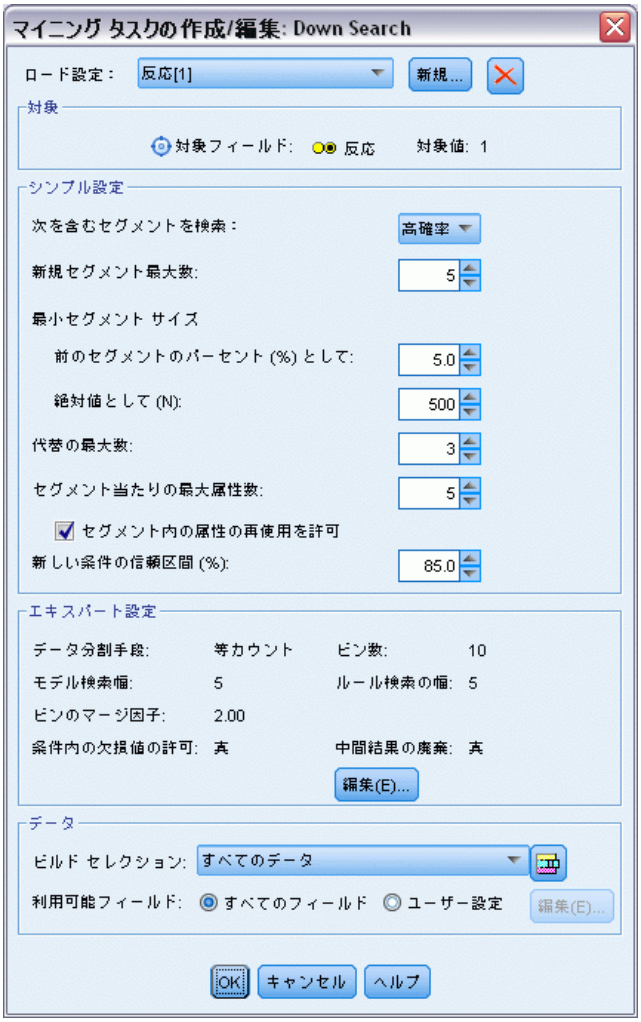
▶ 残りの部分を選択し、[**設定**] をクリックして［マイニング タスクの作成/編集］ダイアログ ボックスを再度開きます。

▶ 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。

▶ [シンプル設定] を編集して、新しいセグメント数を 5 まで増やし、最小セグメント サイズを 500 にします。

▶ [OK] をクリックし、Interactive List Viewer に戻ります。

図 11-18
デフォルトのマイニング タスクの選択

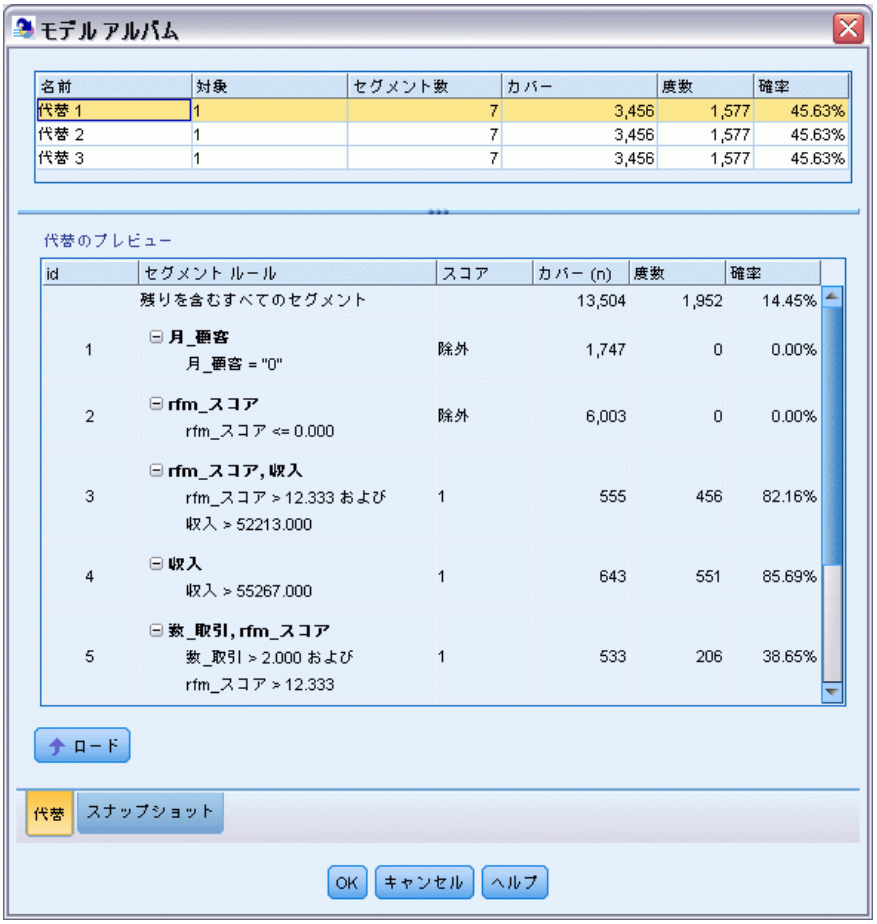


▶ [セグメントの検索] をクリックします。

こうすることで、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あうようになります。回答率が低いセグメントは除外され、Null としてスコアリグされます。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、前のモデルより高いカバー率（3,456 レコードのうち 1,577 ヒット）となります。

図 11-19
結合されてモデルの代替

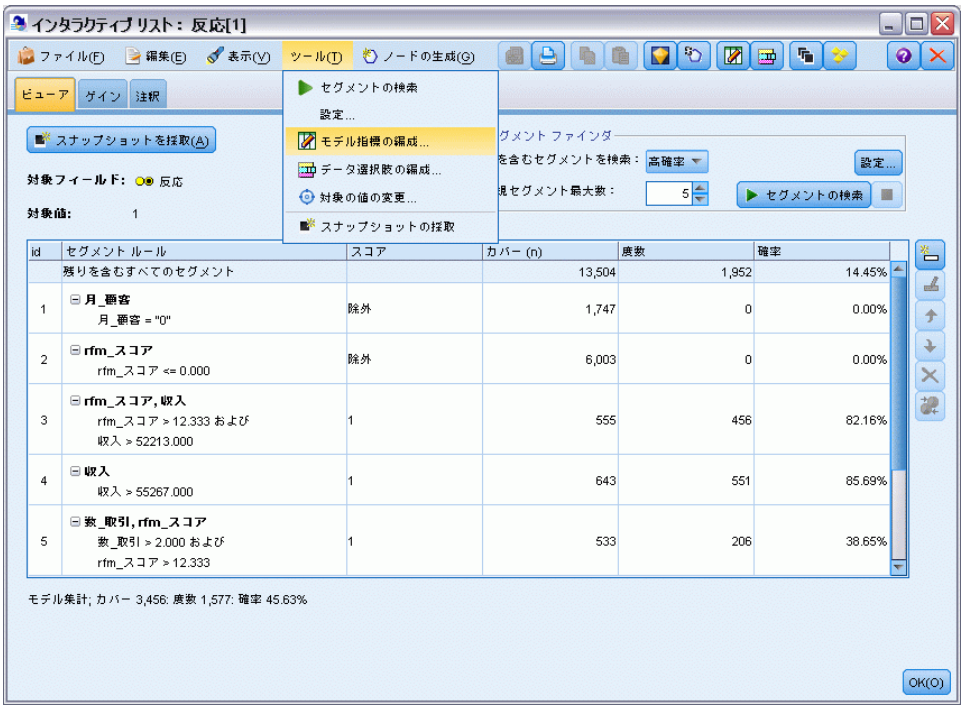


▶ 最初の代替モデルをプレビューして [**ロード**] をクリックして、プレビューしたモデルを作業モデルにします。

Excel を使用したカスタム指標の計算

▶ 実際の問題でモデルがどのように実行されているかについての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 11-20
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使用を計算するかどうかを指

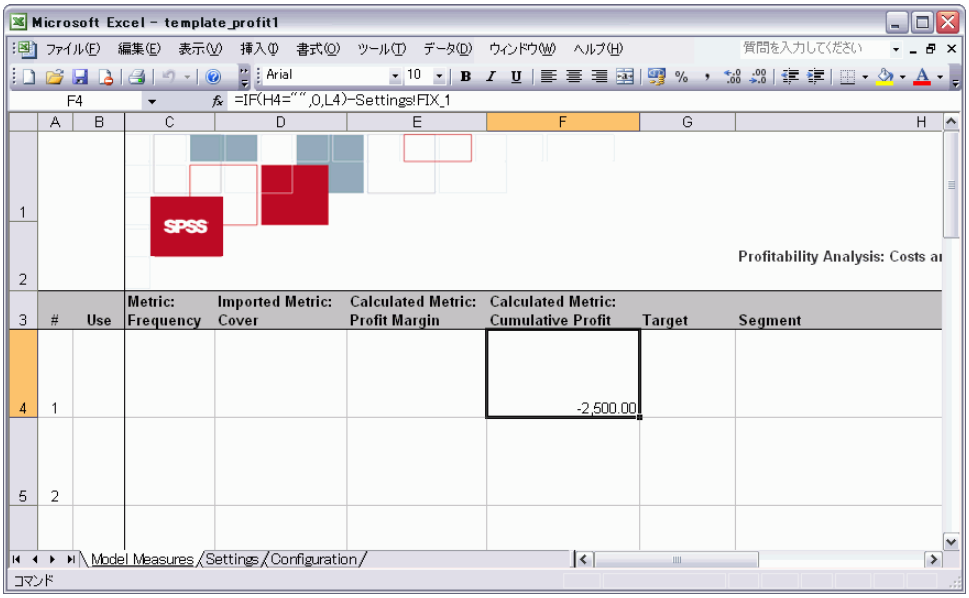
定することができます。さらに状況に応じて、数ではなく円グラフを表示することができます。

図 11-21
[モデル指標の編成] ダイアログ ボックス

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することができます。

- ［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。
- **［Excel (TM) へ接続］** をクリックします。
- IBM® SPSS® Modeler インストール ディレクトリの Demos フォルダの 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 11-22
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、カスタム指標を計算した後エクスポートしてモデルに戻します。
- [**設定**] にはカスタム指標の計算に使用されるパラメータが含まれています。
- [**構成**] では、モデルからインポートしモデルにエクスポートするメトリックを定義します。

モデルにエクスポートされるメトリックは次のとおりです。

- **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。

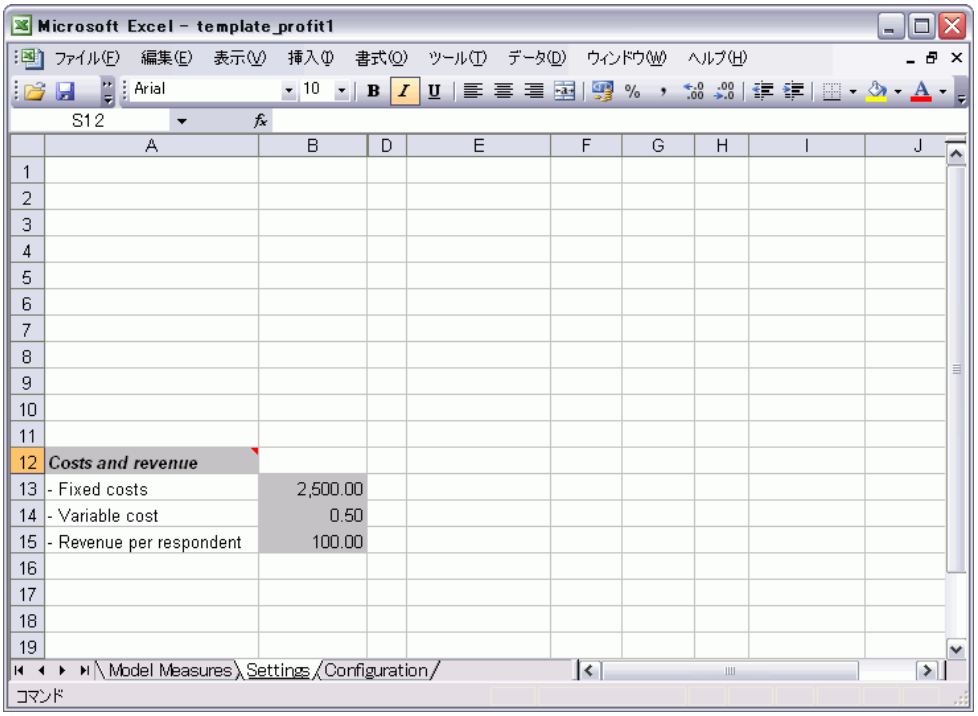
Profit Margin = Frequency * Revenue per respondent - Cover * Variable cost

Cumulative Profit = Total Profit Margin - Fixed cost

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は［設定］ワークシートでユーザーに定義されます。

図 11-23
Excel の［設定］ワークシート



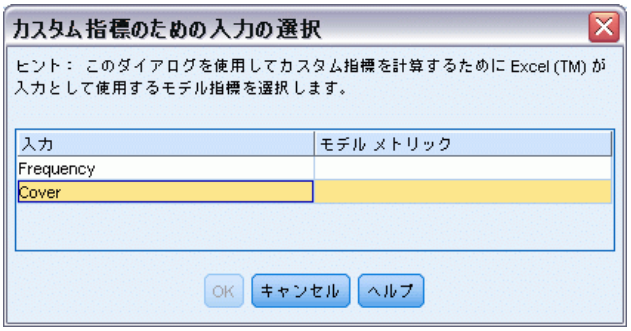
固定コスト は、設計または計画などキャンペーンに設定されたコストです。

変動コストは、封筒や切手など、各顧客にオファーを拡張したコストです。

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。

▶ 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 + Tab キー）Interactive List Viewer に再度移動します。

図 11-24
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、モデルからテンプレートに定義された特定のパラメータに入力を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を表示し、右側の列には、［構成］ワークシートで定義されたように使用をスプレッドシートパラメータにマッピングします。

▶ **［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名「度数」および「カバー (n)」は入力に一致しますが、異なる名前も使用することができます。

▶ ［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 11-25
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メトリック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 11-26
Interactive List Viewer に表示されるカスタム指標

Excel テンプレートを編集すると、多く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 はデフォルトの Excel テンプレートで提供され Interactive List Viewer と共に使用されますが、設定を変更したり独自のテンプレートを変更する必要があることがあります。たとえば、テンプレートのコストは、編成には不適切で、修正の必要があります。

注 ： 既存のテンプレートを変更または独自のテンプレートを作成する場合、.xlt の接尾辞を使用して Excel 2003 でファイルを保存します。

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、新しい図で Interactive List Viewer を更新する手順は次のとおりです。

▶ Interactive List Viewer で、[ツール] メニューから **[モデル指標を編成]** を選択します。

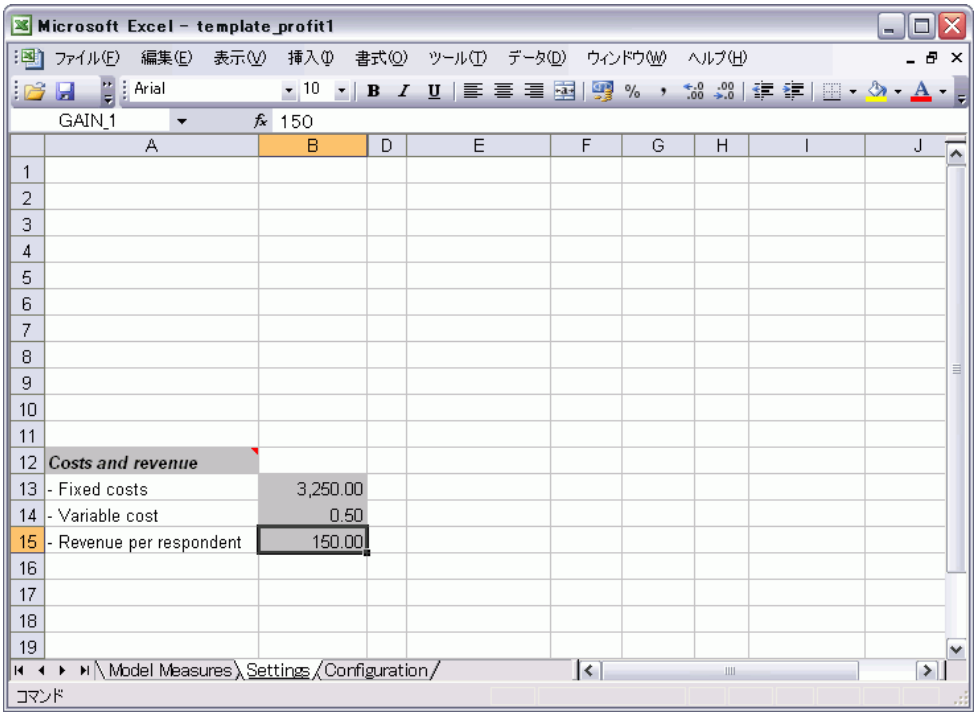
▶ ［モデル指標の編成］ダイアログボックスで、[Excel への接続™] をクリックします。

▶ 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。

▶ ［設定］ワークシートを選択します。

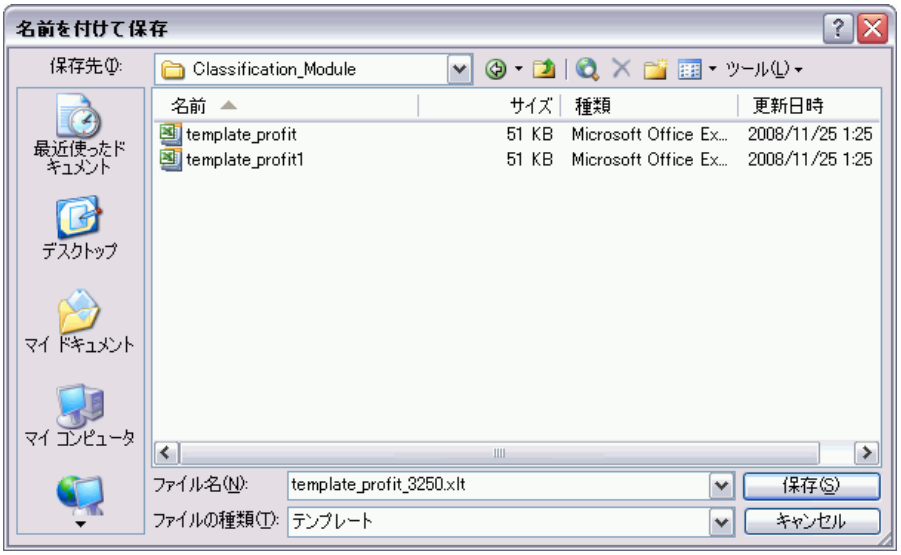
▶ [固定コスト] を 3,250.00 に、[回答者あたりの利益] を 150.00 に編集します。

図 11-27
Excel の［設定］ワークシートの変更された値



▶ 一意で、関連したファイル名で変更されたテンプレートを保存します。Excel 2003 の .xlt 拡張子が使用されていることを確認します。

図 11-28
変更された Excel テンプレートの保存



▶ Windows タスクバーを使用して（または Ait + 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て [OK] をクリックします。

▶ [モデル指標の編成] ダイアログ ボックスで [OK] をクリックし、Interactive List Viewer を更新します。

明らかに、この例では Excel テンプレートの変更のシンプルな 1 つの方法のみ示しています。Interactive List Viewer とデータを受け渡すより詳細な変更を行ったり、Excel を使用してグラフなどの出力を作成することができます。

図 11-29
Interactive List Viewer に表示される変更されたカスタム指標

結果の保存

インタラクティブ セッションの中で後で使用できるよう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ることができます。スナップショットは、[スナップショット] タブに一覧されます。インタラクティブ セッションの中では、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。

この方法を継続することで、追加のセグメントを検索するために、追加のマイニング タスクで実験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、その他の多数の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含有または除外して、それぞれに得点をつける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、得点記録の目的で展開するモデルを生成できます。

また、後日使用するためにインタラクティブ セッションの現在状態を保存するために、[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。こうすることで、マイニング タスク、モデルのスナップショット、データ選択、およびカスタム指標を含めて、ディシジョン リストのモデリング 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 モデル作成ノードの中で **[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ます。 詳細は、 9 章 ディシジョン リスト in IBM SPSS Modeler 15 モデル作成ノード を参照してください。

章
12

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、数値型フィールドではなくカテゴリ フィールドを対象フィールドとします。

たとえば、電気通信プロバイダーがその顧客を、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ライズしているとします。顧客がどのグループに属するかを、人口統計データを使って予測できれば、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_custcat.str という名前のストリームを使用します。このストリームは、telco.sav という名前のデータ ファイルを参照します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、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フォーカスします。対象フィールド custcat は、次のように 4 つの顧客グループに対応して 4 つの値を取ります。

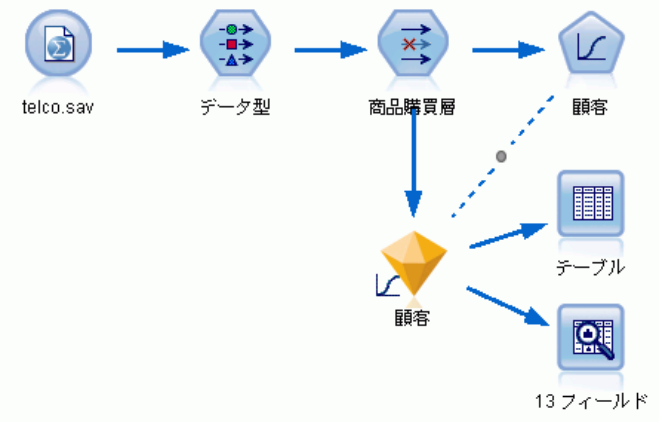
値	Label
1	基本サービス
2	E-サービス
3	プラス サービス
4	トータル サービス

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。 詳細は、 13 章 p. 179 電気通信会社の顧客の解約（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

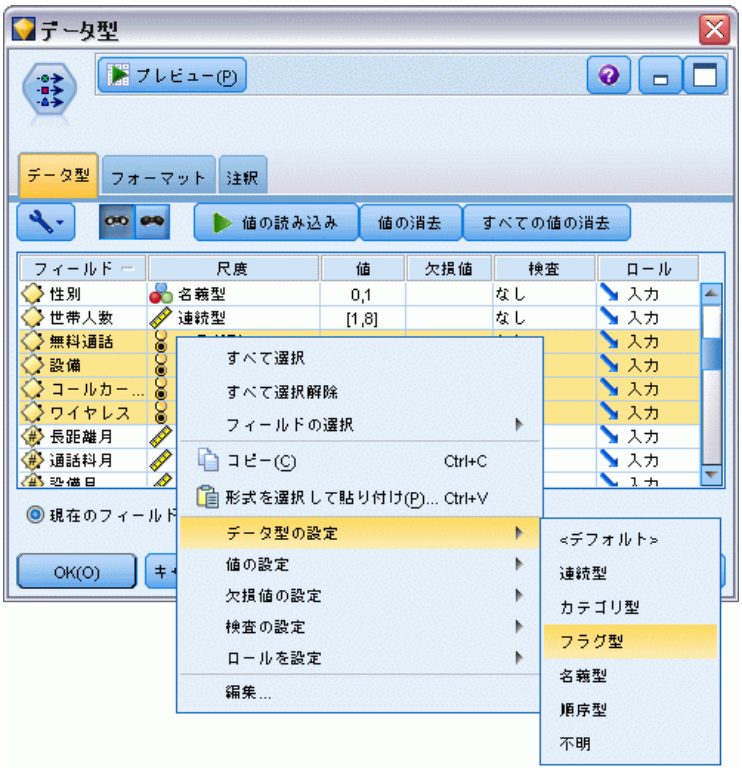
- telco.sav を指定するStatistics ファイル入力ノードを Demos フォルダに追加します。

図 12-1
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



▶ データ型ノードを追加して［**値の読み込み**］をクリックし、すべての測定レベルが正しく設定されているか確認します。たとえば、値が 0 と 1 の多くのフィールドは整数してフラグが付きます。

図 12-2
複数のフィールドの測定レベル設定



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プロパティを変更するには、［値］列のヘッダをクリックしてフィールドを値によってソートし、Shift キーを押したままマウスまたは矢印キーを使って、変更するフィールドをすべて選択します。その後、選択したフィールドの上で右クリックをすると、選択したフィールドの測定レベルまたは属性を変更することができます。

性別はフラグではなく値が 2 つのフィールドとしてより正確に認識されるため、尺度値をは［**名義型**］のままになります。

► 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 12-3
フィールドの役割の設定

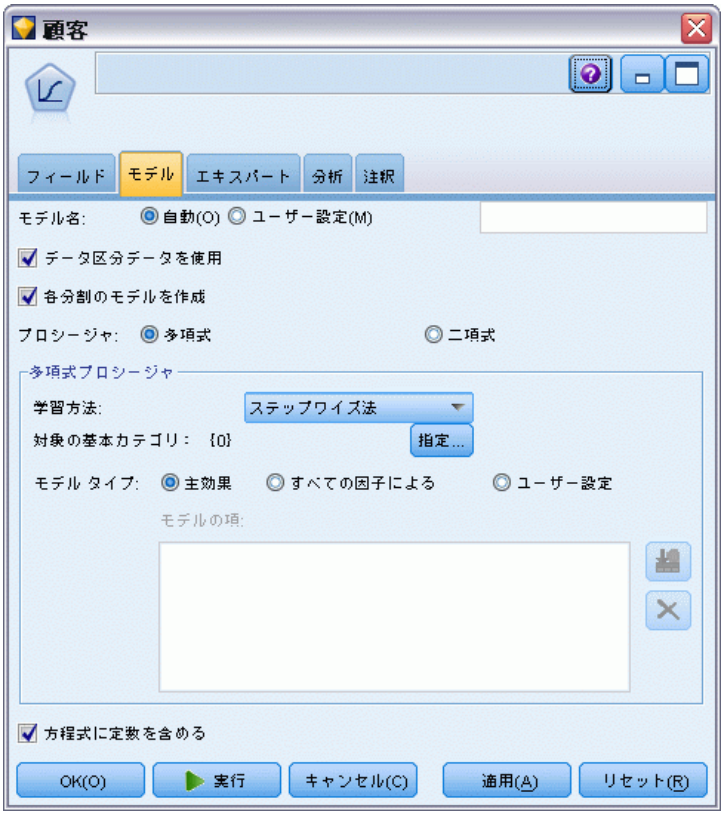
例では人口統計に焦点を当てているため、フィルター ノードを使用して関連フィールドのみを選択します（地域、年齢、 結婚、住所、収入、学歴、職業，退職、性別、居住および custcat）。他のフィールドは、この分析のために除外されることがあります。

図 12-4
人口統計フィールドのフィルタリング

（代わりに、これらのフィールドの役割を除外するのではなく**なし**に変更するか、あるいはモデル作成ノードで使用したいフィールドを選択できます。）

▶ 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] の方法を選択します。[多項式]、[主効果]、および [方程式に定数を含む] も選択します。

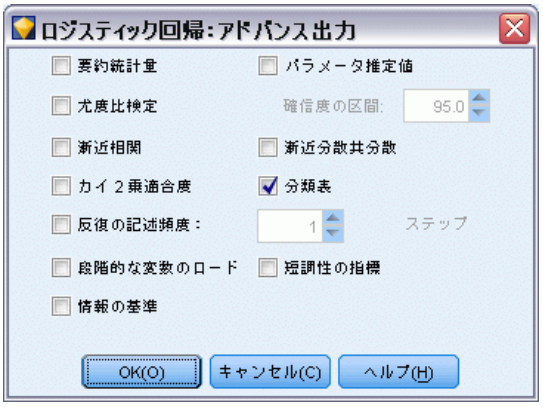
図 12-5
モデル オプションの選択



対象のベース カテゴリを 1 とします。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。

▶ ［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、**[分類表]** を選択します。

図 12-6
出力オプションの選択



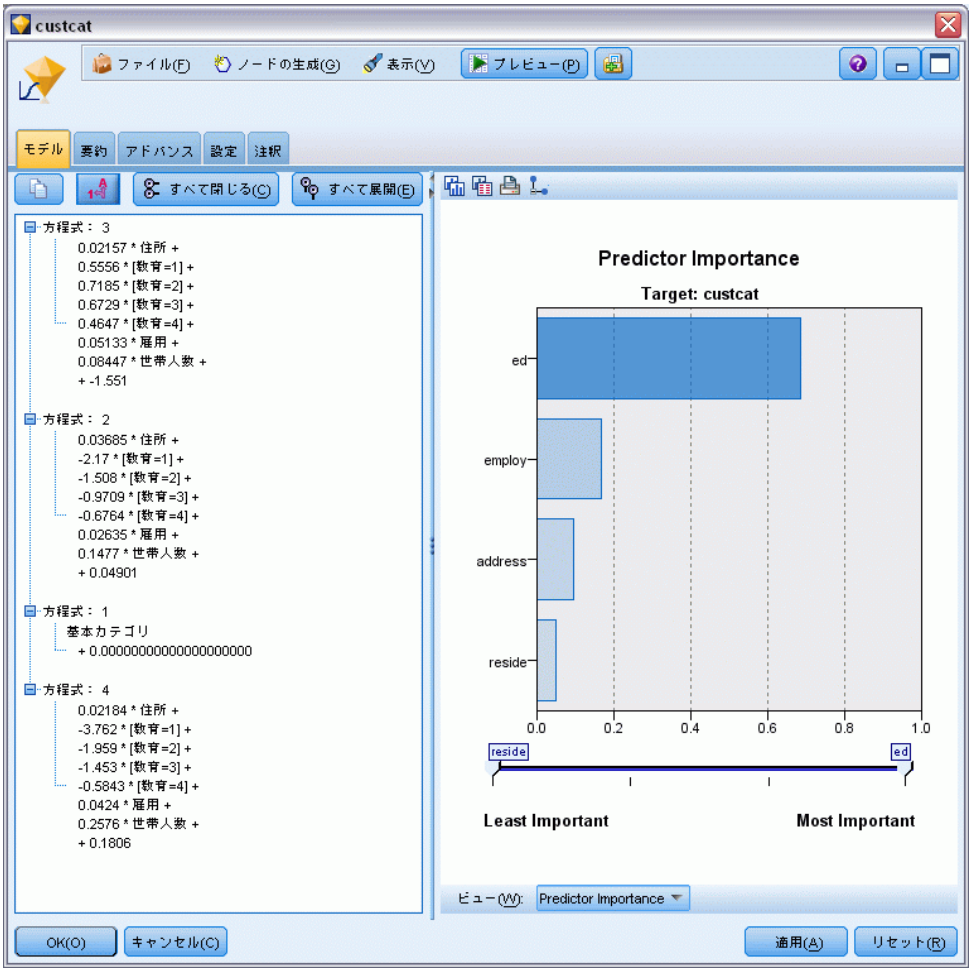
モデルの参照

▶ ノードを実行してモデルを生成します。これは右上角のモデル パレットに追加されます。この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 カテゴリです。

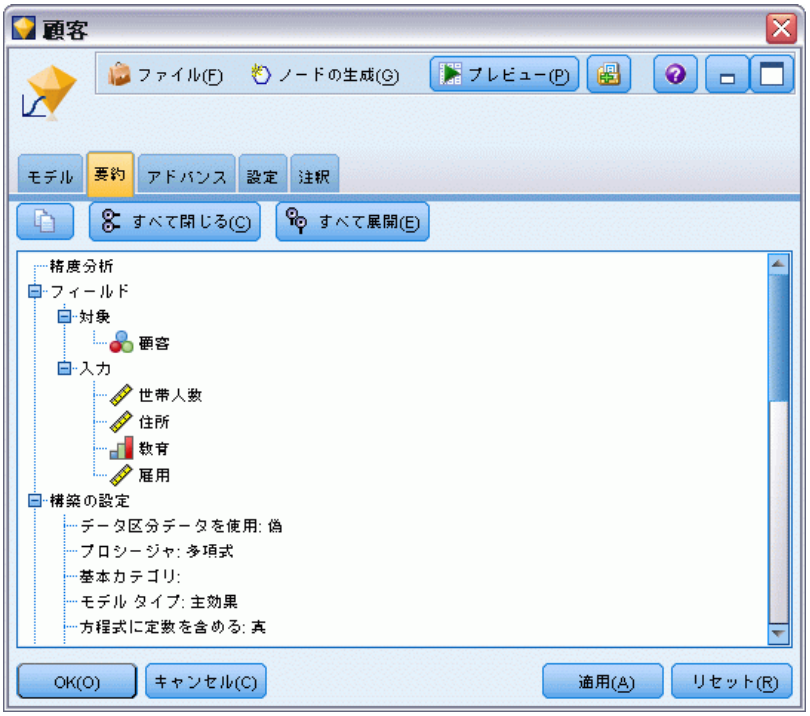
カテゴリ 3 がプラス サービスなどを表すなど、残りの 3 つの方程式についての詳細が表示されています。

図 12-7
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、考慮のために提出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が存在します。

図 12-8
対象および入力フィールドを表示するモデルの要約

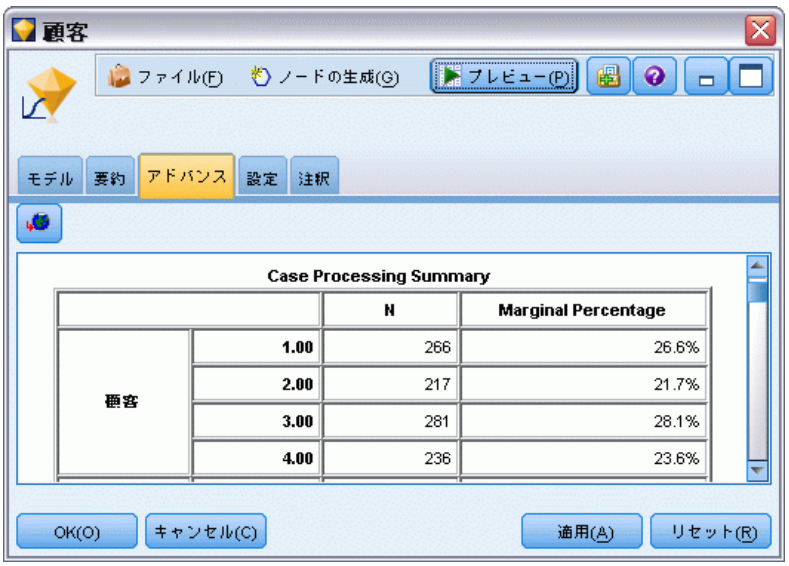


［詳細］タブに表示される項目は、モデリング ノードの［詳細出力］ダイアログ ボックスで選択されたオプションによって異なります。

1 つの項目は、［処理したケースの要約］に常に表示されます。これは、対象フィールドの各カテゴリに該当するレコードのパーセンテージを表します。これにより、比較の基礎として使用するヌル モデルが与えられます。

予測値を使用したモデルを構築しない場合、最良の推測は、1（プラス サービス）というもっとも一般的なグループにすべてを割り当てることです。

図 12-9
ケース処理要約(S)



学習データに基づいて、すべての顧客をヌル モデルに割り当てた場合は、時間の 281/1000 = 28.1% 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。それから、予測をヌル 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検討します。

［詳細］タブの一番下の分類テーブルに、モデルの結果が表示されます。これは時間の 39.9% 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 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 2）を識別する上では非常に劣っています。カテゴリ 2 の顧客に関する精度を向上する場合は、それらを識別するためのほかの予測値を見つける必要があります。

図 12-10
分類表

予測したいものによっては、このモデルが必要なものに完全にふさわしいものになっていることがあります。たとえば、カテゴリ 2 の顧客を識別することを重視しない場合は、このモデルは十分に正確なことがあります。これは、E-サービスがあまり利益をもたらさない特売サービスである場合です。

たとえば、カテゴリ 3 または 4 に当てはまる顧客から投資利益の最大部分を得ている場合は、このモデルから必要な情報が得られています。

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には、モデルを構築するときに、多数の診断方法が［詳細出力］ダイアログ ボックスで使用可能です。詳細は、 10 章 ロジスティック モデル ナゲットの詳細出力 in IBM SPSS Modeler 15 モデル作成ノード を参照してください。IBM® SPSS® Modelerで使用するモデリング メソッドの数学的基礎の説明は、インストール ディスクの ¥Documentation ディレクトリにもある SPSS Modeler 『アルゴリズム ガイド』に記載されています。

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をどれだけうまく一般化しているかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。 詳細は、 4 章 データ区分ノード

in IBM SPSS Modeler 15 入力ノード、プロセス ノード、出力ノード を参照してください。

章
13

電気通信会社の顧客の解約（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、数値型フィールドではなくカテゴリ フィールドを対象フィールドとします。

この例では、telco_churn.str という名前のストリームを使用します。このストリームは、telco.sav という名前のデータ ファイルを参照します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、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 ディレクトリにあります。

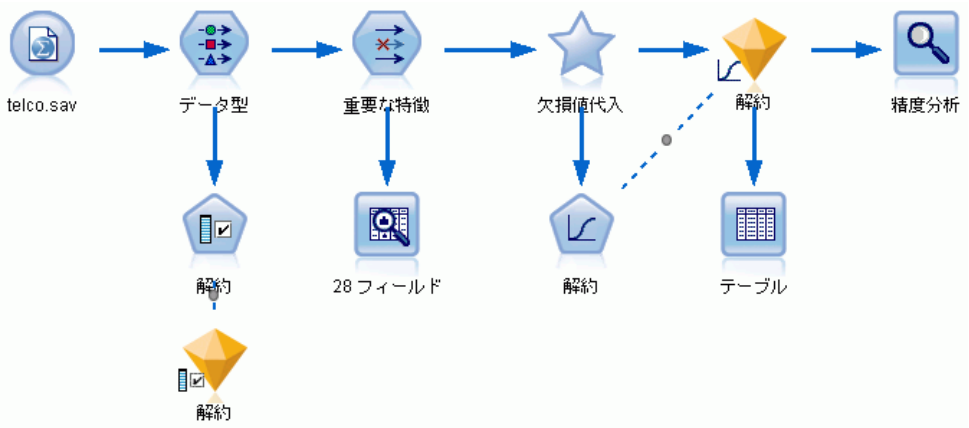
たとえば、競合他社に奪われる顧客に数に関して、電気通信プロバイダーが心配しているとします。サービス使用量データを、どの顧客が他のプロバイダーに移りそうかを予測するために使用できれば、オファーをカスタマイズして、できるだけ多くの顧客を保持することができます。

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フォーカスします。対象には 2 つの明確なカテゴリがあるために、2 項検定モデルを使用します。複数のカテゴリのある対象の場合は、代わりに多項検定モデルを作成します。 詳細は、 12 章 p.167 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

- telco.sav を指定するStatistics ファイル入力ノードを Demos フォルダに追加します。

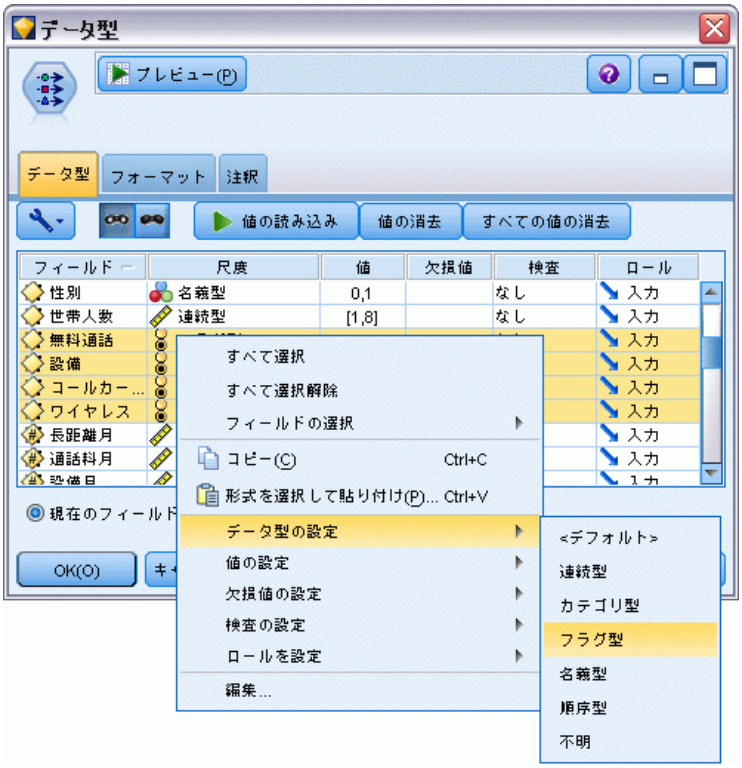
図 13-1
2 項検定ロジスティック回帰を使用して顧客を分類するためのサンプル ストリーム



- データ型ノードを追加してフィールドを定義し、すべての測定レベルが正しく設定されているか確認します。たとえば、値 0 や 1 を持つ多くの

フィールドはフラグとして認識されますが、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としてより正確に認識されます。

図 13-2
複数のフィールドの測定レベル設定



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プロパティを変更するには、[値] 列のヘッダをクリックしてフィールドを値によってソートし、Shift キーを押したままマウスまたは矢印キーを使って、変更するフィールドをすべて選択します。その後、選択したフィールドの上で右クリックをすると、選択したフィールドの測定レベルまたは属性を変更することができます。

▶ ［解約］フィールドの測定レベルを［**フラグ型**］に、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 13-3
解約フィールドの測定レベルおよび役割の設定

▶ フィールド選択モデリング ノードをデータ型に追加します。

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。

▶ ストリームを実行する。

▶ モデル ナゲットを開き、[生成] メニューから [フィルタ] を選択してフィルタ ノードを作成します。

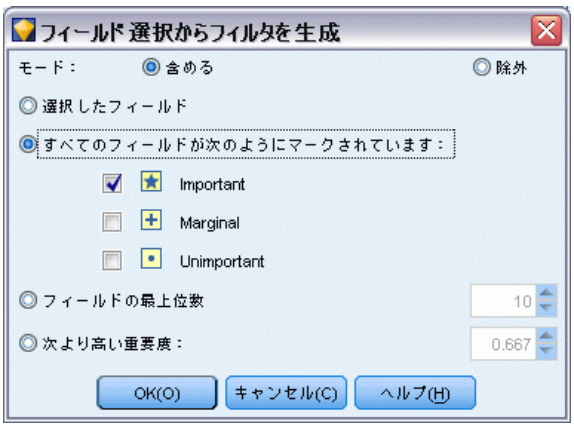
図 13-4
フィールド選択ノードからのフィルタ ノードの生成

telco.sav の中のすべてのデータが解約の予測に有益な訳ではありません。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、フィルタを使用することができます。

▶ [フィルタの生成] ダイアログで [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。

▶ 生成されたフィルタ型ノードをデータ型ノードに接続します。

図 13-5
重要なフィールドの選択



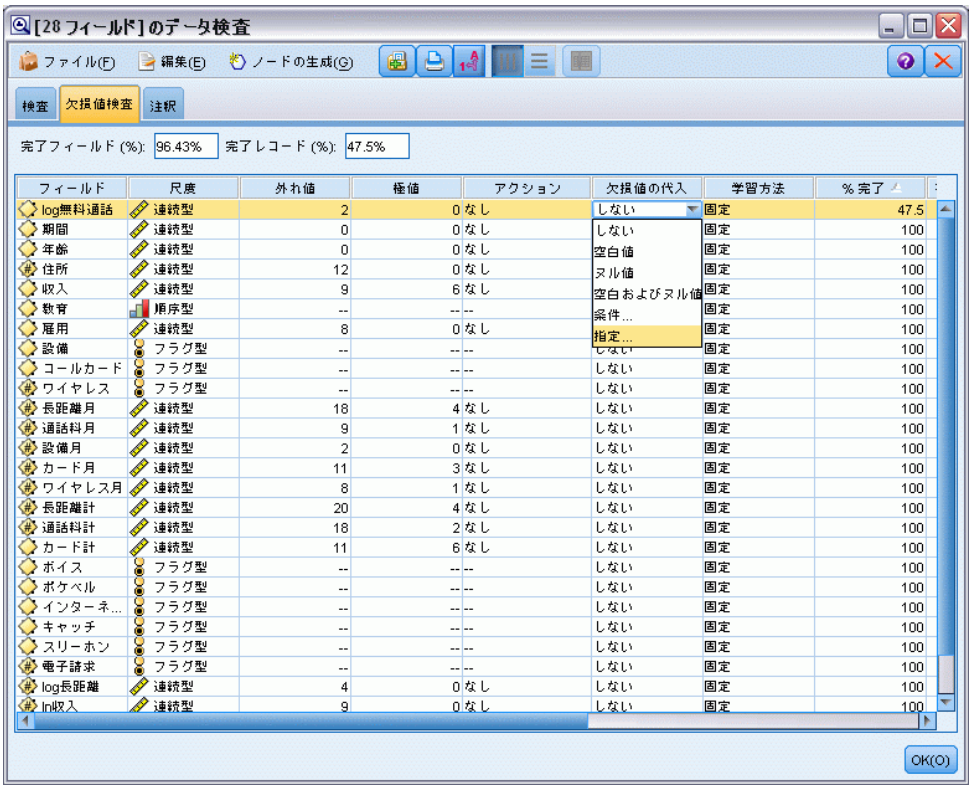
▶ データ検査ノードを生成されたフィルタ ノードに接続します。

データ検査ノードを開いて、[**実行**] をクリックします。

▶ データ検査ブラウザの［欠損値検査］タブで、[% 完了］列をクリックして、数値の昇順に列をソートします。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。これは完全性が 50% 未満です。

► logtoll の［欠損値の代入］列で、［指定］をクリックします。

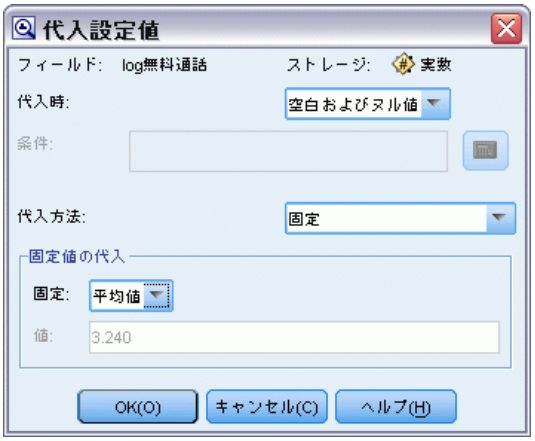
図 13-6
logtoll の欠損値の代入



► ［代入時］について、［空白値とヌル値］を選択します。［固定］について、［平均値］を選択し、［OK］をクリックします。

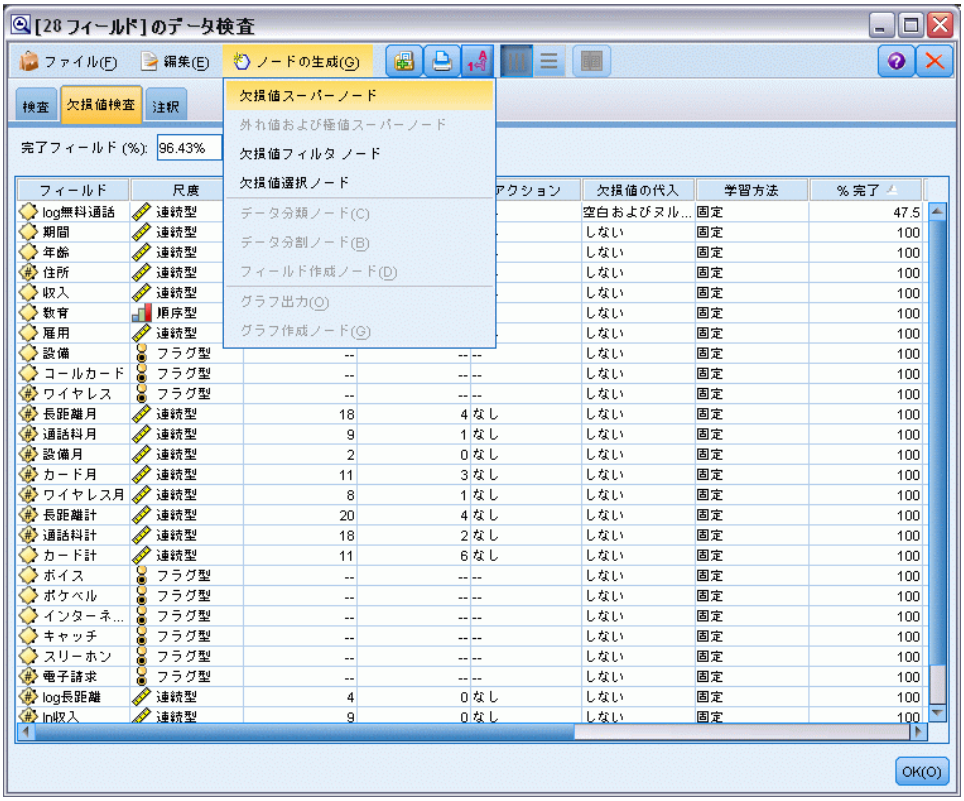
[平均値] を選択すると、代入した値が、全体データの中のすべての値の平均に悪影響を及ぼしません。

図 13-7
代入設定の選択



▶ データ検査ブラウザの［欠損値検査］タブで、欠損値スーパーノード を生成します。これを行うためには、メニューから次の項目を選択します。
ノードの生成 > 欠損値スーパーノード

図 13-8
欠損値スーパーノードの生成

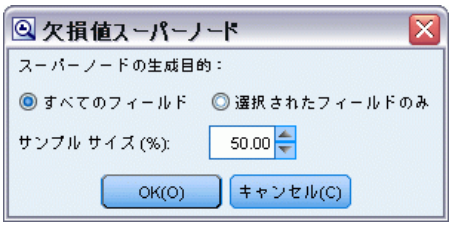


［欠損値スーパーノード］ダイアログ ボックスで、**［サンプル サイズ］**を 50% に増加し、**［OK］**をクリックします。

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 キャンバスに表示されます。

▶ スーパーノードをフィルタ ノードに接続します。

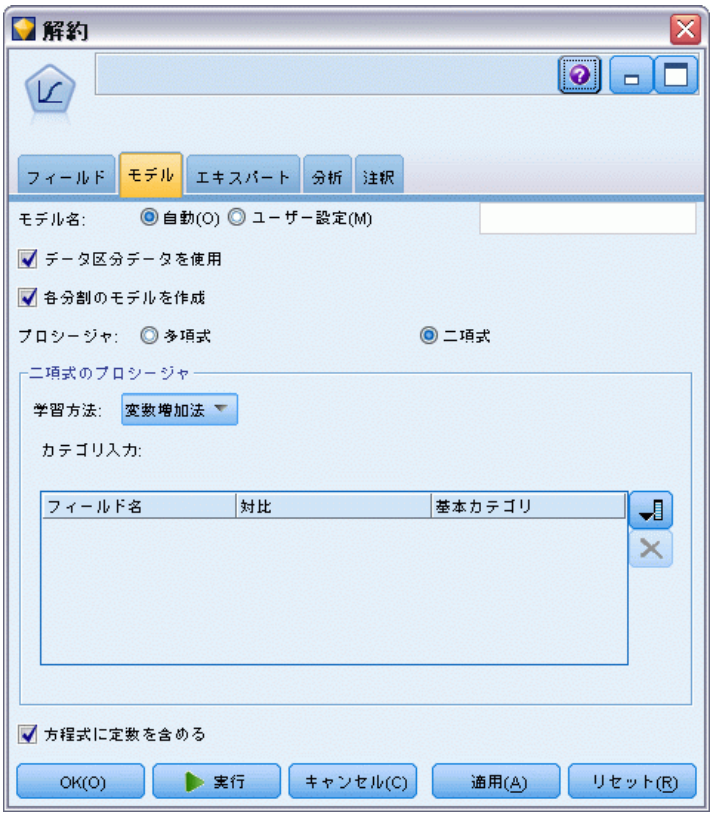
図 13-9
サンプル サイズの指定



▶ ロジスティック ノードをスーパーノードに追加します。

▶ ロジスティック ノードで、[モデル] タブをクリックし、**[二項検定]** 手続きを選択します。[二項検定手続き] エリアで、**[変数増加法]** の方法を選択します。

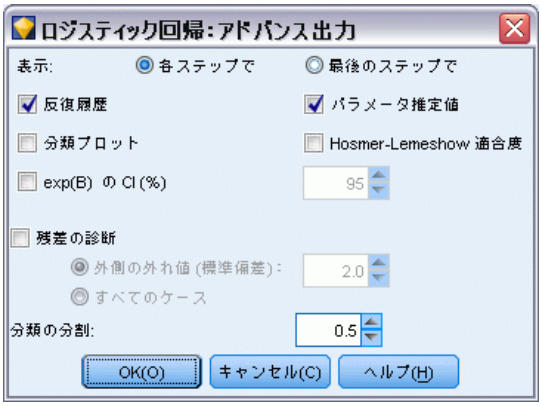
図 13-10
モデル オプションの選択



▶ [エキスパート] タブで、**[エキスパート]** モードを選択し、それから **[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。

▶ ［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 **[パラメータ推定値]** を選択し、[OK] をクリックします。

図 13-11
出力オプションの選択



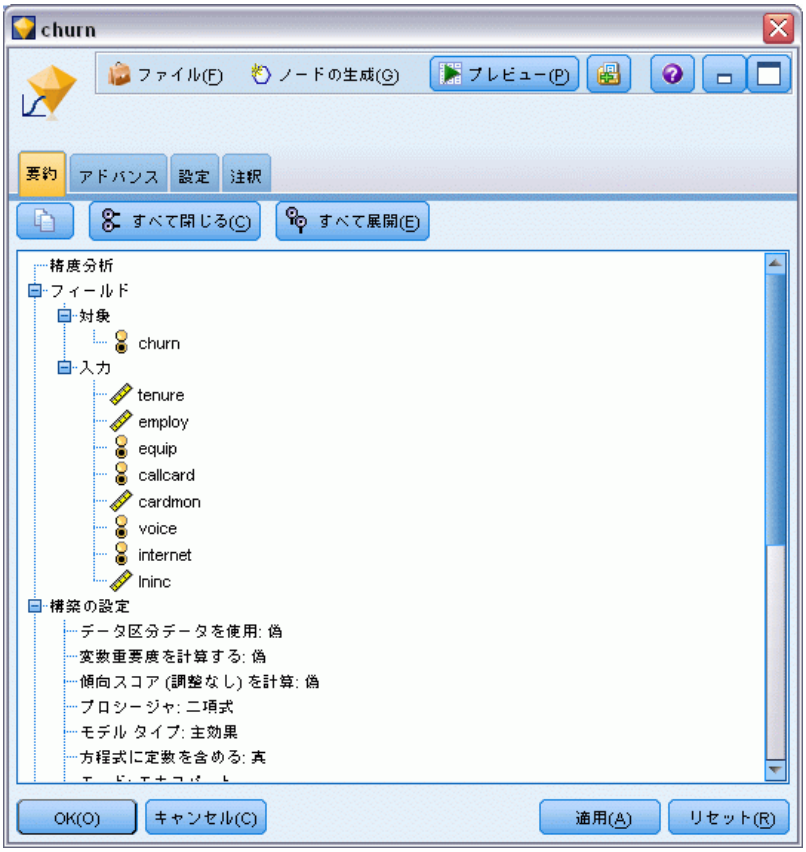
モデルの参照

▶ 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、および右上の［モデル］パレットに追加されます。モデル ナゲットの詳細を表示するには、モデル ナゲットを右クリックして、**[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、考慮のために提出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が存在します。

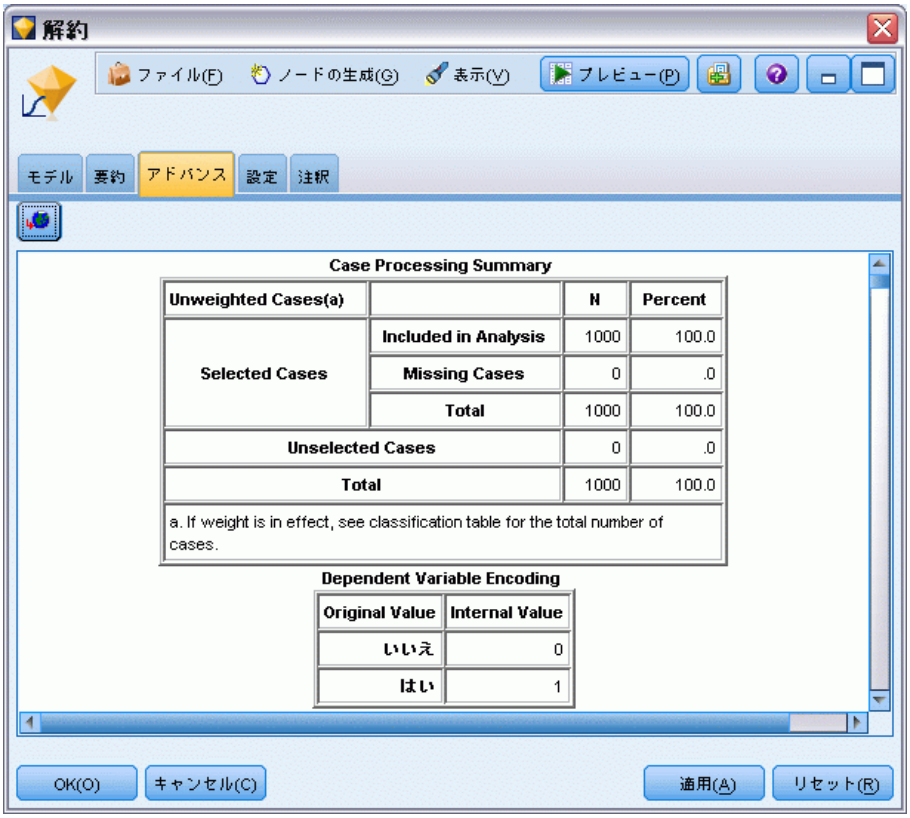
図 13-12
対象および入力フィールドを表示するモデルの要約



［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パーセンテージを表示します。さらに、これ

は、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、その数を一覧します。また選択されなかったケースも一覧します。

図 13-13
ケース処理要約(S)

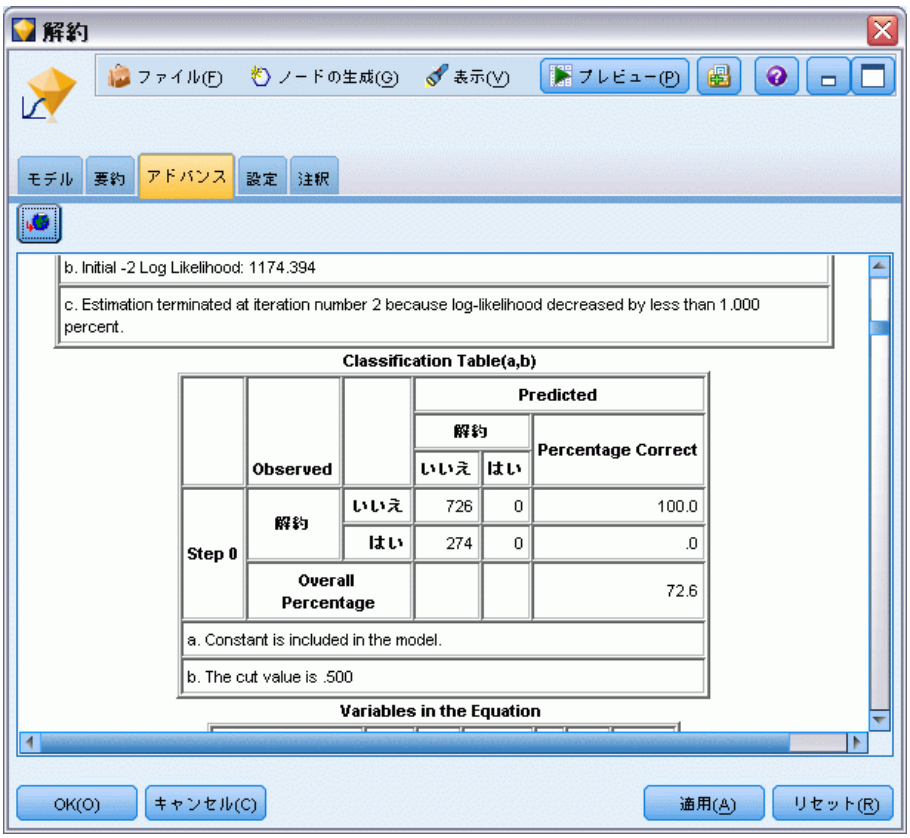


▶ [処理したケースの要約] をスクロール ダウンして、ブロック 0: 開始ブロックの下の分類表を表示します。

変数増加ステップワイズ法はヌル モデルから開始します。これは、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。ヌル 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、ヌル モデルは 72.6% の精度です。これは単純に 726 人の顧客が解約

しなかったことが正しく予測されるためです。しかし、解約しなかった顧客は、まったく正しく予測されません。

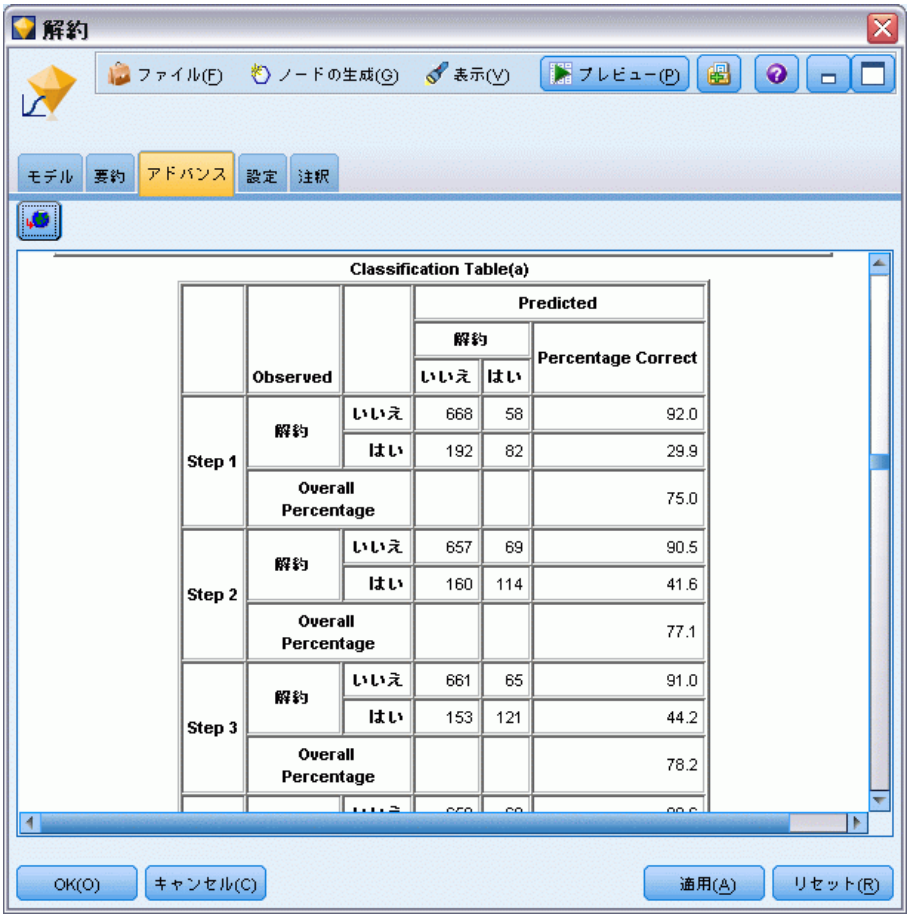
図 13-14
分類表 - ブロック 0 の開始



▶ 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。

この分類表には、予測値としてのモデルのための結果が各ステップで追加されることが示されます。すでに、最初のステップで、ただ 1 つの予測値を使用した後で、モデルは解約予測の精度を 0.0% から 29.9% に増加しています。

図 13-15
分類表 - ブロック 1



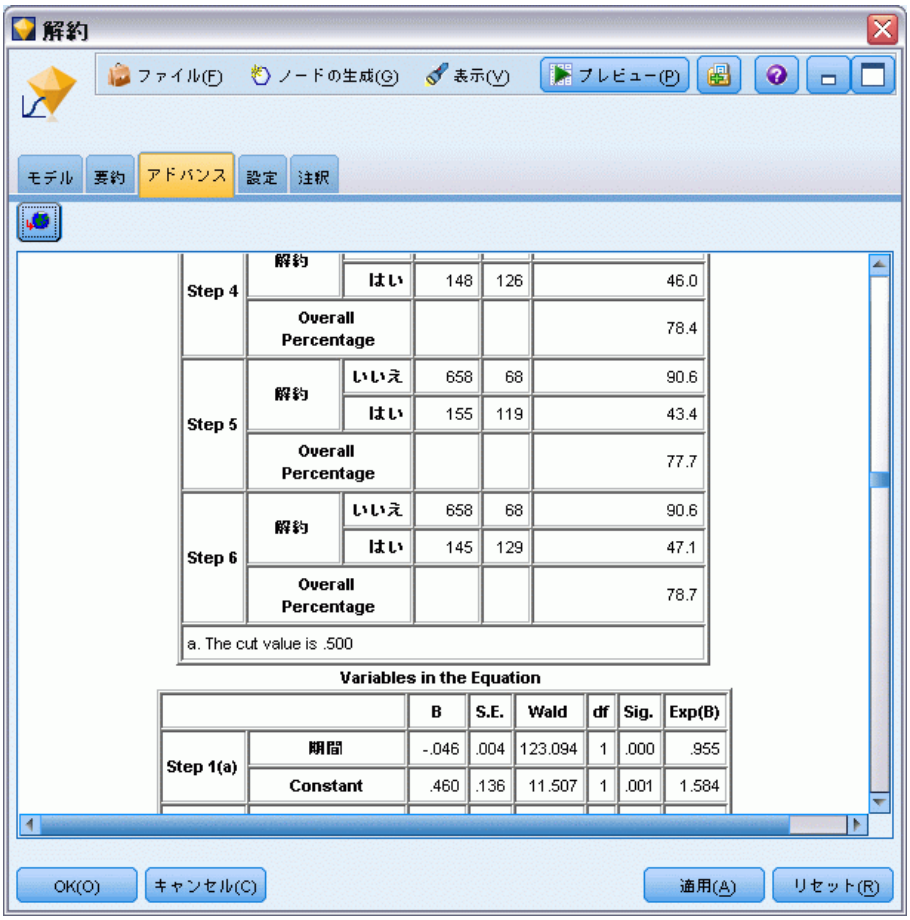
	Observed		Predicted		
			解約		Percentage Correct
			いいえ	はい	
Step 1	解約	いいえ	668	58	92.0
		はい	192	82	29.9
	Overall Percentage				75.0
Step 2	解約	いいえ	657	69	90.5
		はい	160	114	41.6
	Overall Percentage				77.1
Step 3	解約	いいえ	661	65	91.0
		はい	153	121	44.2
	Overall Percentage				78.2

▶ この分類表をスクロール ダウンします。

分類テーブル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は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、解約しなかっ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ヌル モデルから大幅な向上です。

図 13-16
分類表 - ブロック 1



解約を減らしたい顧客については、それを約半分に減らすことができるのならば、それらの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、全体のパーセンテージをモデルの精度のガイドとして使用することが、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。元のヌル 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には、モデルを構築するときに、多数の診断方法が［詳細出力］ダイアログ ボックスで使用可能です。詳細は、 10 章 ロジスティック モデル ナゲットの詳細

出力 in IBM SPSS Modeler 15 モデル作成ノード を参照してください。 IBM® SPSS® Modelerで使用するモデリング メソッドの数学的基礎の説明は、インストール ディスクの ¥Documentation ディレクトリにもある SPSS Modeler 『アルゴリズム ガイド』に記載されています。

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をどれだけうまく一般化しているかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。 詳細は、 4 章 データ区分ノード in IBM SPSS Modeler 15 入力ノード、プロセス ノード、出力ノード を参照してください。

章
14

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

ここでは、全国規模のブロードバンド プロバイダから依頼を受けた分析担当者が、帯域の利用状況を予測するために、ユーザー契約数の予測値を求めるというシナリオを設定します。予測は、全国的な加入者を構成する各地域の市場向けに必要です。多数の地方市場の今後 3 か月の予測を行うために時系列モデル作成を使用します。2 番目の例は、入力データが時系列ノードに入力するための正しいフォーマットでない場合のデータの変換方法を示しています。

この例では、broadband_create_models.str というストリームを使用します。これは、broadband_1.sav というデータ ファイルを参照します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos フォルダにあります。これは、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 フォルダにあります。

最後の例は、予測をさらに 3 か月拡張するために、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図示しています。

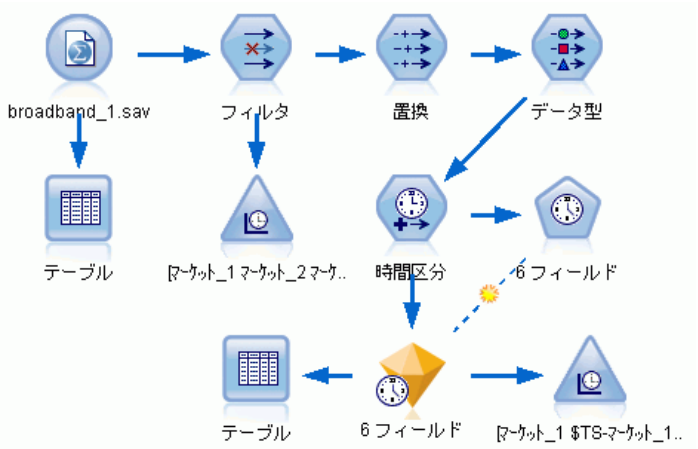
SPSS Modeler では、1 回の操作で複数の時系列モデルを作成できます。単純化するには、すべての市場の合計に 5 種類を加えてモデル化するだけでいいのですが、使用するソース ファイルには 85 のさまざまな市場の時系列データが含まれています。

broadband_1.sav データ ファイルには、85 ヶ所の地域の市場の月間利用データが含まれています。この例の目的のために、最初の 5 種類の系列のみを使用します。全体に加えて、これらの 5 種類の系列のそれぞれについて個別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つけるために時間区分ノードで使用します。データ フィールドは文字列として SPSS Modeler に読み込まれますが、SPSS Modeler でこのフィールドを

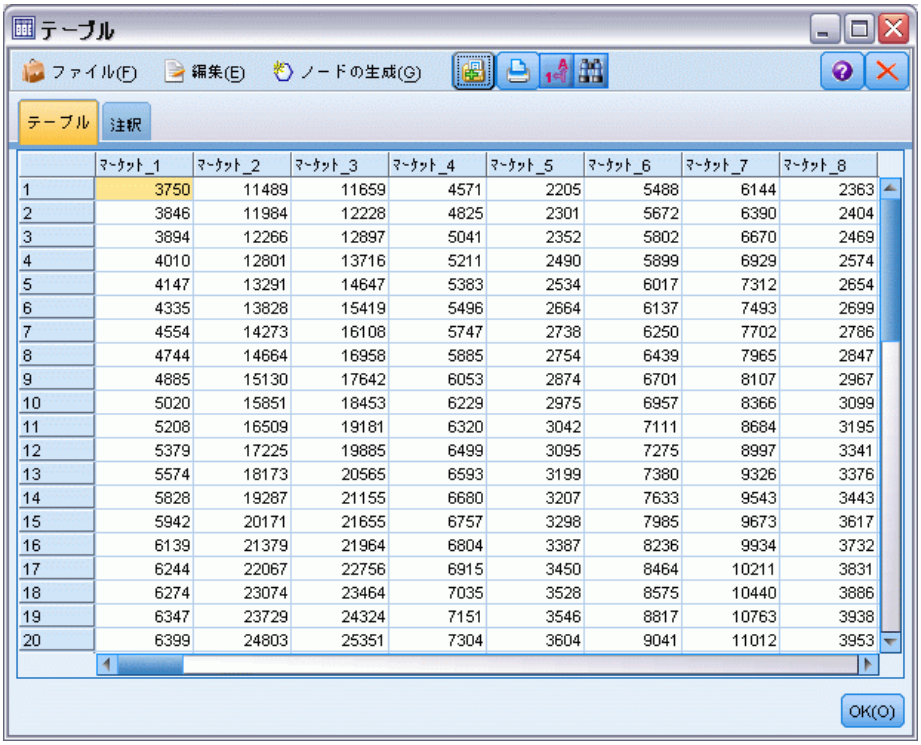
使用するには、置換ノードを使用してストレージ タイプを数値による日付のフォーマットに変換します。

図 14-1
時系列モデル作成を示すサンプル ストリーム



時系列ノードは、それぞれの時系列が各区間に対応する行と、別の列になっている必要があります。SPSS Modeler は、必要に応じてこのフォーマットに適合するようにデータを変換する方法を提供します。

図 14-2
ブロードバンド地方市場の月次購読データ



	マーケット_1	マーケット_2	マーケット_3	マーケット_4	マーケット_5	マーケット_6	マーケット_7	マーケット_8
1	3750	11489	11659	4571	2205	5488	6144	2363
2	3846	11984	12228	4825	2301	5672	6390	2404
3	3894	12266	12897	5041	2352	5802	6670	2469
4	4010	12801	13716	5211	2490	5899	6929	2574
5	4147	13291	14647	5383	2534	6017	7312	2654
6	4335	13828	15419	5496	2664	6137	7493	2699
7	4554	14273	16108	5747	2738	6250	7702	2786
8	4744	14664	16958	5885	2754	6439	7965	2847
9	4885	15130	17642	6053	2874	6701	8107	2967
10	5020	15851	18453	6229	2975	6957	8366	3099
11	5208	16509	19181	6320	3042	7111	8684	3195
12	5379	17225	19885	6499	3095	7275	8997	3341
13	5574	18173	20565	6593	3199	7380	9326	3376
14	5828	19287	21155	6680	3207	7633	9543	3443
15	5942	20171	21655	6757	3298	7985	9673	3617
16	6139	21379	21964	6804	3387	8236	9934	3732
17	6244	22067	22756	6915	3450	8464	10211	3831
18	6274	23074	23464	7035	3528	8575	10440	3886
19	6347	23729	24324	7151	3546	8817	10763	3938
20	6399	24803	25351	7304	3604	9041	11012	3953

ストリームの作成

- 新規のストリームを作成し、broadband_1.sav を示す Statistics ファイル入力ノードを追加します。
- モデルを単純化するために、フィルタ ノードを使用して [Market_6] から [Market_85] フィールドまでを除外し、さらに [MONTH_] と [YEAR_] フィールドを除外します。

  ヒント ：隣接する複数のフィールドを 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たまま [Market_85] フィールドまでマウスをドラッグします。選択したフィール

ドは、青色で強調表示されます。別のフィールドを追加するには、Ctrl キーを押しながら［MONTH_］と［YEAR_］フィールドをクリックします。

図 14-3
モデルの単純化

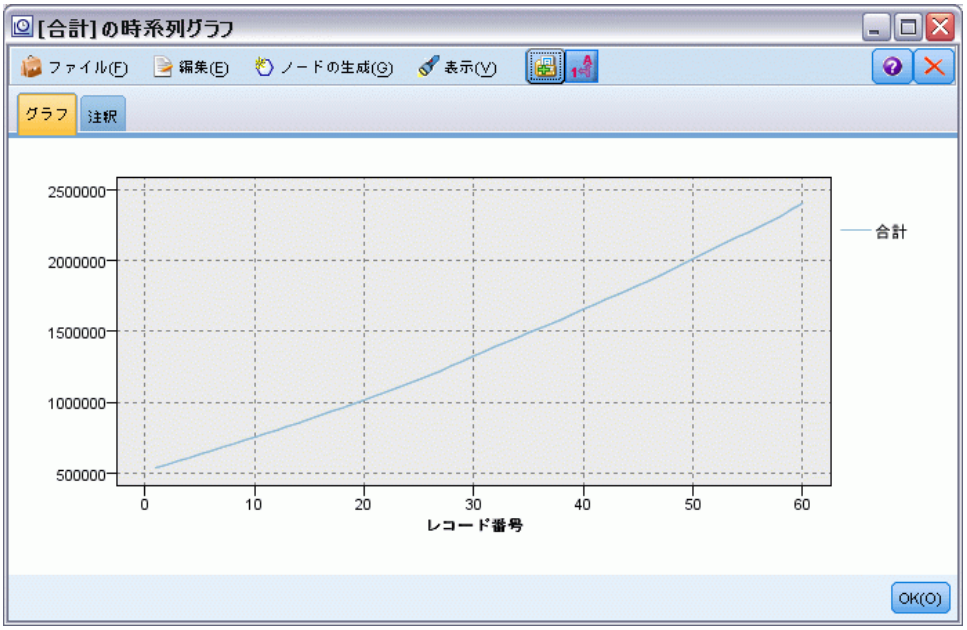
データの調査

モデルを作成する前にデータの特性を知っておくのは良い考えです。データは季節変動を示しますか ?エキスパート モデラーは各系列について最適な季節性モデルまたは非季節性モデルを自動的に検索できますが、データに季節性が存在しない場合は、検索を非季節性モデルに限定することによって結果が早く得られます。85 ヶ所の地域の市場のそれぞれについてデータを検証しなくても、すべての市場の加入者総数をプロットすることによって季節性の有無の概略を知ることができます。

図 14-4
加入者総数のプロット

- ［グラフ］パレットから、時系列ノードをフィルタ ノードに接続します。
- ［合計］フィールド を［系列］リストに追加します。
- **［別のパネルに時系列を表示］**と**［正規化］**のチェック ボックスの選択を解除します。
- **［実行］**をクリックします。

図 14-5
[合計] フィールドの時系列

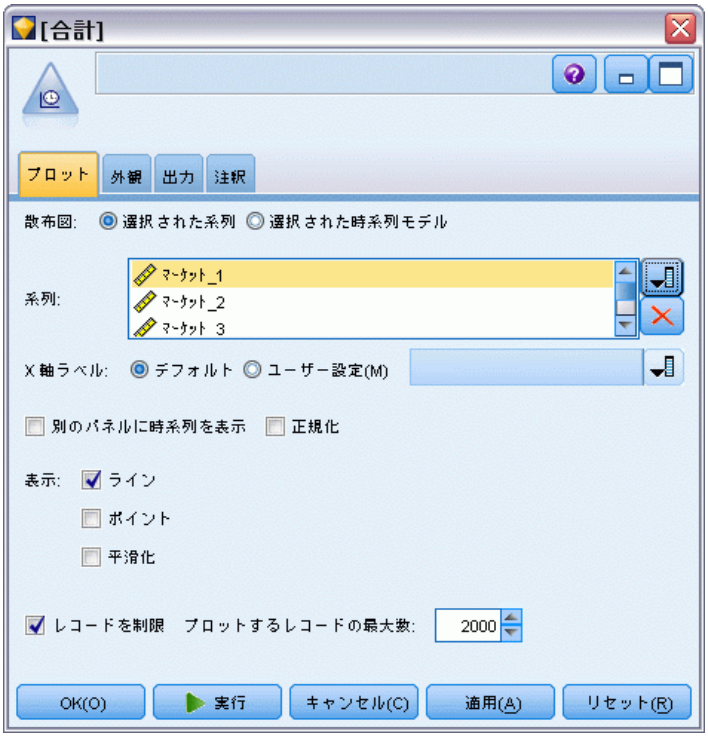


この系列は、非常に滑らかな上昇傾向を示しており、季節変動の存在を示すものはありません。個々の市場について見れば、季節変動を持つ系列も存在する可能性はありますが、データ全般では季節性は顕著な特徴ではないと考えられます。

もちろん、季節モデルを除外してしまう前に、個々の系列を調べることは必要です。それにより、季節性が現れる系列を抽出し、それらを別個にモデル化することができます。

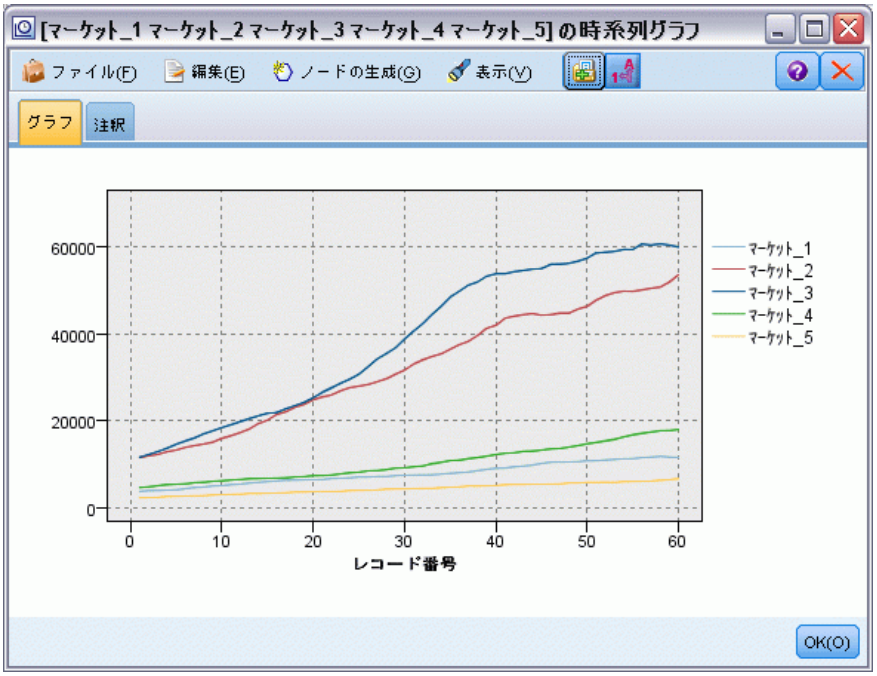
IBM® SPSS® Modeler では、複数の系列をまとめて簡単にプロットできます。

図 14-6
複数の時系列のプロット



- 時系列ノードを開き直します。
- ［系列］リストから［合計］フィールドを削除します（フィールドを選択して赤い X ボタンをクリックします）。
- ［Market_1］から［Market_5］までのフィールドをリストに追加します。
- ［実行］をクリックします。

図 14-7
複数のフィールドの時系列



それぞれの市場を検査すると、各ケースにおける安定した増加傾向が明らかになります。いくつかの市場は他の市場に比べてやや不安定ですが、季節性が見受けられる徴候はありません。

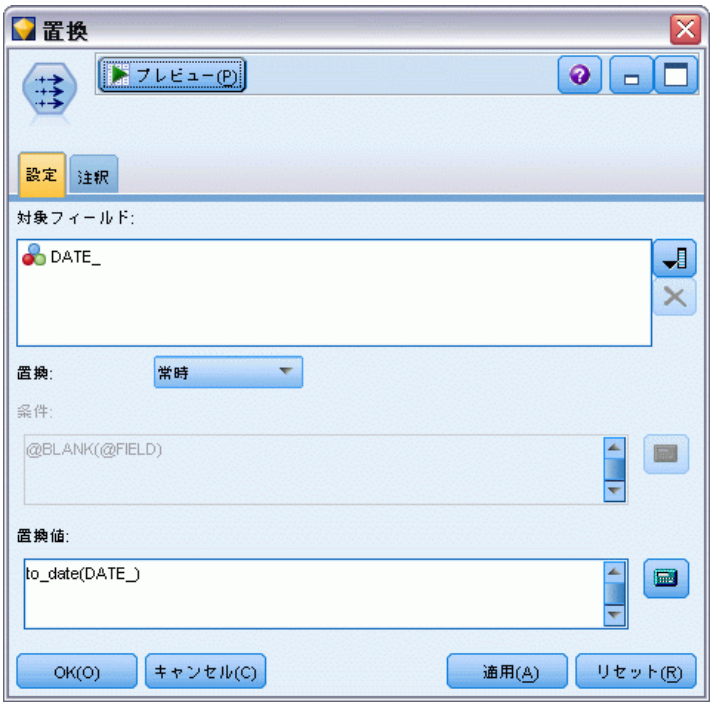
日付の定義

ここで、[DATE_] フィールドのストレージ タイプを日付のフォーマットに変更する必要があります。

- 置換ノードをフィルタ ノードに接続します。
- 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。
- ［DATE_］を選択して［**対象フィールド**］に追加します。
- ［**置換**］条件を［**常時**］に設定します。

▶ ［置換値］を［to_date(DATE_)］に設定します。

図 14-8
日付ストレージ タイプの設定



デフォルトの日付フォーマットを［日付］フィールドのフォーマットに適合するように変更します。これは、予測した通りに機能するように［日付］フィールドを変換するのに必要です。

▶ メニューで、**[ツール]** > **[ストリームのプロパティ]** > **[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。

▶ デフォルトの［**日付フォーマット**］を［MON YYYY］に設定します。

図 14-9
日付フォーマットの設定

対象の定義

▶ データ型ノードを追加し、［DATE_］フィールドに対しては、役割を［**なし**］に設定します。それ以外のフィールド（［Market_n］フィールドおよび［合計］フィールド）に対しては、役割を［**対象**］に設定します。

▶ ［**値の読み込み**］ボタンをクリックして［値］列を読み込みます。

図 14-10
複数のフィールドに対する役割の設定

時間区分の設定

▶ 時間区分ノード（[フィールド設定］パレット）を追加します。

▶ ［区間］タブで、時間区分として［**月**］を選択します。

▶ ［**データから構築**］オプションを選択します。

▶ ビルド フィールドとして［DATE_］を選択します。

図 14-11
時間区分の設定

▶ ［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。

▶ 値を 3 に設定します。

▶ [OK] をクリックします。

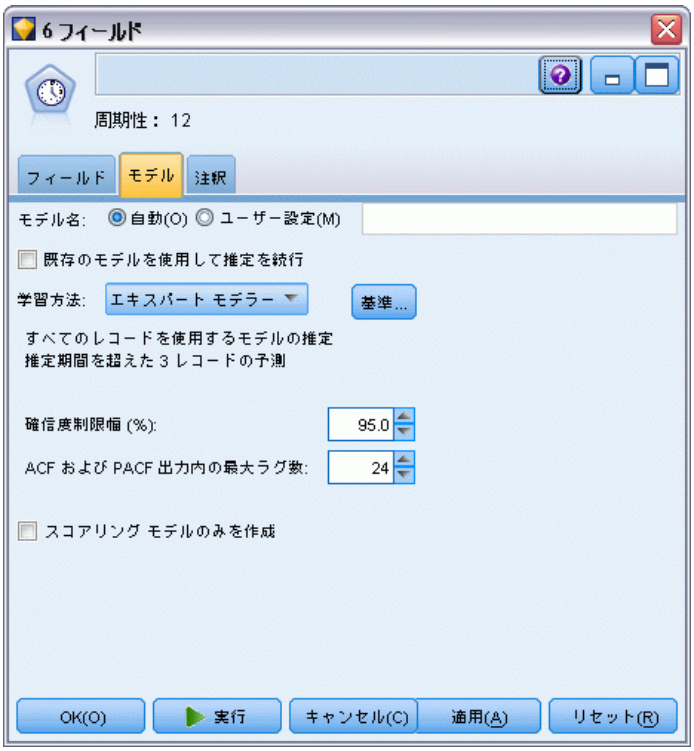
図 14-12
予測期間の設定

モデルの作成

▶ ［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。

▶ 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。

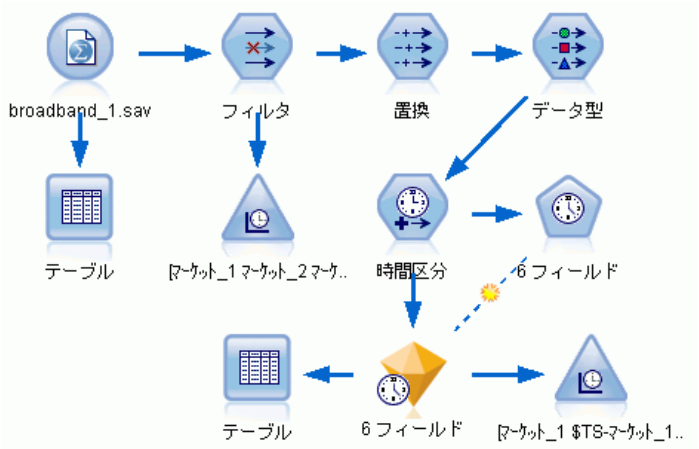
図 14-13
時系列用のエキスパート モデルの選択



▶ 時系列モデル ナゲットを時間区分ノードに接続します。

▶ 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 14-14
時系列モデル作成を示すサンプル ストリーム



これで3種類の新しい行（61 から 63）がオリジナルのデータに付け加えられます。これは予測期間用の行であり、この場合、2004年の1月から3月までです。

新しい列もいくつか表示されます。すなわち、時間区分ノードが追加した列が $TI_ 個、時系列ノードが追加した列が $TS- 個です。これらの列は、それぞれの行について以下の事項を示します（つまり、時系列データにおける各区分）。

列	説明
$TI_TimeIndex	この行の時間区分指標値。
$TI_TimeLabel	この行の時間区分ラベル。
$TI_Year	この行の生成されたデータの年と月の指標。
$TI_Month	
$TI_Count	この行の新規データの判断に関連するレコード数。
$TI_Future	この行が予測データを含んでいるかどうかを示します。
$TS-colname	元のデータの各列に対応する生成されたモデル データ。
$TSLCI-colname	生成されたモデル データの各列の、信頼区間の初めの値。
$TSUCI-colname	生成されたモデル データの各列の、信頼区間の終わりの値。
$TS-Total	この行の $TS-colname 値の合計。

列	説明
$TSLCI-Total	この行の $TSLCI-colname 値の合計。
$TSUCI-Total	この行の $TSUCI-colname 値の合計。

予測操作のための最も重要な列は、$TS-Market_n、$TSLCI-Market_n、および $TSUCI-Market_n の列です。特に、行 61 から 63 におけるこれらの列は、各地方市場のユーザー購読予測データと信頼区間を含んでいます。

モデルの検証

- 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。

他の市場のために生成したタイプから市場 5 について 別のタイプのモデルを生成することを、エキスパート モデラーがどのようにして選択したかということに留意してください。

図 14-15
市場のために生成された時系列モデル

[予測値] の列は、対象ごとの予測値として使用するフィールドの数を示しています。この場合は、[なし] です。

このビューのそれ以外の列は、モデルごとの適合度の測定結果を示します。[StationaryR**2] の列は固定 R-squared の値を示します。この統計は、モデルが説明す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になります。

［Q］、［df］、および［Sig.］列は、モデルの残差エラーの無作為のテストの Ljung-Box 統計に関連し 、エラーが無作為であるほど、そのモデルは良好です。［Q］列は Ljung-Box 統計そのもので、［df］列（自由度）は特定の対象を推定する場合さまざまなモデル パラメータ数を示します。

［Sig.］の列は Ljung-Box 統計の有意確率を示しています。これはモデルが適切に指定されているかどうかの別の指標になります。0.05 未満の有意確率は、残差エラーが無作為でなく、モデルが説明しない観測系列に構造があるということを意味します。

固定 R 2 乗および有意確率の両方を考慮に入れると、エキスパート モデラーが Market_1、Market_3、および Market_5 を選択していたモデルは特に良好です。Market_2 と Market_4 の Sig. 値はどちらも 0.05 未満であり、これはこれらの市場にさらに適合するモデルによる実験が必要かもしれないということを意味します。

下部に表示されている要約値は、すべてのモデルの統計の分布に関する情報を提供します。たとえば、すべてのモデルの固定 R 2 乗の平均値は 0.247 であるのに対し、その最小値は 0.049（全モデルの値）であり、最大値は 0.544（Market_5 の値）です。

SE とは、統計ごとのすべてのモデルに関する標準エラーを意味します。たとえば、すべてのモデルに関する固定 R 2 乗の標準エラーは 0.169 です。

要約のセクションは、モデルの統計の分布に関する情報を示すパーセンタイル値も含んでいます。各パーセンタイルについて、その割合（パーセント）に当たるモデルでは、適合度統計量の値が表示されている値より低くなります。

したがって、たとえば、0.121 未満の固定 R 2 乗の値を示すモデルはわずか 25% です。

▶ ［表示］ドロップダウン リストをクリックして［**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、モデルで説明することができる時系列における総変動の推定です。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 14-16
時系列モデルの詳細表示

RMSE は 2 乗平均平方誤差、つまりモデルが予測する値から系列の実際の値がどれほど異なるかの尺度で、系列自体に使用される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を望みます。最初は、これまで確認してきた統計によれば適切である一方で、Market_2 および Market_3 のモデルは他の 3 つの使用のモデルに比べて良好ではありません。

追加の適合度測定結果は、平均絶対パーセント誤差（MAPE）と最大絶対パーセント誤差（MaxAPE）を含みます。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どれほど変動するのかということに関する尺度で、パーセンテージ値で表されます。すべてのモデルに関する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 値はすべてのモデルが非常に低い、1% を下回る平均不確定要素を表すことを示します。MaxAPE の値は最大絶対パーセント誤差を示し、予測に関して最悪のシナリオを想定するのに役立ちます。これは、それぞれのモデルの最大パーセント誤差がおよそ 1.8 ～ 2.5% の範囲内に収まることを示し、非常に低い数字のセットです。

MAE（絶対平均誤差）値は、予測の誤差の絶対値の平均を表示します。RMSE 値と同様、系列自体に使用される単位と同じ単位で表されます。MaxAE は、同じ単位で最大予測誤差を示し、予測に関して最悪のシナリオを示します。

これらの絶対値に焦点を当てると、対象系列は規模が変動する市場の加入者数を示すため、パーセント誤差の値（MAPE and MaxAPE）はこのケースでより役に立ちます。

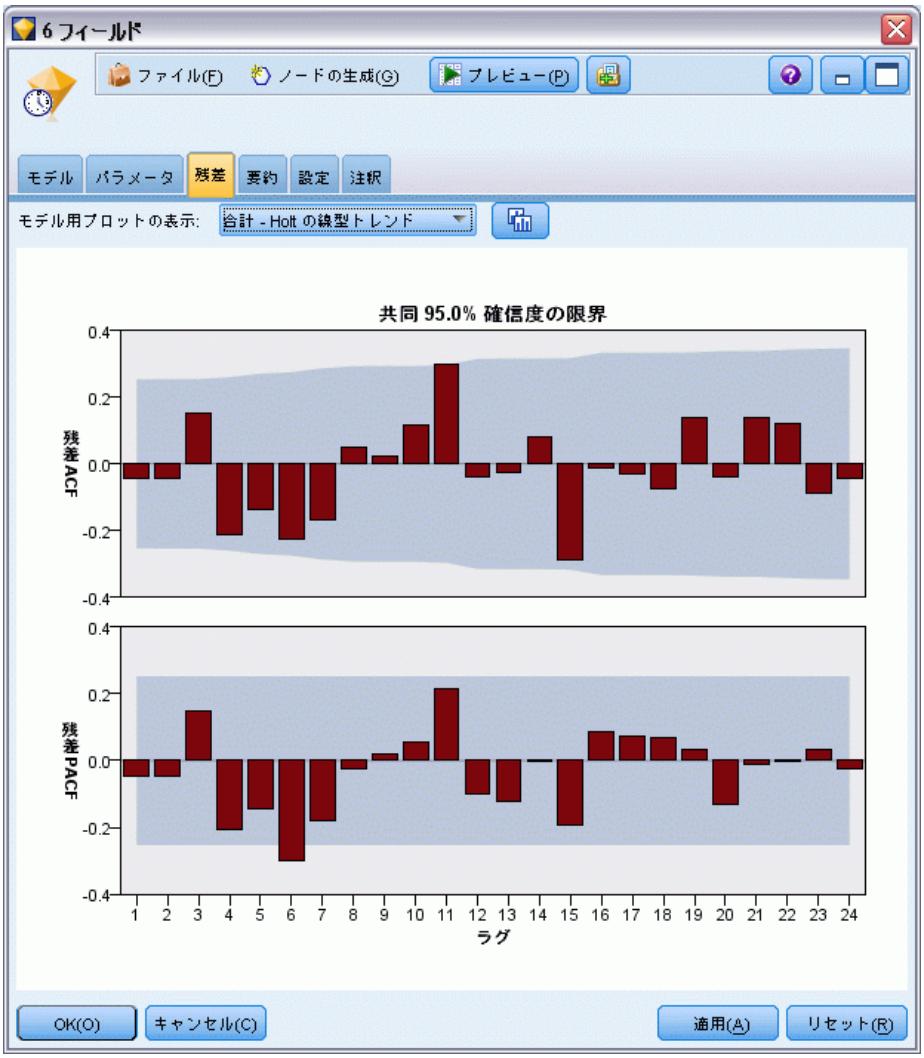
MAPE 値および MaxAPE 値は妥当な程度の不確定要素を表しますか ?もちろん非常に低い値です。これは、許容できるリスクが問題ごとに変化するため、ビジネス センスが働き始める状況です。適合度統計は適切な範囲内に収まり、残差エラーの確認を継続すると想定します。

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）を検証することにより、ただ単に適合度統計を表示するよりも、モデルに対するさらに定量的な洞察が示されます。

指定された時系列は、季節性、トレンド、循環性、その他重要な要素など、無作為でないすべての変動を取得します。この場合、時間の経過に伴って誤差は自己相関しません。これらの自己相関関数のいずれかにおける著しい構造は、基礎的モデルが不完全であるということを意味します。

▶ [残差] タブをクリックして、最初の地方市場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 14-17
市場に関する ACF と PACF の値



これらのプロットでは、誤差変数の元の値は、最大 24 時間遅延し、元の値と比較して時間の経過に伴って相関があるかどうかを確認します。適切なモデルの場合、ACF プロットには正（上）または負（下）のいずれの方向であっても、色の付いた領域を超えるバーはありません。

万一領域を超えた場合、PACF プロットで構造が確定されているかどうかを確認します。PACF プロットは、時間ポイントの干渉で系列値を制御した後相関を調査します。

Market_1 の値はすべて色の付いた領域内であるため、継続して他の市場の値をチェックすることができます。

- [**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、他の市場や全市場についてこれらの値を表示します。

  Market_2 および Market_4 の値には、考えられる原因が少しあり、Sig. 値から以前予測したものを確認します。いずれかのポイントでこれらの市場のさまざまなモデルを検証し、より良い適合度を取得しているかどうかを確認する必要がありますが、この例の残りについては、Market_1 モデルで他に学習できることについて考えます。

- [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。

- [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を外します。

- [**系列**] リストで、フィールド選択ボタンをクリックして [Market_1] と [$TS-Market_1] のフィールドを選択し、[OK] をクリックしてそれらをリストに追加します。

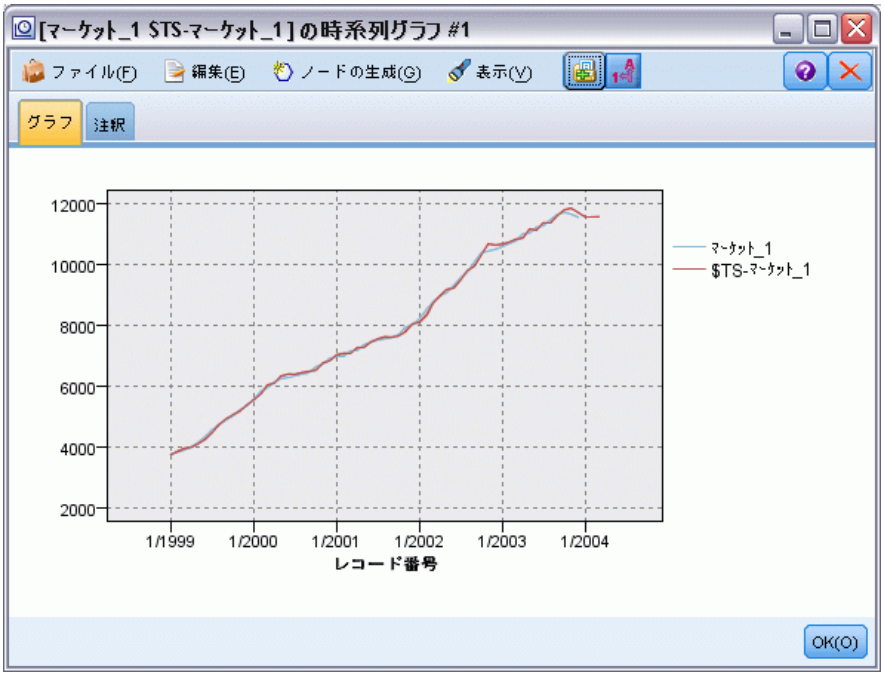
▶ ［実行］をクリックして、最初の市場と地方の市場の実際のデータと予測データの折れ線グラフを表示します。

図 14-18
プロットするフィールドを選択

予測の線（$TS-Market_1）が実際のデータの終わりを超えてどのようにして伸びているのかに留意してください。これで、この市場の次の 3 ヶ月の見込み需要を予測できます。

時系列全体の実際のデータの線と予測データの線はグラフ上でひじょうに接近しており、これは特定の時系列について信頼できるモデルであることを示しています。

図 14-19
Market_1 の実際のデータと予測データの時系列



将来の事例で使用するために、モデルをファイルに保存します。

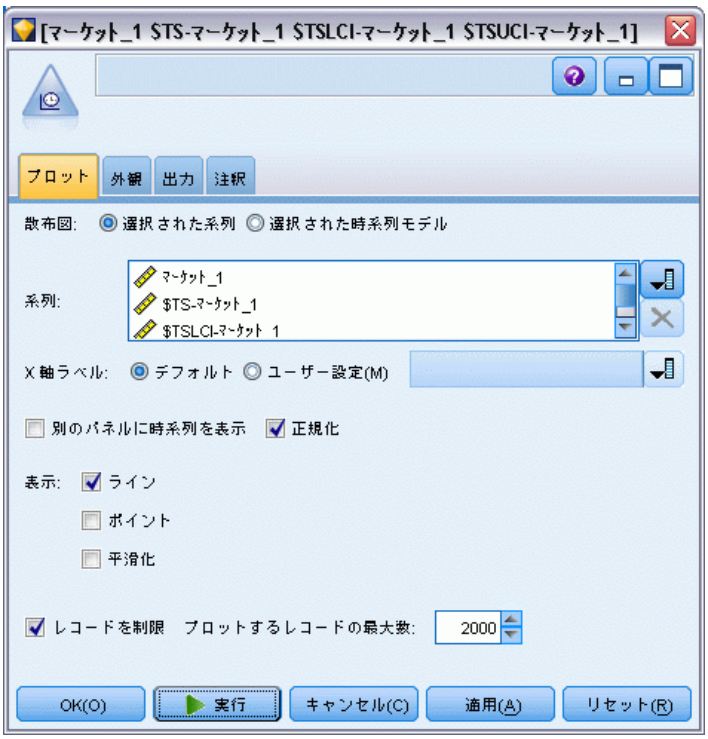
- ［OK］をクリックして、現在のグラフを閉じます。
- 時系列モデル ナゲットを開きます。
- ［**ファイル**］>［**ノードの保存**］を選択し、ファイルの場所を指定します。
- ［**保存**］をクリックします。

  この特定の市場に関する信頼できるモデルが得られましたが、予測はどのような誤差許容範囲を見込んでいますか ?信頼区間を検証することにより、この指標が得られます。

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

▶ [実行] をクリックします。

図 14-20
さらにフィールドをプロットに追加

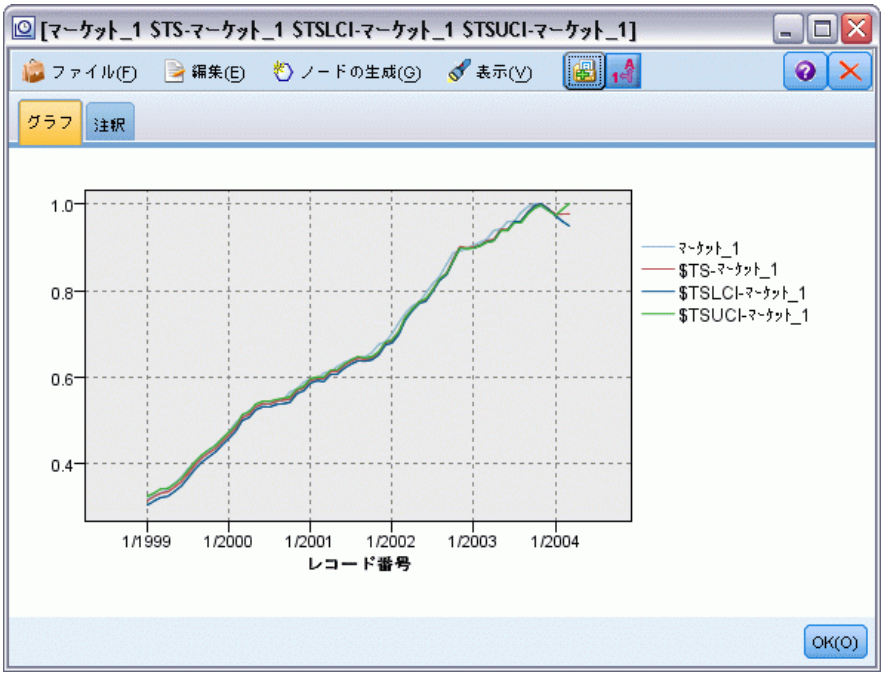


以前と同じグラフができましたが、追加した信頼区分には上限（$TSUCI）と下限（$TSLCI）があります。

信頼区分の境界が、将来に向けて予測する場合の不確定要素の増大を示しながら予測期間を超えてどのように伸びているのかに留意してください。

ただし、それぞれの期間が経過するにつれて、予測の基礎になるさらに 1 か月分（この場合）の実際の使用データが得られます。新しいデータをストリームに読み込んで、信頼できるとわかっているモデルを再適用できます。詳細は、 p. 221 時系列モデルの再適用 を参照してください。

図 14-21
追加された信頼区分による時系列



要約

エキスパート モデラーを使用して複数の時系列について予測を行う方法を学び、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フォーマット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、最初の時系列の事例の時系列モデルを再適用しますが、単独で使用することもできます。詳細は、 p. 196 時系列ノードによる予測 を参照してください。

本来のシナリオでは、全国的なブロードバンド プロバイダのアナリストは、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加入ユーザーの月次予測を行うように要求されます。ただし、すでにエキスパート モデラーを使用してモデルは作成されており、今後 3 か月の予測結果も出ています。

データ ウェアハウスは本来の予測期間の実際のデータで更新されましたので、そのデータを使用して予測期間をさらに 3 ヶ月拡張します。

この例では、broadband_2.sav という名前を持つ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。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos フォルダにあります。これは、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作成します。モデルを保存していなくても問題はありません。モデルは Demos フォルダに用意してあります。

▶ Demos の下の streams フォルダから ストリーム broadband_apply_models.str を開きます。

図 14-22
ストリームを開く

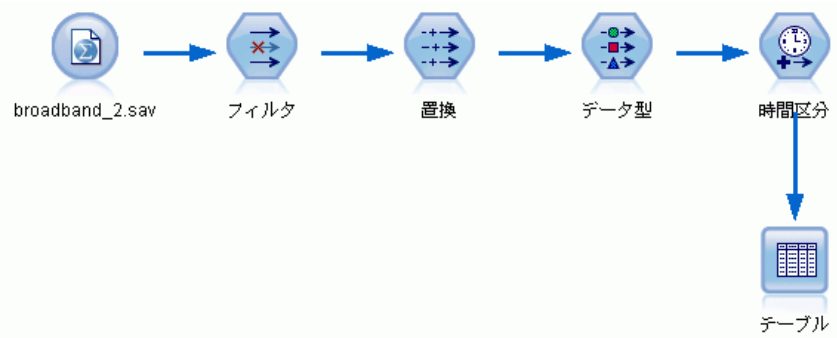
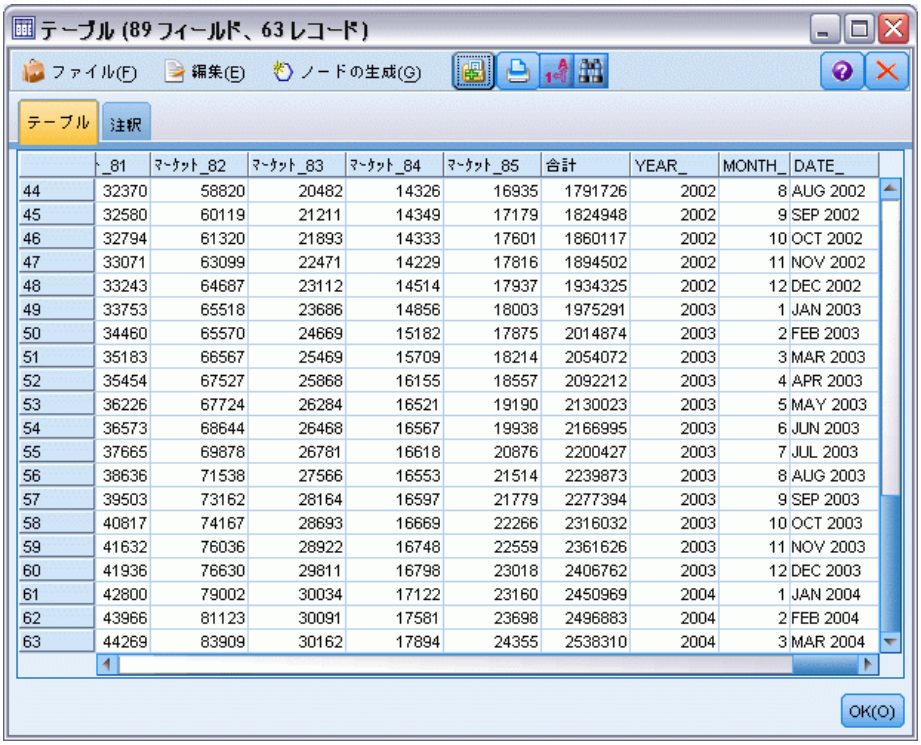


図 14-23
更新された販売データ



	_81	マーケット_82	マーケット_83	マーケット_84	マーケット_85	合計	YEAR_	MONTH_	DATE_
44	32370	58820	20482	14326	16935	1791726	2002	8	AUG 2002
45	32580	60119	21211	14349	17179	1824948	2002	9	SEP 2002
46	32794	61320	21893	14333	17601	1860117	2002	10	OCT 2002
47	33071	63099	22471	14229	17816	1894502	2002	11	NOV 2002
48	33243	64687	23112	14514	17937	1934325	2002	12	DEC 2002
49	33753	65518	23686	14856	18003	1975291	2003	1	JAN 2003
50	34460	65570	24669	15182	17875	2014874	2003	2	FEB 2003
51	35183	66567	25469	15709	18214	2054072	2003	3	MAR 2003
52	35454	67527	25868	16155	18557	2092212	2003	4	APR 2003
53	36226	67724	26284	16521	19190	2130023	2003	5	MAY 2003
54	36573	68644	26468	16567	19938	2166995	2003	6	JUN 2003
55	37665	69878	26781	16618	20876	2200427	2003	7	JUL 2003
56	38636	71538	27566	16553	21514	2239873	2003	8	AUG 2003
57	39503	73162	28164	16597	21779	2277394	2003	9	SEP 2003
58	40817	74167	28693	16669	22266	2316032	2003	10	OCT 2003
59	41632	76036	28922	16748	22559	2361626	2003	11	NOV 2003
60	41936	76630	29811	16798	23018	2406762	2003	12	DEC 2003
61	42800	79002	30034	17122	23160	2450969	2004	1	JAN 2004
62	43966	81123	30091	17581	23698	2496883	2004	2	FEB 2004
63	44269	83909	30162	17894	24355	2538310	2004	3	MAR 2004

更新した月次データは broadband_2.sav に集められます。

- 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、行 61 ～ 63 の 2004 年の 1 月から 3 月の間の実際の販売データで更新されました。

- ストリームの時間区分ノードを開きます。
- ［**予測**］タブをクリックします。

▶ ［**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。

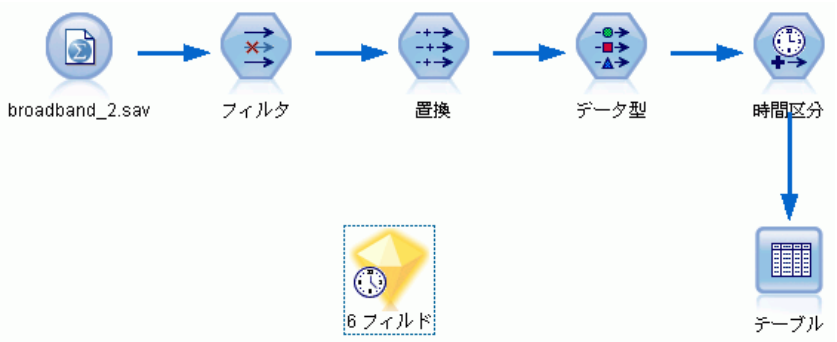
図 14-24
予測期間の設定をチェック

保存されたモデルの取得

▶ IBM® SPSS® Modeler メニューで、［**挿入**］>［**ファイルからノード**］を選択し、Demos フォルダから TSmodel.nod ファイルを選択します（あるいは、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 14-25
モデル ナゲットの追加

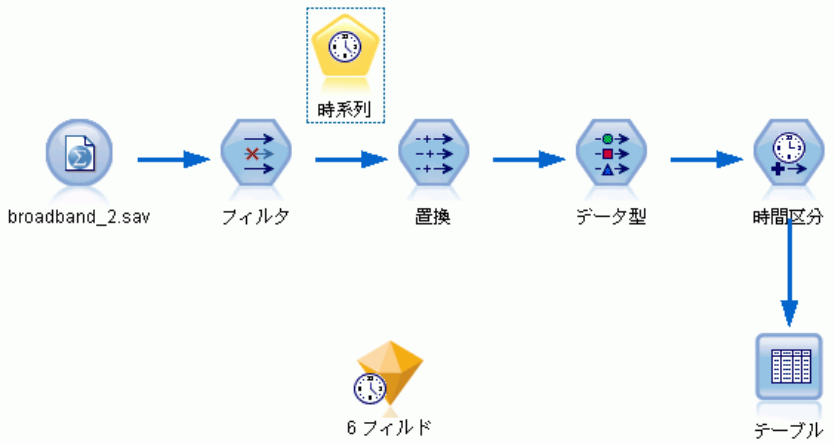


モデル作成ノードの生成

▶ 時系列モデル ナゲットを開いて、［生成］>［モデル作成ノードの生成］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 14-26
モデル ナゲットからモデル作成ノードを生成



新規モデルの生成

▶ 時系列モデル ナゲットを閉じて、領域から削除します。

古いモデルは、60 行のデータに構築されました。更新した販売データ（63 行）に基づいて新規モデルを生成する必要があります。

▶ 新たに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 14-27
モデル作成ノードをストリームに接続

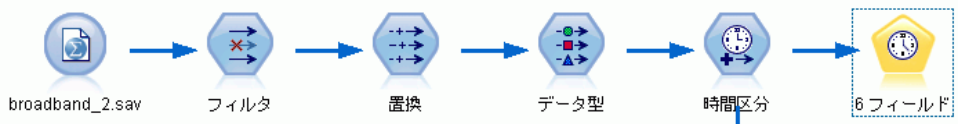
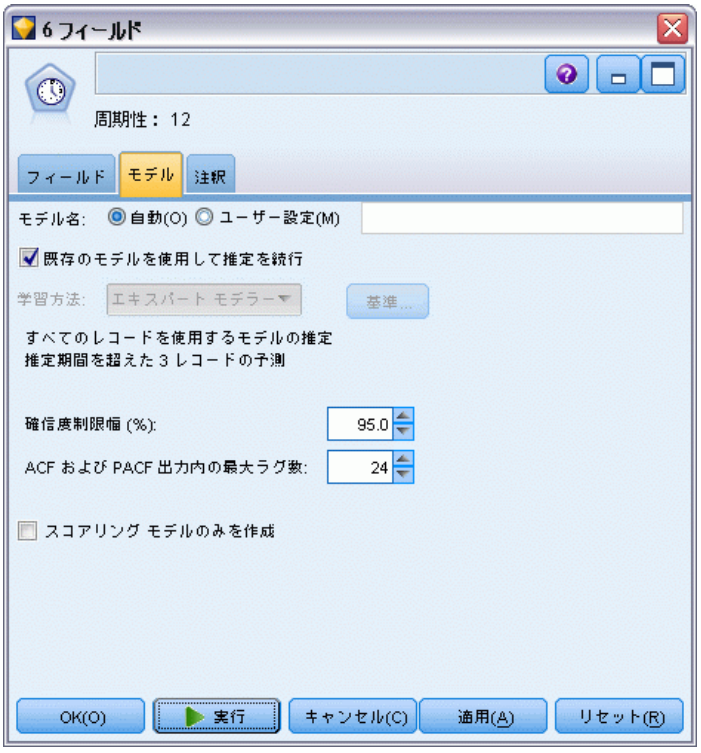


図 14-28
時系列モデルのための格納済み設定値の再使用



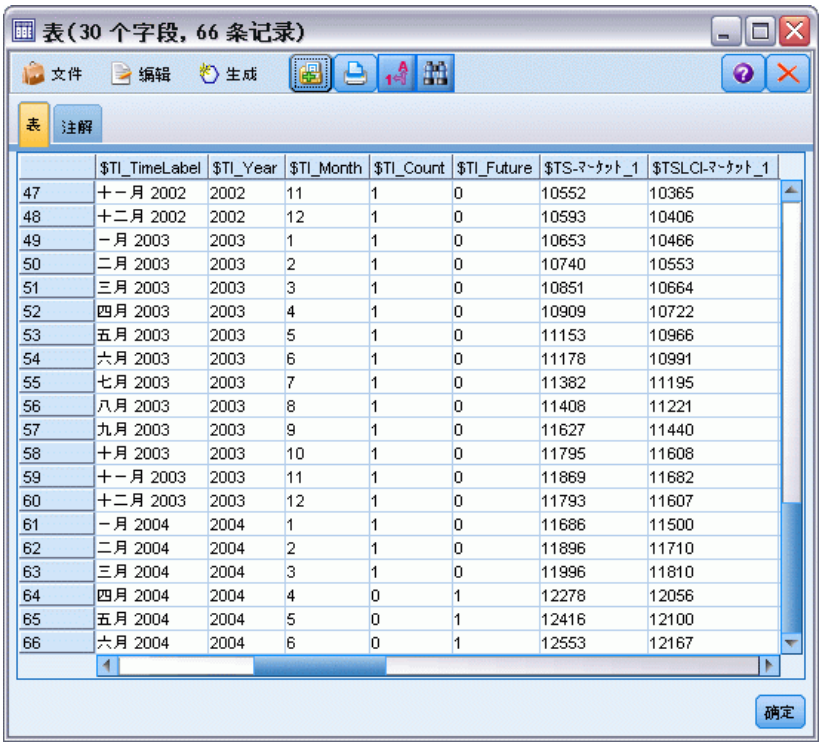
▶ 時系列ノードを開きます。

▶ ［**モデル** タブで、［**既存のモデルを使用して推定を続行**］がチェックされていることを確認します。

▶ ［**実行**］をクリックして新規モデルナゲットを領域と［モデル］パレットに配置します。

新規モデルの検証

図 14-29
新規予測を示すテーブル



	$TI_TimeLabel	$TI_Year	$TI_Month	$TI_Count	$TI_Future	$TS-ﾏｰｹｯﾄ_1	$TSLCI-ﾏｰｹｯﾄ_1
47	十一月 2002	2002	11	1	0	10552	10365
48	十二月 2002	2002	12	1	0	10593	10406
49	一月 2003	2003	1	1	0	10653	10466
50	二月 2003	2003	2	1	0	10740	10553
51	三月 2003	2003	3	1	0	10851	10664
52	四月 2003	2003	4	1	0	10909	10722
53	五月 2003	2003	5	1	0	11153	10966
54	六月 2003	2003	6	1	0	11178	10991
55	七月 2003	2003	7	1	0	11382	11195
56	八月 2003	2003	8	1	0	11408	11221
57	九月 2003	2003	9	1	0	11627	11440
58	十月 2003	2003	10	1	0	11795	11608
59	十一月 2003	2003	11	1	0	11869	11682
60	十二月 2003	2003	12	1	0	11793	11607
61	一月 2004	2004	1	1	0	11686	11500
62	二月 2004	2004	2	1	0	11896	11710
63	三月 2004	2004	3	1	0	11996	11810
64	四月 2004	2004	4	0	1	12278	12056
65	五月 2004	2004	5	0	1	12416	12100
66	六月 2004	2004	6	0	1	12553	12167

- 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。
- テーブルノードを開いて、[**実行**] をクリックします。

  格納済み設定値を再使用しているので、新規モデルは依然として 3 ヶ月先を予測します。ただし、推定期間（時間区分ノードで指定）は 1 月ではなく 3 月に終了するため、今回は 4 月から 6 月までを予測します。

図 14-30
プロットするフィールドの指定

- 時系列モデル ナゲットに時系列グラフ ノードを接続します。

  今回は、特に時系列モデルのために設計された時系列表示を使用します。

- ［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。

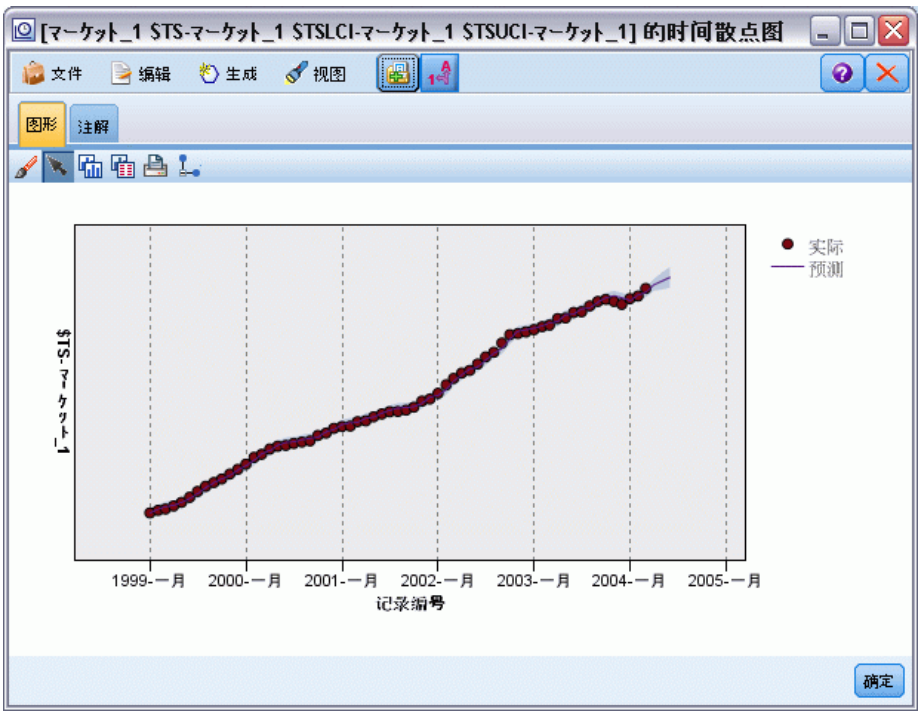
- ［**系列**］リストで、フィールド選択ボタンをクリックして[$TS-Market_1]フィールドを選択し、［OK］をクリックしてそれをリストに追加します。

- ［**実行**］をクリックします。

  2004年6月までの予測（予想）売り上げおよび信頼区分（青色で示された領域）とともに、2004 年 3 月までの Market_1 の実際の売り上げを示すグラフが作成されます。

最初の例の場合のように、予測値はその期間中の実際のデータに従い、優れたモデルを持っていることを再度示します。

図 14-31
6 月まで拡張した予測



要約

現在の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を学びました。さらに、モデルを再構築することなくこれを実行しました。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。

章
15
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。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、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 ディレクトリにあります。

初期の例では、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調べました。ここでは、モデルを選択する場合に利用できる 2 つの方法、すなわち指数平滑化と ARIMA について検討します。

適切なモデルの決定を支援するために、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。

- 系列は包括的なトレンドを示しますか ?そのような場合、そのトレンドは一定のように見えますが、それとも時間の経過とともに減衰するように思われますか ？
- 系列は季節性を示しますか ?そのような場合、季節的変動は時間の経過とともに大きくなるようですか、それとも一定の状態で継続するように見えますか？

ストリームの作成

▶ 新規のストリームを作成し、catalog_seasfac.sav を示す Statistics ファイル入力ノードを追加します。

図 15-1
カタログ販売の予測

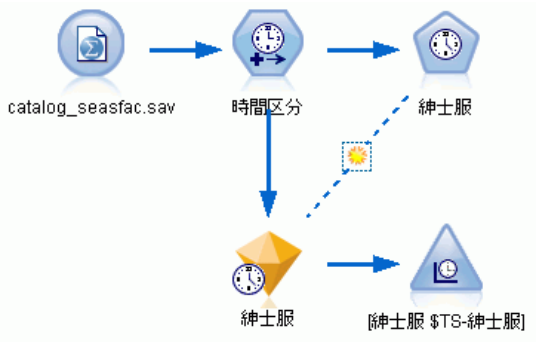


図 15-2
対象フィールドの指定

- IBM® SPSS® Statistics ファイル入力ノードを開いて［タイプ］タブを選択します。
- **［値の読み込み］**をクリックし、次に［OK］をクリックします。
- ［男性］フィールドの［役割］の列をクリックし、役割を**［対象］**に設定します。
- その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 15-3
時間区分の設定

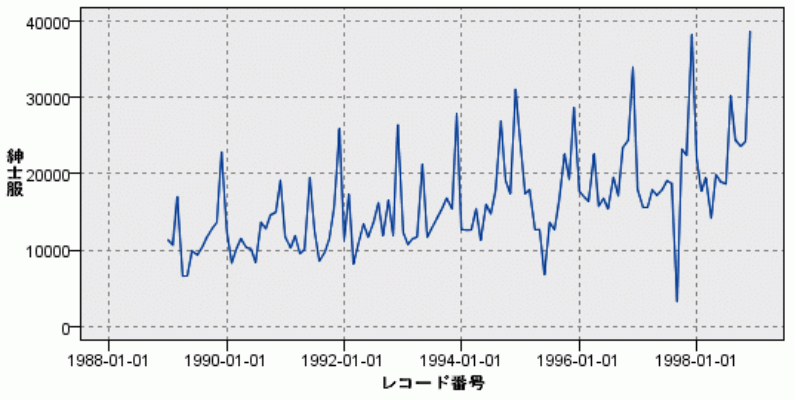
- 時間区分ノードを SPSS Statistics ファイル入力ノードに接続します。
- 時間区分ノードを開き、[**時間区分**] を [**月**] に設定します。
- [**データから構築**] を選択します。
- [**フィールド**] を [**日付**] に設定し、[OK] をクリックします。

図 15-4
時系列のプロット

- 時系列ノードを時間区分ノードに接続します。
- ［プロット］タブで［男性］を［系列］リストに追加します。
- ［正規化］チェック ボックスの選択を解除します。
- ［実行］をクリックします。

データの調査

図 15-5
紳士服の実際の売り上げ



系列は一般的な増加傾向を示します。すなわち、系列値は時間と共に増加する傾向があります。増加傾向は表面的には一定であり、線型トレンドを示します。

系列は、グラフの垂直線に現れているように、12 月に売り上げが増えるという特異な季節性パターンも示します。季節変動は系列の増加傾向と共に大きくなるように思われ、これは相加的というよりもむしろ相乗的な季節性を示唆しています。

▶ [OK] をクリックして、プロットを閉じます。

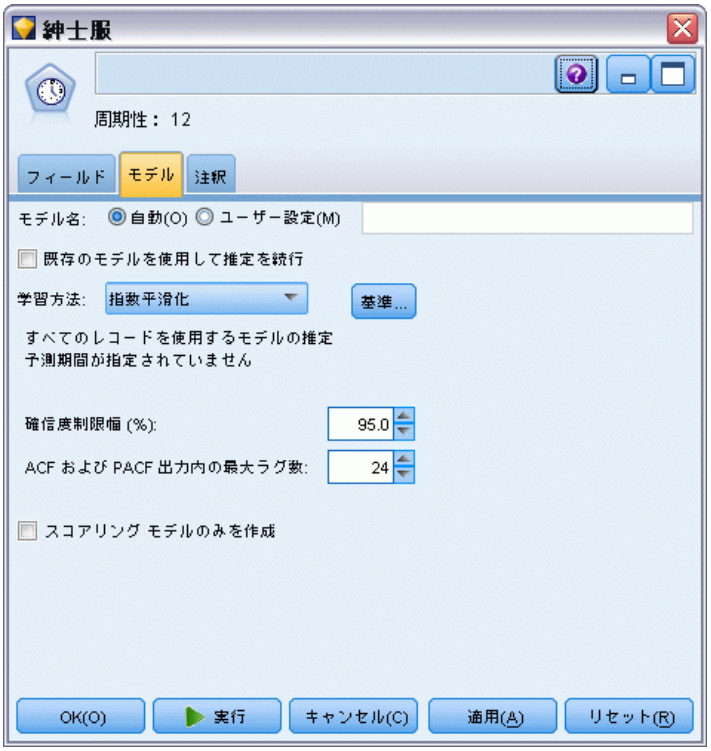
系列の特性を確認したので、それをモデル化する用意ができました。トレンド、季節性またはその両方を示す時系列を予測する上で、指数平滑化法が役に立ちます。すでに確認したように、データは両方の特性を示します。

指数平滑法

最適な指数平滑化モデルの構築には、モデルはトレンドを含む必要があるのか、季節性を含む必要があるのか、それとも両方を含む必要があるのかにかかわらず、モデル タイプの決定および選択したモデルの最適なパラメータの取得が含まれます。

時間の経過に伴う紳士服の売り上げのプロットは、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によるモデルを示唆しました。これは Winters モデルを意味します。ただし、まず、シンプルなモデル（トレンドも季節性もない）を調査し、次に、Holt（線型トレンドを組み込んでいるが季節性はない）モデルを調べます。これにより、モデルがデータに適合しない場合にモデルの構築を成功させるのに不可欠な技術を特定する事例が示されます。

図 15-6
指数平滑化の指定



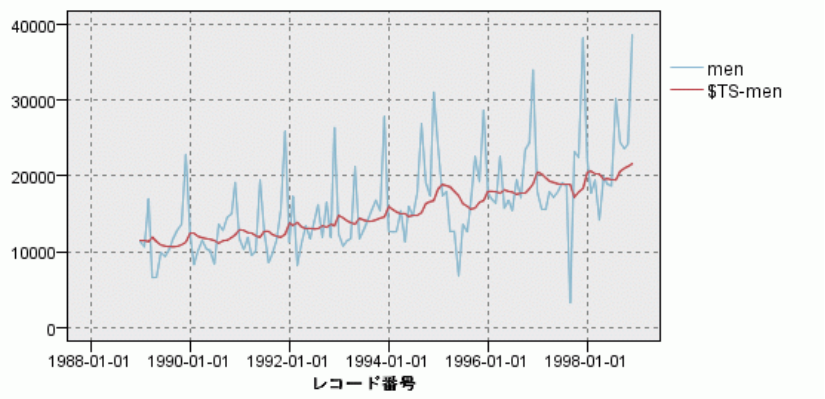
シンプルな指数平滑化モデルから始めます。

- 時系列ノードを時間区分ノードに接続します。
- ［**モデル**］タブで、［**方法**］を［**指数平滑化**］に設定します。
- ［**実行**］をクリックして、モデル ナゲットを作成します。

図 15-7
時系列モデルのプロット

- 時系列ノードをモデル ナゲットに接続します。
- [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。
- [別のパネルに時系列を表示] と [正規化] のチェック ボックスの選択を解除します。
- [実行] をクリックします。

図 15-8
シンプルな指数平滑化モデル

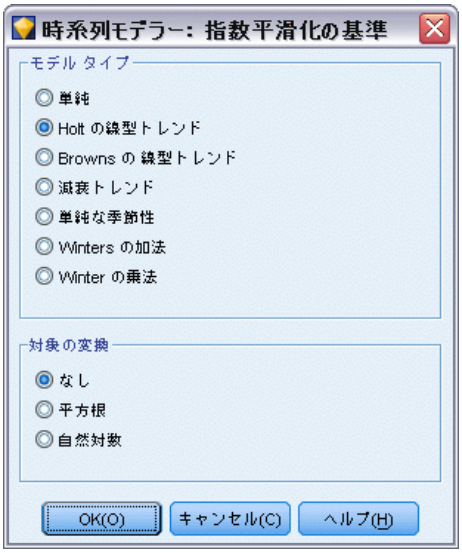


[男性] のプロットが実際のデータを示すのに対し、[$TS-men] は時系列モデルを表します。

シンプルなモデル ノードは実際には漸進的な（しかもかなり冗長的な）増加傾向を示しますが、季節性は無視されます。このモデルは安全に却下できます。

▶ [OK] をクリックして、時系列ウィンドウを閉じます。

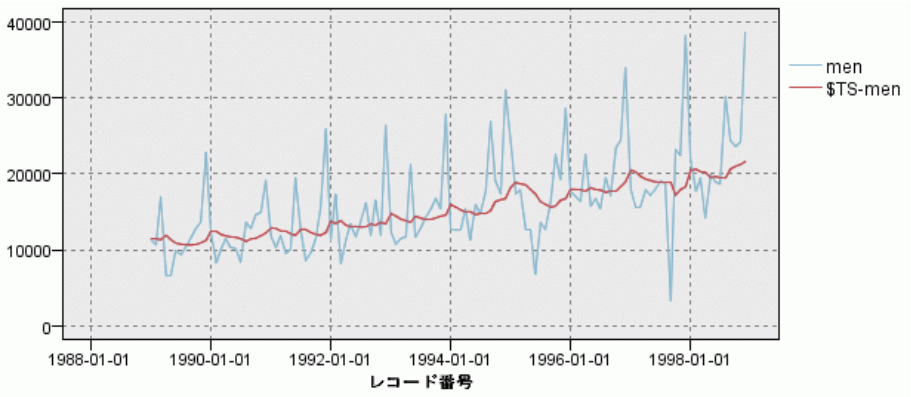
図 15-9
Holt のモデルの選択



Holt の線型モデルを試してみましょう。季節性は取得できそうもありませんが、少なくともシンプルなモデルよりも良好にトレンドをモデル化する必要があります。

- 時系列ノードを開き直します。
- [モデル] タブで、[指数平滑化] を選択した状態で [基準] をクリックします。
- ［指数平滑化基準］ダイアログ ボックスで、[Holts 線形トレンド] を選択します。
- [OK] をクリックして、ダイアログ ボックスを閉じます。
- ［実行］をクリックして、モデル ナゲットを再作成します。
- 時系列ノードを再度開いて、[実行] をクリックします。

図 15-10
Holt の線型トレンド モデル

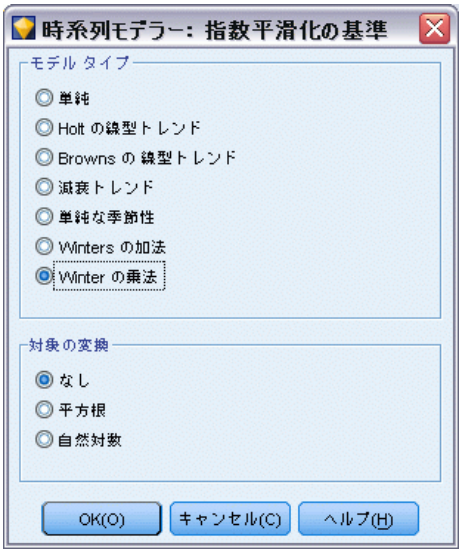


Holt のモデルはシンプルなモデルよりもスムーズな増加傾向を示しますが、季節性は無視されるので、これを破棄することもできます。

- 時系列ウィンドウを閉じます。

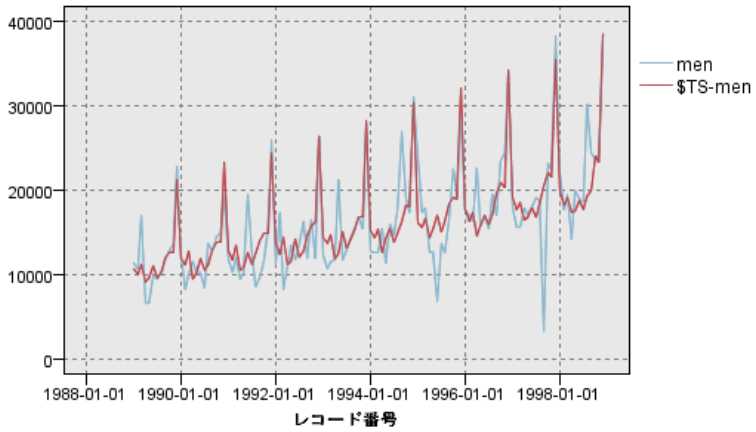
時間の経過に伴う紳士服の売り上げの初期プロットが線型トレンドと相乗的季節性を組み込んでいるモデルを示唆したということを思い出すかもしれません。したがって、より適した候補は Winters のモデルと考えられます。

図 15-11
Winters のモデルの選択



- 時系列ノードを開き直します。
- [**モデル**] タブで、[**指数平滑化**] を選択した状態で [**基準**] をクリックします。
- ［指数平滑化基準］ダイアログ ボックスで、[**Winters 相乗モデル**] を選択します。
- [OK] をクリックして、ダイアログ ボックスを閉じます。
- ［**実行**］をクリックして、モデル ナゲットを再作成します。
- 時系列ノードを開いて、[**実行**] をクリックします。

図 15-12
Winters の相乗モデル



これは良さそうです。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。

データ セットは 10 年の期間を対象とし、毎年 12 月に発生する 10 種類の季節的ピークを含んでいます。予測結果に表示される 10 種類のピークは、実際のデータの 10 種類の年次ピークにうまく適合します。

しかし、この結果は指数平滑化手順の制約も強調しています。上向きと下向きの山形部分に注目すると、無視できない顕著な構造があることがわかります。

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のならば、指数平滑化は良い選択かもしれません。この場合のようにより複雑な構造をモデル化するには、ARIMA 手順の使用を考慮する必要があります。

ARIMA 分析

ARIMA 手順では、時系列の微調整モデリングに適した自己回帰統合移動平均（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のコンポーネントのモデル作成に指数平滑法モデルよりも洗練された方法が用意されており、モデル内に予測変数を含める利点が追加されています。

予測モデルを開発したいと願っているカタログ会社の例を継続しながら、売り上げの変動を説明するのに使用できるいくつかの系列と共に紳士服の月間売り上げに関するデータをその会社がどのようにして集めたのかを論じました。考えられる予測値には、郵送するカタログの数、カタログのページ数、受注用の電話回線の数、広告印刷物に要した費用、およびカスタマー サービス担当者の数が含まれます。

これらの中に、予測に有効な予測変数はあるでしょうか。予測値によるモデルは予測値を使用しないモデルよりも本当に優れていますか ?ARIMA 手順を使用すると、予測値による予測モデルが作成でき、また、予測値なしの指数平滑化モデルに関する予測能力に顕著な相異があるかどうかを確認できます。

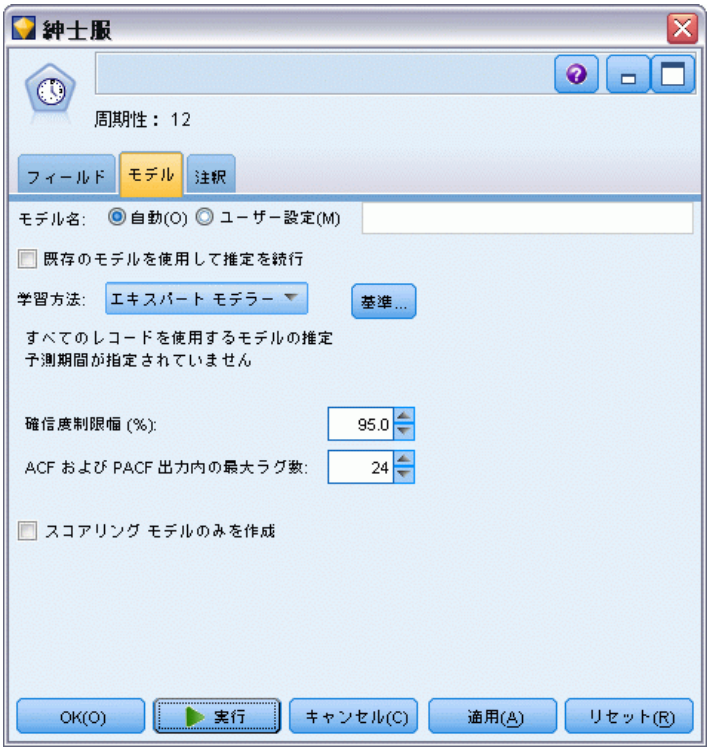
ARIMA 法では、自己回帰、差別化、および移動平均の順序、およびもちろんこれらのコンポーネントの季節性の同等物の順序も指定することによりモデルを微調整できます。これらのコンポーネントの最適値を手作業で決めることは、多くの試行錯誤を伴う時間浪費プロセスであり、この例については、エキスパート モデラーに ARIMA モデルを選択させることにします。

予測値変数としてデータ セット内の他の変数をいくつか処理することにより、優れたモデルを試して構築します。予測値として含めるのに最も役立つと思われるものは、郵送するカタログの数（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）、広告印刷物に要した費用（印刷）、およびカスタマー サービス担当者の数（サービス）です。

図 15-13
予測値 フィールドの設定

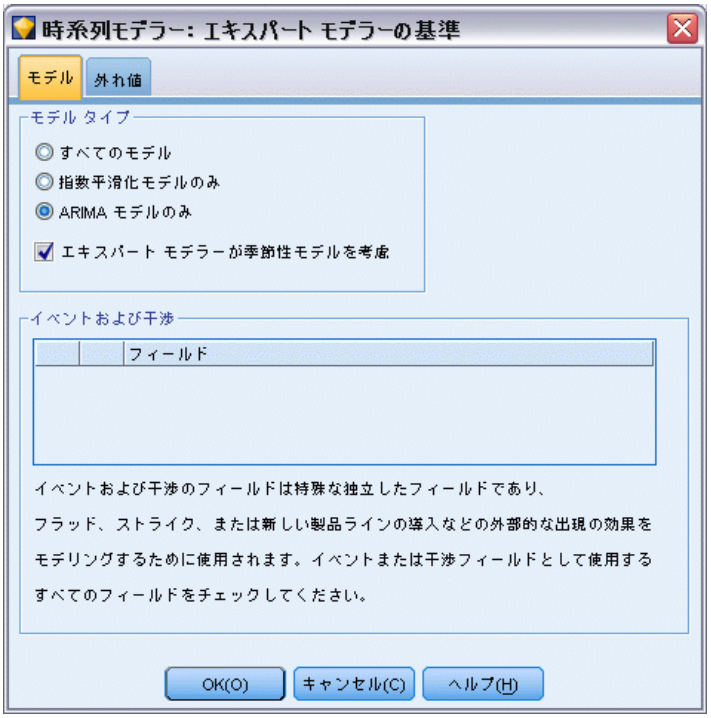
- IBM® SPSS® Statistics ファイル入力ノードを開きます。
- ［タイプ］タブで、［郵送］の［役割］、［ページ］、［電話］、［印刷］、および［サービス］を［**入力**］に設定します。
- ［**男性**］の役割が［**対象**］に設定されていることおよびその他のすべてのフィールドが［**なし**］に設定されていることを確認します。
- ［OK］をクリックします。

図 15-14
エキスパート モデラーの選択



- 時系列ノードを開きます。
- ［モデル］タブで、**[方法]** を **[エキスパート モデラー]** に設定して、**[基準]** をクリックします。

図 15-15
ARIMA モデルのみの選択



- ［エキスパート モデラー］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]** がチェックされていることを確認します。
- [OK] をクリックして、ダイアログ ボックスを閉じます。
- ［モデル］タブの［**実行**］をクリックして、モデル ナゲットを再作成します。

図 15-16
エキスパート モデラーによる 2 種類の予測値の選択

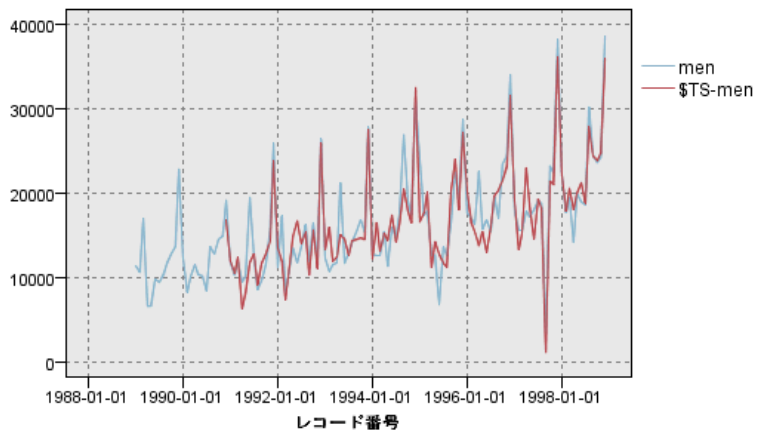
▶ モデル ナゲットを開きます。

モデルにとって重要な 5 個の指定予測値のうちの 2 個のみをエキスパート モデラーがどのようにして選択したのかについて留意してください。

▶ [OK] をクリックして、モデル ナゲットを閉じます。

▶ 時系列ノードを開いて、[実行] をクリックします。

図 15-17
指定された予測値による ARIMA モデル



このモデルは、大きな下向きの山形を取得してそれを最適な状態にすることにより、以前のモデルを改善します。

モデルはさらに改良できますが、この時点からの改善は最小限になると思われます。予測値による ARIMA モデルが望ましいことを立証したので、構築したばかりのモデルを使用してみましょう。この例の目的に合わせて、来年の売り上げを予測します。

▶ [OK] をクリックして、時系列ウィンドウを閉じます。

▶ 時間区分ノードを開いて、［予測］タブを選択します。

▶ ［レコードの将来への拡張］チェック ボックスを選択して、その値を 12 に設定します。

予測を行う場合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、予測期間中にそれらのフィールドの推定値を指定する必要があります。

図 15-18
予測フィールド用の将来の値の指定

▶ **［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。

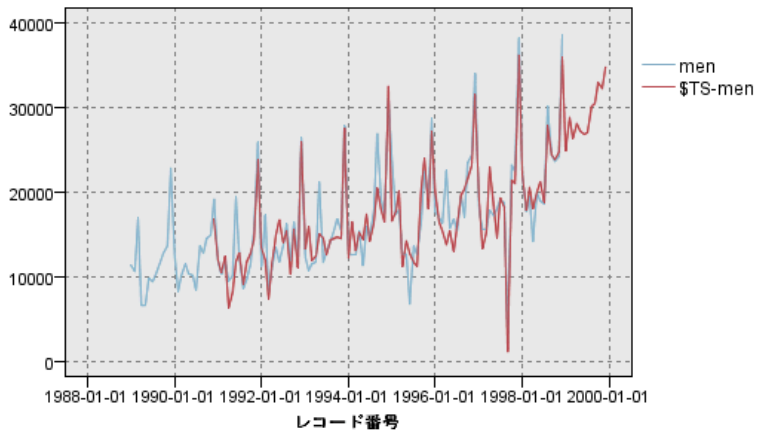
▶ ［フィールドの選択］ダイアログで **［郵送］**から **［サービス］**までを選択し、［OK］をクリックします。

実世界では、これらの 5 種類の予測値はすべて制御下にあるアイテムに関連しているため、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ついて 12 種類の値を指定するという無駄をなくします。（この例にさらに精通しているのであれば、モデルに与える影響を調べるためにさまざまな将来の値で実験を試してみるとよいでしょう。）

- 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。
- [OK] をクリックします。
- 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。
- 時系列ノードを開いて、[実行] をクリックします。

  1999 年度の予測は予想通り良好のように見え、12 月のピークに続いて通常の販売レベルの利益があり、一般的には前年度よりも著しく高い売り上げになるため、その年の第 2 半期は堅実な増加傾向を示しています。

図 15-19
指定予測値による売り上げ予測



要約

増加傾向だけでなく季節変動やその他の変動も組み込んでいる複雑な時系列を問題なくモデル化しました。試行錯誤により、将来の売り上げを予測するのに使用した正確なモデルに徐々に近づくことができます。

実際には、実際の販売データが更新された場合は（たとえば、月ごとあるいは四半期ごとの例）、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。 詳細は、 14 章 p. 221 時系列モデルの再適用 を参照してください。

章
16
顧客へのオファー提供（自己学習）

自己学習応答モデル (SLRM) ノードは、どのオファーが顧客に最も適しているか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、オファーが受け入れられる確率を表示します。これらの種類のモデルは、マーケティング アプリケーションやコール センターなどの顧客対応管理で最も役立ちます。

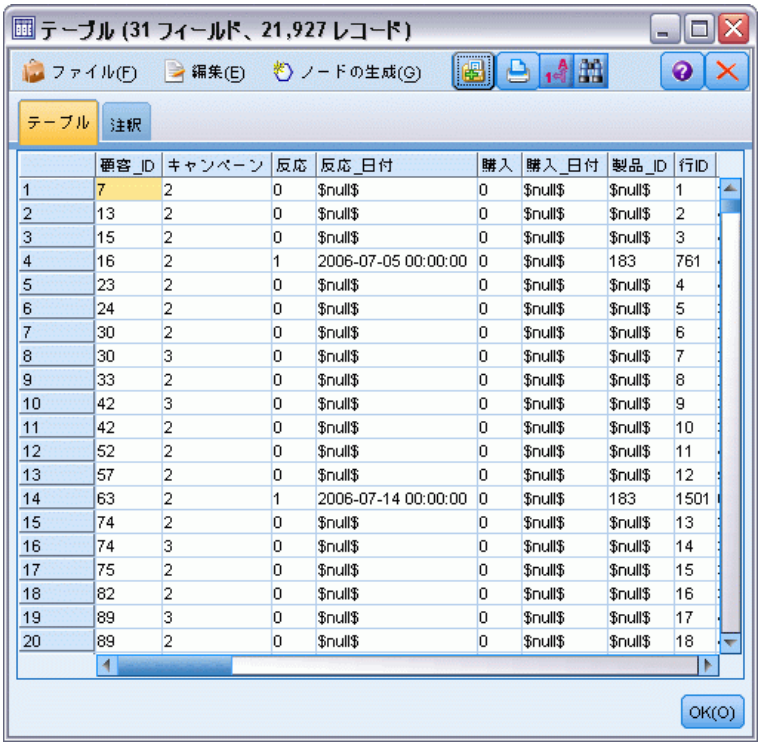
この例では、架空の銀行を使用します。マーケティング部門では、それぞれの顧客に合った適切な金融サービスの提案を行うことで、今後さらに収益を上げることを望んでいます。特に、この例では、自己学習応答モデルを使用して、以前のオファーおよび応答を基に、顧客が最も好意的な反応を示す特徴を識別し、その結果に基づいて最良の現在のオファーを作成します。

この例では、ストリーム pm_selflearn.str を使用し、データ ファイル pm_customer_train1.sav、pm_customer_train2.sav および pm_customer_train3.sav を参照します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos フォルダにあります。これは、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オファー、およびそれらのオファー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。

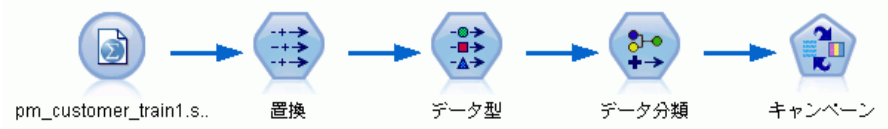
図 16-1
以前のオファーに対する応答



ストリームの構築

- IBM® SPSS® Modeler インストール フォルダの Demos フォルダにある、pm_customer_train1.sav を示す Statistics ファイル入力ノードを追加します。

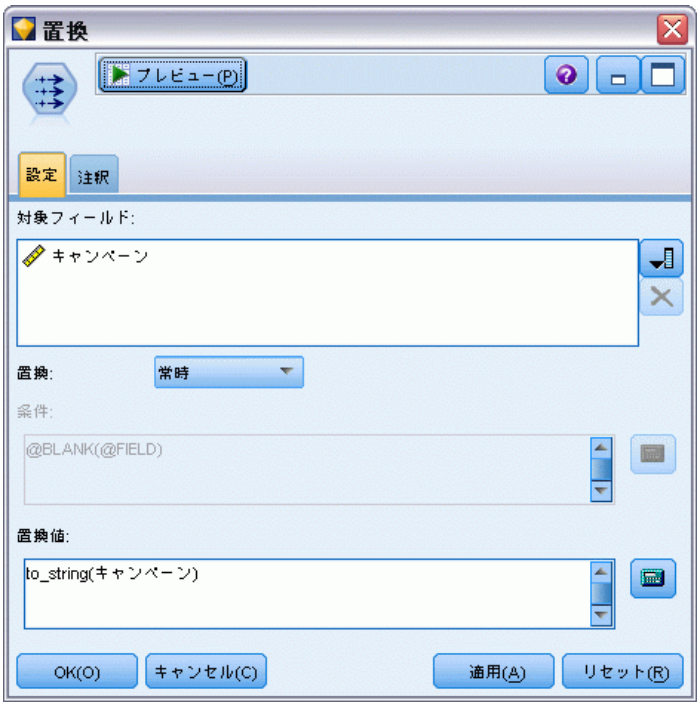
図 16-2
SLRM サンプル ストリーム



- 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。

- ▶ **常時** の置換タイプを選択します。
- ▶ ［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 16-3
キャンペーン フィールドの作成



▶ 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 フィールドの 役割を [なし] に設定します。

図 16-4
データ型の設定変更

▶ campaign 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 [尺度] を [フラグ型] に設定します。

▶ [値の読み込み] をクリックし、次に [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 および 4) として表示されるため、フィールドを再分類してより重要なタイトルをつけることができます。

▶ データ分類ノードをデータ型ノードに追加します。

▶ [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。

▶ [データ分類フィールド] リストで、[キャンペーン] を選択します。

▶ [取得] ボタンをクリックします。キャンペーン値が [元の値] 列に追加されます。

▶ [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**

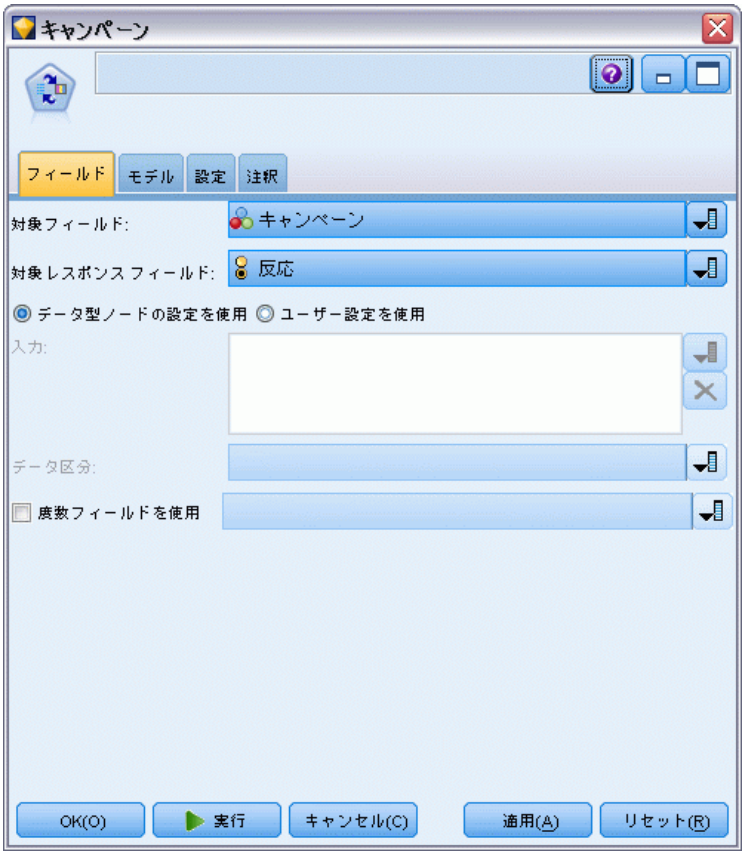
- カー ローン
- 貯金
- 年金

▶ [OK] をクリックします。

図 16-5
キャンペーン名のデータ分類

▶ SLRM モデル作成ノードをデータ分類ノードに添付します。[フィールド] タブで、対象フィールドに [**キャンペーン**] を選択し、対象レスポンス フィールドに [**レスポンス**] を選択します。

図 16-6
対象および対象レスポンスの選択

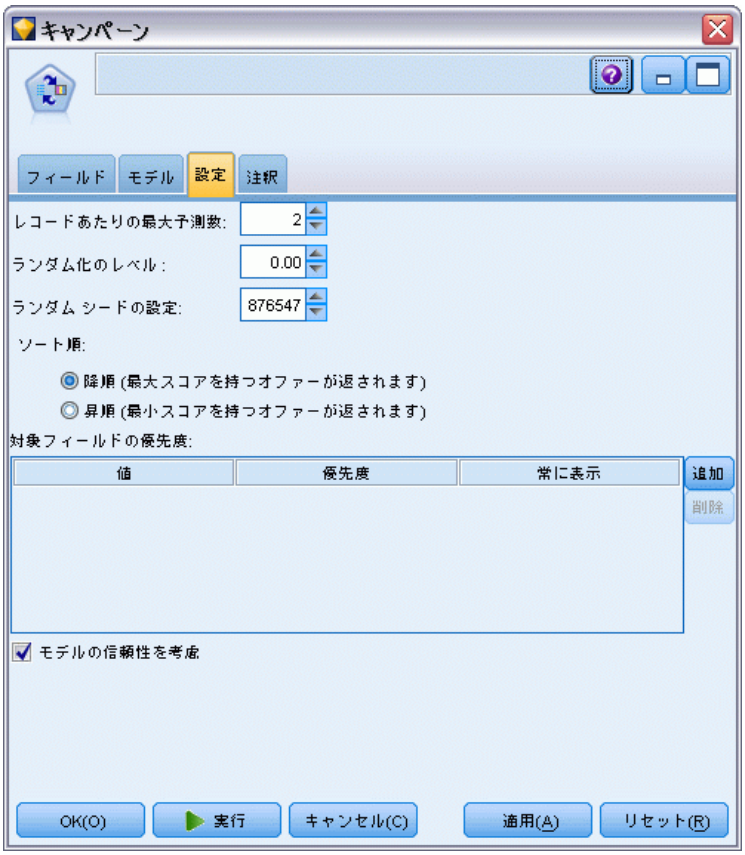


▶ [設定] タブの [レコードあたりの最大予測数] フィールドで数を 2 に減少させます。

これにより、各顧客に対し受け入れられる最も高い確率を持つと特定された 2 つのオファーが存在することになります。

► [モデルの信頼性を考慮] を選択して、[実行] をクリックします。

図 16-7
SLRM ノード設定



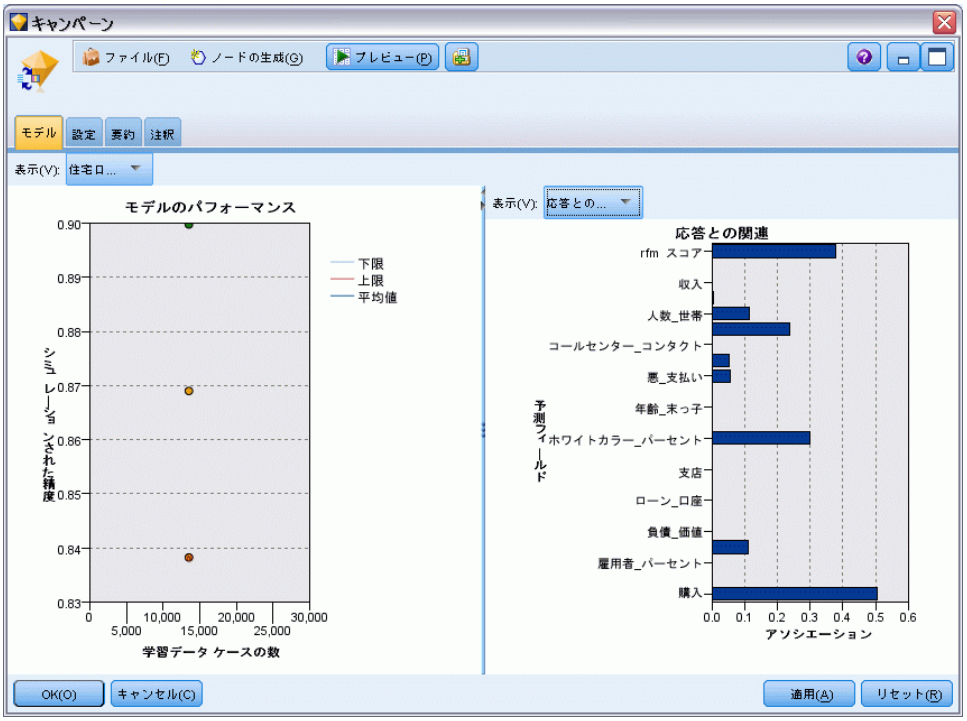
モデルの参照

- モデル ナゲットを開きます。[モデル] タブには最初、各オファーの予測フィールドの推定された精度およびモデル推定時の予測フィールドの相対重要度を表示します。

  各予測値と対象変数の相関を表示するには、右側のパネルの [**表示**] リストから [**応答との関連**] を選択します。

- 予測を実行した4 つの各オファーの間で表示を切り替えるには、右側のパネルにある [**表示**] リストから必要なオファーを選択します。

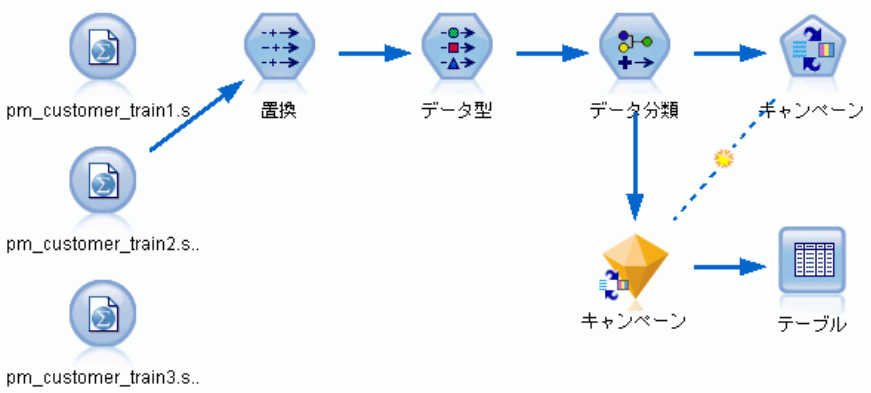
図 16-8
SLRM モデル ナゲット



- モデル ナゲット ウィンドウを閉じます。
- 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。

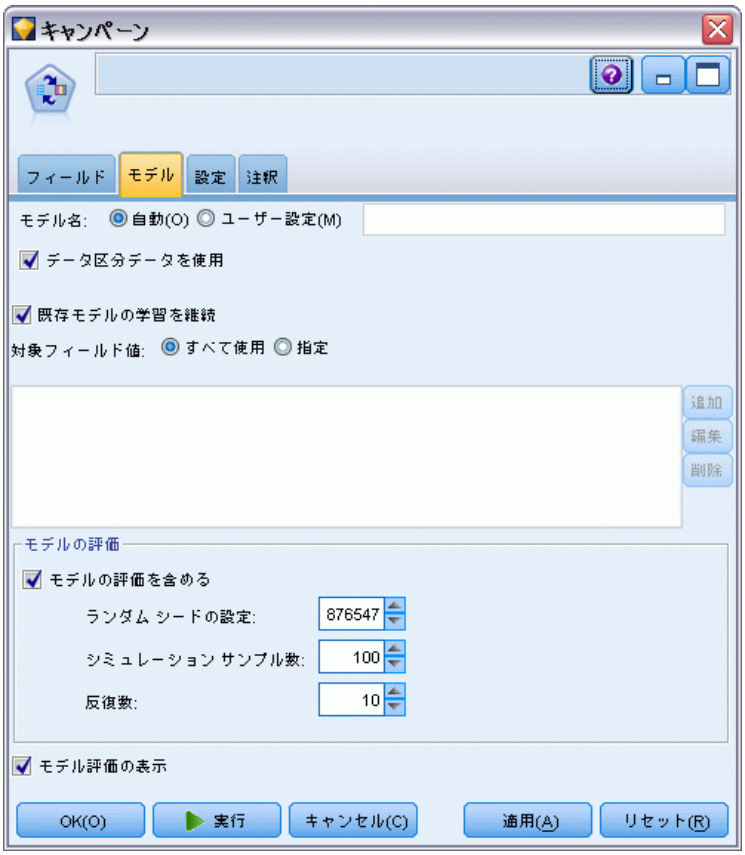
▶ IBM® SPSS® Modeler インストール フォルダの Demos フォルダにある、pm_customer_train2.sav を示す Statistics ファイル入力ノードを追加し、置換ノードに接続します。

図 16-9
2 番目のデータ ソースを SLRM ストリームに追加します。



▶ 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 16-10
モデル学習の継続

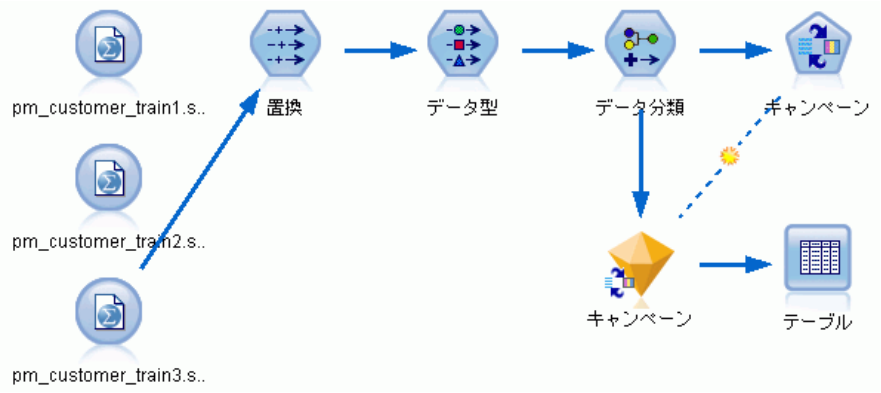


▶ **［実行］**をクリックして、モデル ナゲットを再作成します。詳細を表示するには、領域のナゲットをダブルクリックします。

［モデル］タブでは、各オファーの予測フィールドに関する精度の取得された推定値を表示します。

- SPSS Modeler インストール フォルダの Demos フォルダにある、pm_customer_train3.sav を示す Statistics ファイル入力ノードを追加し、置換ノードに接続します。

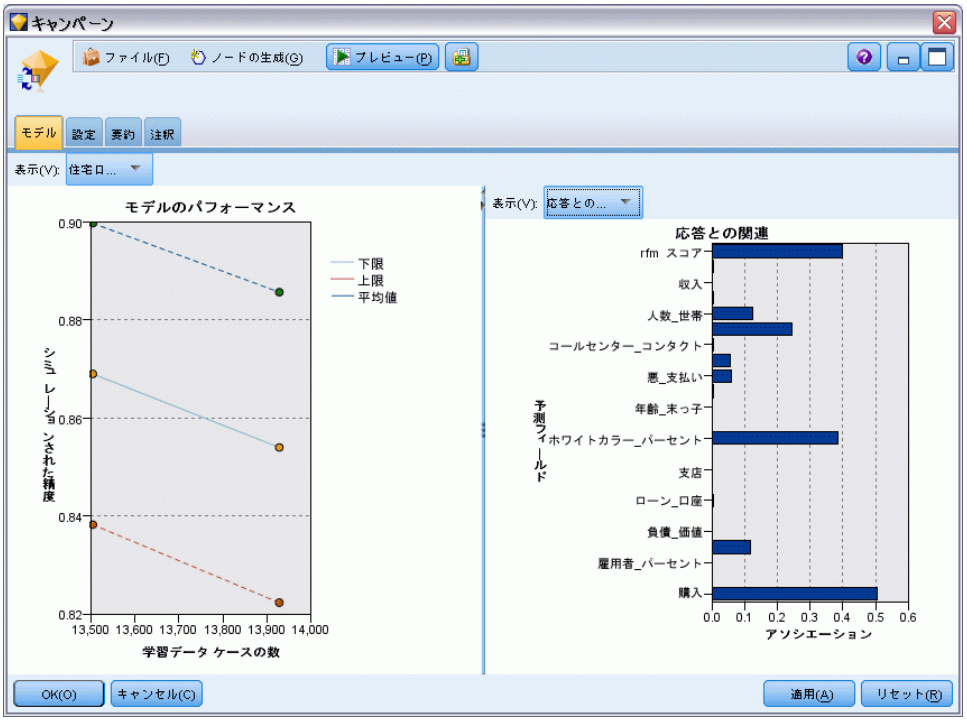
図 16-11
3 番目のデータ ソースを SLRM ストリームに追加します。



- ［**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。
- ［モデル］タブでは、各オファーの予測フィールドに関する精度の最終的な推定値を表示します。

表示されたとおり、平均精度は追加したデータソースの分わずかに減少します（86.9% から 85.4%）。ただし、この変動は最小値で、利用可能なデータ内のわずかな矛盾によるものです。

図 16-12
更新された SLRM モデル ナゲット

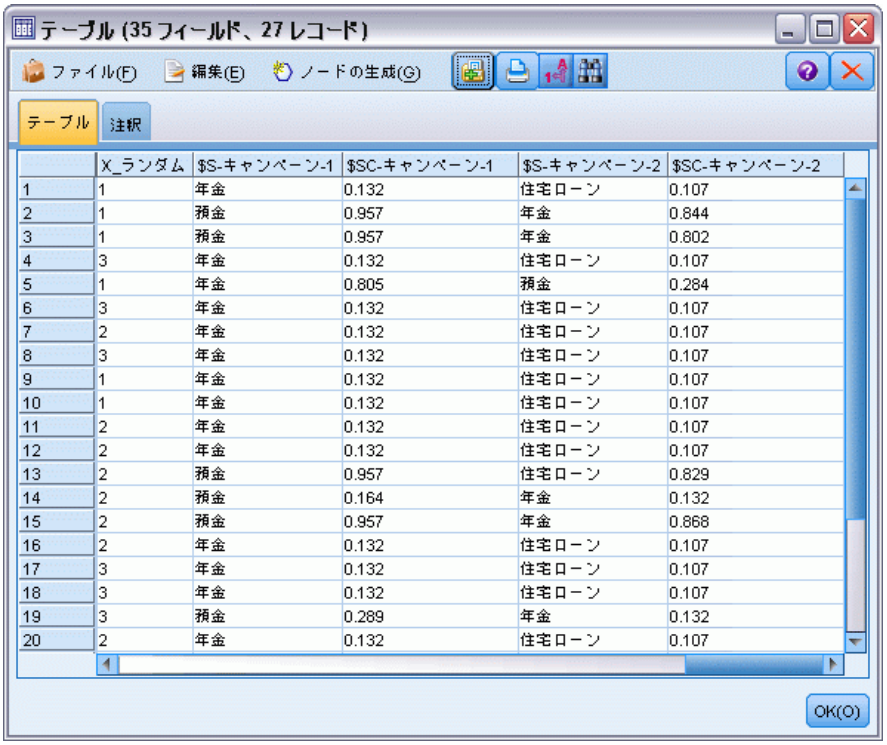


- テーブル ノードを最後（3 番目）に生成されたモデルに接続し、テーブル ノードを実行します。
- テーブルの右へスクロールします。予測フィールドでは、各顧客の詳細に応じて、顧客が最も受け入れるオファーおよび受け入れる確信度を表示します。

  たとえば、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オファーした場合に受け入れる割合はわずか 13.2% の確信度（$SC-campaign-1 列の値 0.132）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表示されています。彼らのケースでは、確信度が 95.7%で、同じ履歴データの顧客はオファーを

受けた場合預金口座を開設しています。そして 80%を超える確信度で、彼らは年金を受け取っています。

図 16-13
モデル出力 - 予測されたオファーと確信度


テーブル (35 フィールド、27 レコード)

ファイル(F) 編集(E) ノードの生成(G)

テーブル 注釈

	X_ランダム	$S-キャンペーン-1	$SC-キャンペーン-1	$S-キャンペーン-2	$SC-キャンペーン-2
1	1	年金	0.132	住宅ローン	0.107
2	1	預金	0.957	年金	0.844
3	1	預金	0.957	年金	0.802
4	3	年金	0.132	住宅ローン	0.107
5	1	年金	0.805	預金	0.284
6	3	年金	0.132	住宅ローン	0.107
7	2	年金	0.132	住宅ローン	0.107
8	3	年金	0.132	住宅ローン	0.107
9	1	年金	0.132	住宅ローン	0.107
10	1	年金	0.132	住宅ローン	0.107
11	2	年金	0.132	住宅ローン	0.107
12	2	年金	0.132	住宅ローン	0.107
13	2	預金	0.957	住宅ローン	0.829
14	2	預金	0.164	年金	0.132
15	2	預金	0.957	年金	0.868
16	2	年金	0.132	住宅ローン	0.107
17	3	年金	0.132	住宅ローン	0.107
18	3	年金	0.132	住宅ローン	0.107
19	3	預金	0.289	年金	0.132
20	2	年金	0.132	住宅ローン	0.107

OK(O)

SPSS Modeler で使用されるモデル作成方法の数学的な基礎については、製品 DVD の ¥Documentation ディレクトリにある『SPSS Modeler アルゴリズム ガイド』で説明されています。

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をどれだけうまく一般化しているかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。詳細は、 4 章 データ区分ノード in IBM SPSS Modeler 15 入力ノード、プロセス ノード、出力ノード を参照してください。SLRM ノードの詳細は、ノード リファレンスの 14 章 を参照してください。

章

17

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」という実際の知識を組み合わせることによって確率モデルを作成し、表面的にはリンクしていない属性を使用して発生の尤度を確立できます。

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 インストールの Demos ディレクトリから使用でき、Windows [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 ディレクトリにあります。

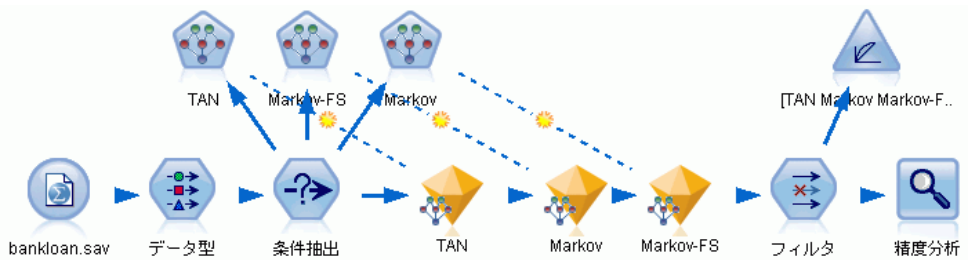
たとえば、銀行がローンが返済されない可能性について心配しているとします。以前のローン返済不能データを使用してローン返済について問題のあると考えられる顧客を予測することができる場合、こうした「リスクのある」顧客の融資を減らしたり、代わりの製品を提案したりすることができます。

この例は、既存のローン返済不能データの使用に焦点を当てて将来の返済不能の可能性を予測、3 つの異なる Bayesian Network 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

- bankloan.sav を指定する Statistics ファイルの入力ノードを Demos フォルダに追加します。

図 17-1
Bayesian Network のサンプル ストリーム



- データ型ノードを入力ノードに追加して、**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。

▶ [値の読み込み] ボタンをクリックして［値］列を読み込みます。

図 17-2
対象フィールドの選択

対象がヌル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。

▶ 条件抽出ノードをデータ型ノードに追加します。

▶ モードに対し、[破棄] を選択します。

▶ ［条件］ボックスに default = ‘$null$’ と入力します。

図 17-3
ヌル値を持つ対象の破棄

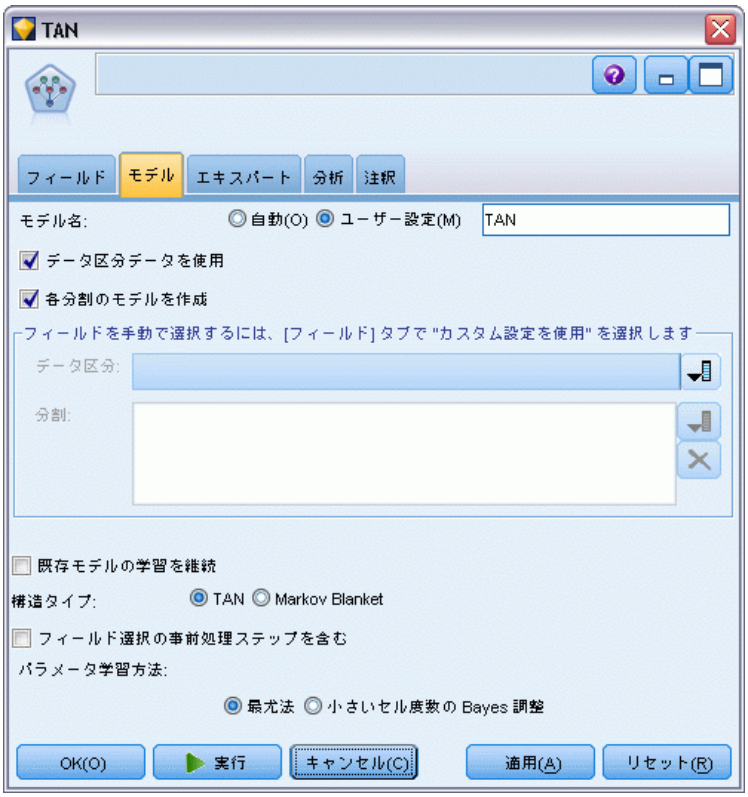
さまざまなタイプのベイズ ネットワークを構築できるため、複数のモデルを比較して最善の予測を提供するモデルを確認する価値があります。Tree Augmented Naïve Bayes (TAN) モデルを最初に作成します。

▶ Bayesian Network ノードを条件抽出ノードに接続します。

▶ ［モデル］タブでモデル名について、**[カスタム]** を選択してテキスト ボックスに **TAN** と入力します。

▶ 構造タイプについて、TAN を選択し、[OK] をクリックします。

図 17-4
Tree Augmented Naïve Bayes モデルの作成



2 番目に構築するモデル タイプは Markov Blanket 構造です。

▶ 2 番目の Bayesian Network ノードを条件抽出ノードに接続します。

▶ ［モデル］タブでモデル名について、[**カスタム**] を選択してテキスト ボックスに Markov と入力します。

▶ 構造タイプについて、Markov Blanket を選択し、[OK] をクリックします。

図 17-5
Markov Blanket モデルの作成

3 番目に構築するのは Markov Blanket 構造で、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。

▶ 3 番目の Bayesian Network ノードを条件抽出ノードに接続します。

▶ [モデル] タブでモデル名について、[**カスタム**] を選択してテキスト ボックスに Markov-FS と入力します。

▶ 構造タイプについて、[Markov Blanket] を選択します。

▶ [フィールド選択の事前処理ステップを含む] を選択して [OK] をクリックします。

図 17-6
フィールド選択事前処理を使用する Markov Blanket の作成

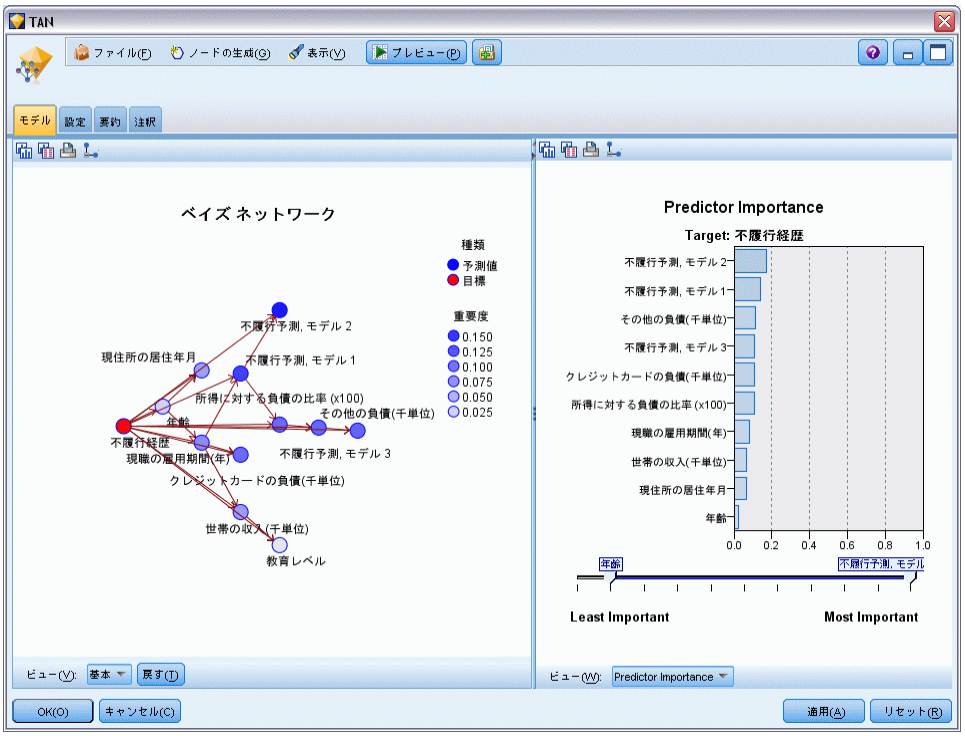
モデルの参照

▶ ストリームを実行してモデル ナゲットを作成します。ストリームおよび右上の［モデル］パレットに追加されます。この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。

モデル ナゲットの［モデル］タブは、次の 2 つのパネルに分けられます。左側のパネルには、予測値間の関係に加え、対象値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。

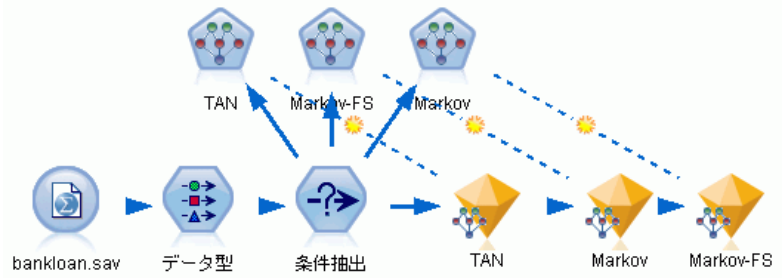
右側のパネルにはモデル推定時に各予測値の相対重要度を示す予測値の重要度、または各ノード値の条件確率値や親ノードの値の組み合わせを含む条件確率が表示されます。

図 17-7
Tree Augmented Naïve Bayes モデルの表示



- 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）。
- 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）。
- 3 つのナゲットを見やすいように条件抽出ノードで配置します。

図 17-8
ストリーム内のナゲットの配置



- 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 ノードを Markov-FS モデル ナゲットに接続します。
- 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。

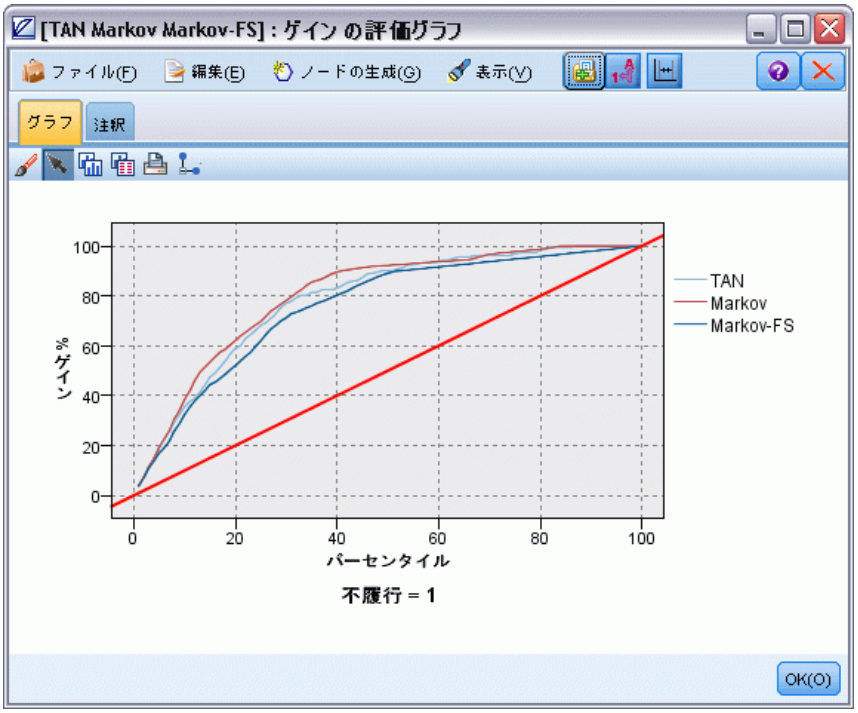
図 17-9
モデル フィールド名の変更

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。

- 評価グラフ ノードをフィルタ 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ますが、Markov モデルが少し良い結果を作成しています。

図 17-10
モデルの精度の評価



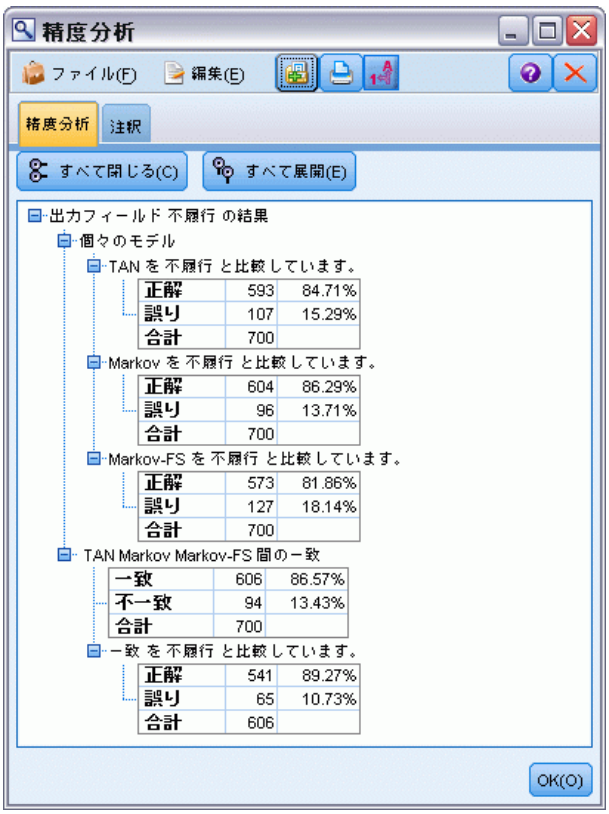
各モデルがどれほど良い予測を行っているかを確認するために、評価グラフではなく精度分析ノードを使用します。これにより、正しい予測と不正な予測の両方の割合によって精度を示します。

▶ 精度分析ノードをフィルタ 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、Markov モデルの方が正しい予測を行っていることを示しています。ただし、Markov-FS モデルが Markov モデルより数 % だけ上回っています。この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ます。

図 17-11
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については、インストール ディスクの ¥Documentation ディレクトリにある『SPSS Modeler アルゴリズム ガイド』で説明されています。

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をどれだけうまく一般化しているかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。 詳細は、 4 章 データ区分ノード in IBM SPSS Modeler 15 入力ノード、プロセス ノード、出力ノード を参照してください。

章
18
毎月ベースのモデルの再学習 (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」という実際の知識を組み合わせることによって確率モデルを作成し、表面的にはリンクしていない属性を使用して発生の尤度を確立できます。

この例では、bayes_churn_retrain.str という名前のストリームを使用します。このストリームは、telco_Jan.sav および telco_Feb.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 インストールの Demos ディレクトリから使用でき、Windows [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 ディレクトリにあります。

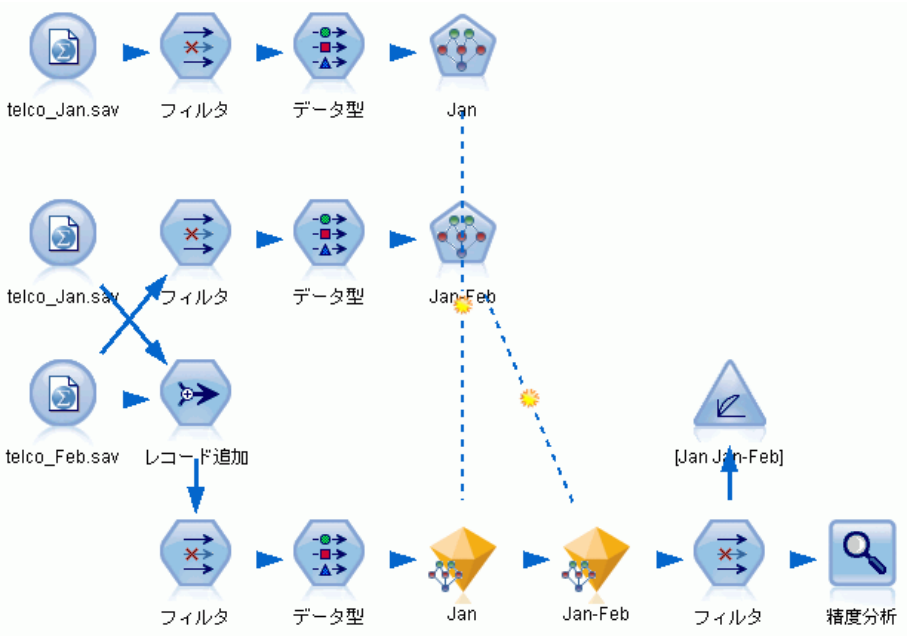
たとえば、競合他社に奪われる顧客の数（解約）に関して、電気通信プロバイダが心配しているとします。過去の顧客データを使用して将来解約すると考えられる顧客を予測することができる場合、これらの顧客を誘導、または他社のサービス プロバイダへの移行を思いとどまらせるような提案の対象とすることができます。

この例では、将来解約すると考えられる顧客を予測するための既存の月ごとの解約データの使用、および来月のデータを追加してモデルを再調整および再学習することに焦点を当てています。

ストリームの構築

- telco_Jan.sav を指定するStatistics ファイル入力ノードを Demos フォルダに追加します。

図 18-1
Bayesian Network のサンプル ストリーム



以前の分析は、解約の予測時にいくつかのデータ フィールドは重要度が少ないことを示していました。これらのフィールドをデータ セットから除外して、モデルの構築およびスコアリング時に処理の速度を向上することができます。

- フィルタ ノードを入力ノードに追加します。
- 住所、年齢、解約、custcat、学歴、職業、性別、結婚、居住、退職、および 保有期間 を除くすべてのフィールドを除外します。

▶ [OK] をクリックします。

図 18-2
不要なフィールドのフィルタリング

▶ データ型ノードをフィルタ ノードに追加します。

▶ 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］列を読み込みます。

▶ 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、[解約] フィールドの測定レベルを [**フラグ型**] 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 18-3
対象フィールドの選択

複数の種類のベイズ 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 Naïve Bayes（TAN）モデルを構築します。これは、大きな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、強力な初期モデルを構築します。

▶ 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。

▶ [モデル] タブでモデル名について、[**カスタム**] を選択してテキスト ボックスに Jan と入力します。

▶ パラメータ学習方法で、[**小さいセルの度数のベイズ調整**] を選択します。

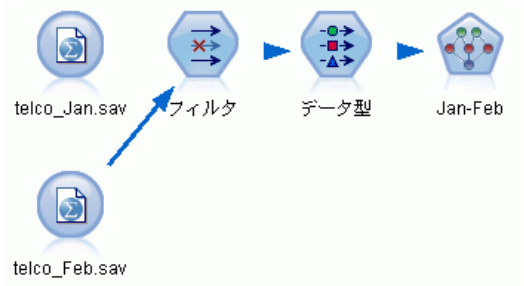
▶ [実行] をクリックします。モデル ナゲットがストリーム、および右上の［モデル］パレットに追加されます。

図 18-4
Tree Augmented Naïve Bayes モデルの作成

▶ telco_Feb.sav を指定するStatistics ファイル入力ノードを Demos フォルダに追加します。

▶ この新しい入力ノードをフィルタ 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 18-5
2 番目の月のデータの追加



- Bayesian Network ノードの［モデル］タブでモデル名について、[カスタム] を選択してテキスト ボックスに Jan-Feb と入力します。
- [既存モデルの学習を継続] を選択します。
- [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、右上の［モデル］パレットにも追加されます。

図 18-6
モデルの再学習

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。

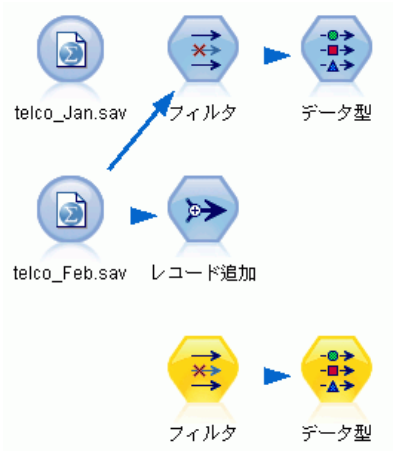
▶ 結合ノードを追加して、telco_Jan.sav および telco_Feb.sav 入力ノードを接続します。

図 18-7
2 つのデータソースの結合

▶ ストリームの初期のフィルタ 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。

▶ 結合ノードを新しくコピーされたフィルタノードに接続します。

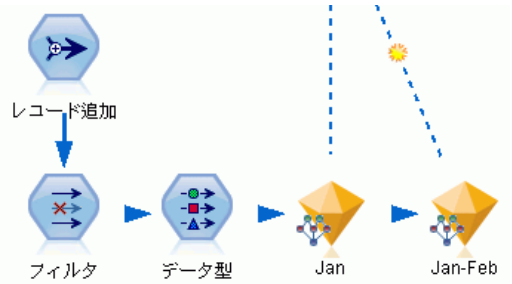
図 18-8
コピーされたノードのストリームへの貼り付け



2 つの Bayesian Network モデルのナゲットは、右上の［モデル］パレットにあります。

- 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接続します。
- ストリーム内にすでにある Jan-Feb モデル ナゲットを Jan モデル ナゲットに接続します。
- Jan モデル ナゲットを開きます。

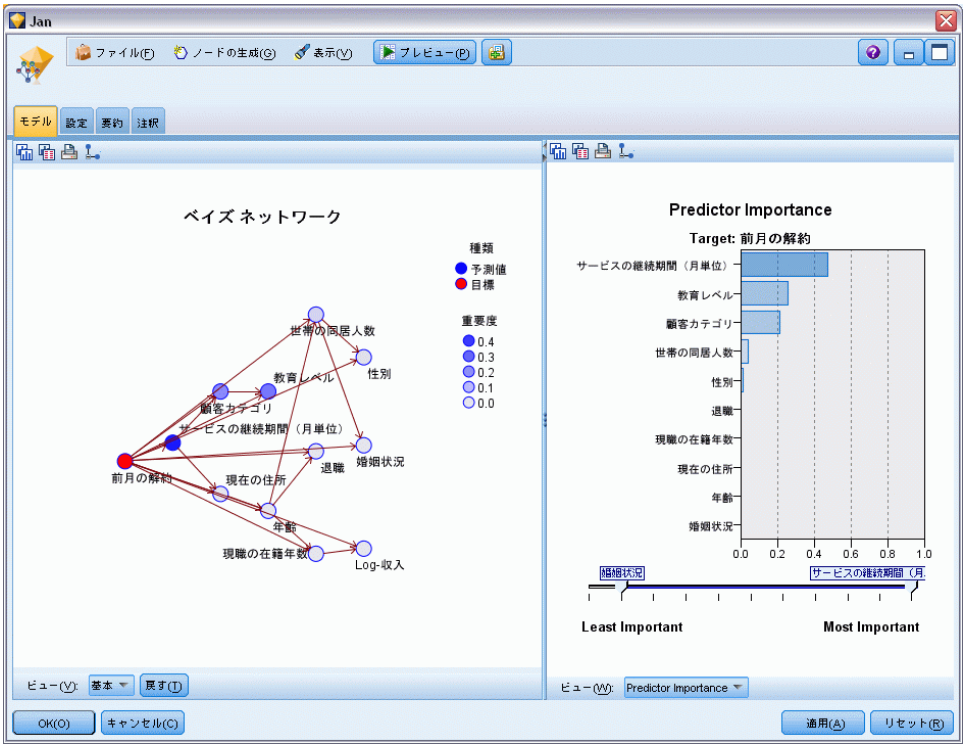
図 18-9
ナゲットのストリームへの追加



Bayesian Network モデルの［モデル］タブは、2 つの列に分けられます。左側の列には、予測値間の関係に加え、対象値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、または各ノード値の条件確率値や親ノードの値の組み合わせを含む条件確率が表示されます。

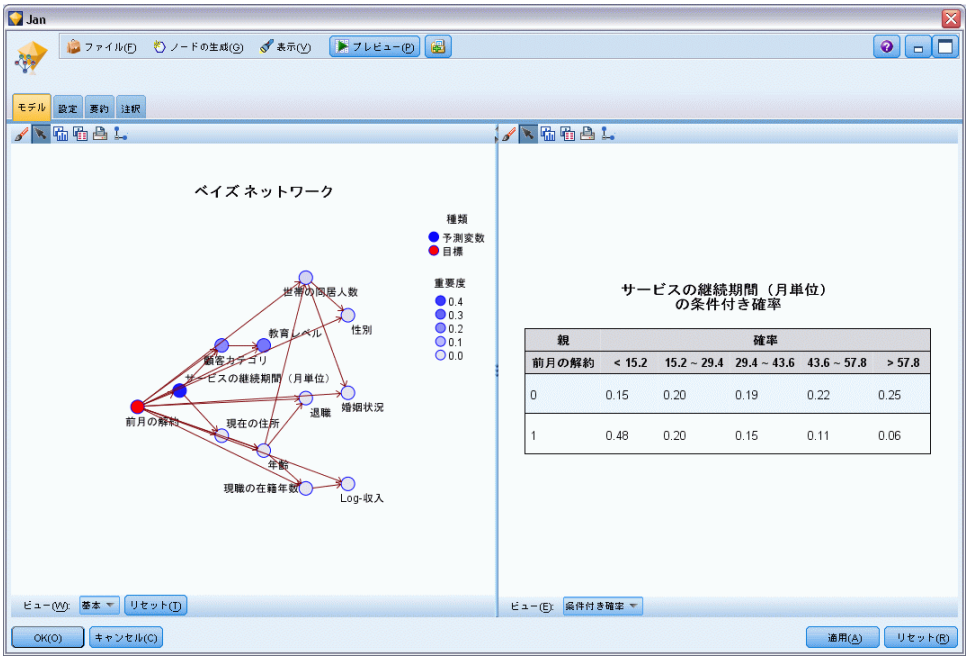
図 18-10
予測値の重要度を示す Bayesian Network モデル



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。右側の列を更新して、必要な詳細を表示します。

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 18-11
条件確率を表示する Bayesian Network モデル



- 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 ノードを Jan-Feb モデル ナゲットに接続します。

▶ 右側の［フィールド］列で、$B-churn を Jan に、$B1-churn を Jan-Feb に名前を変更します。

図 18-12
モデル フィールド名の変更

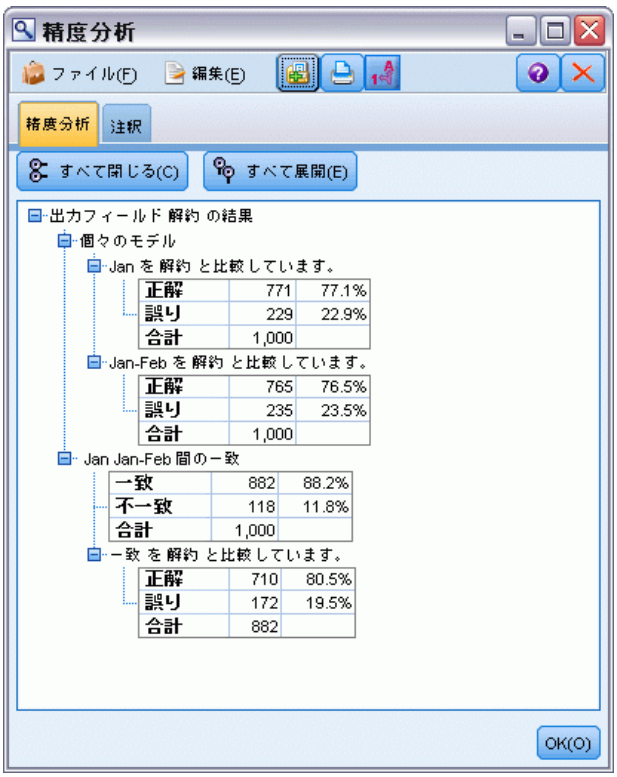
各モデルが解約をいかに正確に予測しているかチェックするには、精度分析ノードを使用します。これにより、正しい予測および不正な予測の割合によって精度が示されます。

▶ 精度分析ノードをフィルタ ノードに接続します。

▶ 精度分析ノードを開いて、[実行] をクリックします。

これにより、解約を予測する場合 2 つのモデルが同じ精度であることを示します。

図 18-13
モデルの精度分析



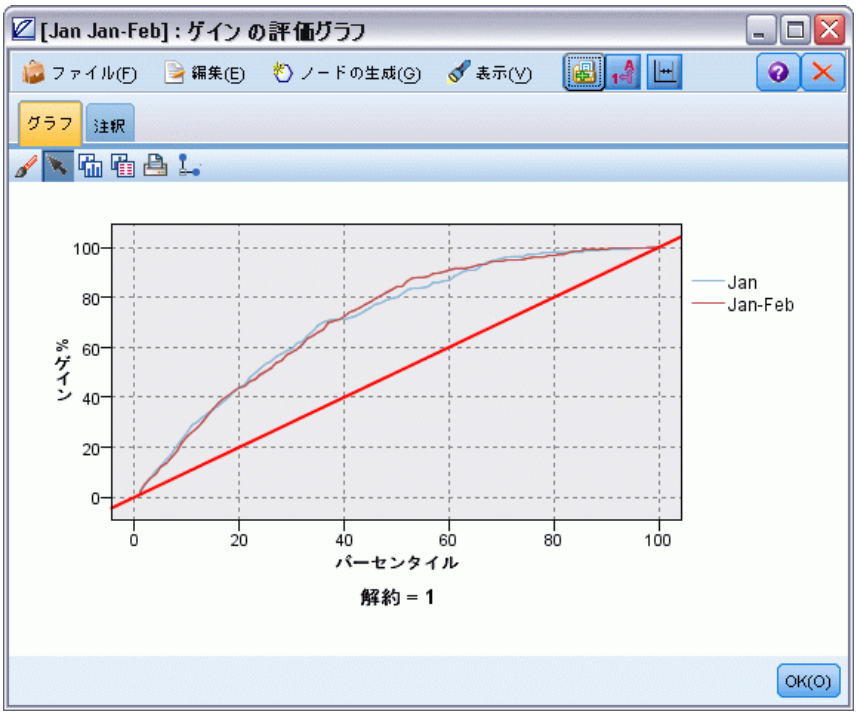
精度分析ノードの代わりとして、評価グラフを使用し、ゲイン グラフを作成したモデルの予測制度を比較することができます。

▶ 評価グラフ ノードをフィルタ ノードに接続します。

デフォルト設定を使用してグラフ ノードを実行します。

精度分析ノードと同様に、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ますが、2 か月のデータを使用する再学習モデルが予測の確信度が高いため、少し良い結果を作成しています。

図 18-14
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については、インストール ディスクの ¥Documentation ディレクトリにある『SPSS Modeler アルゴリズム ガイド』で説明されています。

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をどれだけうまく一般化しているかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。 詳細は、 4 章 データ区分ノード in IBM SPSS Modeler 15 入力ノード、プロセス ノード、出力ノード を参照してください。

章
19

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したデータを扱います(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。

この例は、goodsplot.str、goodslearn.str という名前のストリームを使用します。これらは GOODS1n および GOODS2n という名前のデータ ファイルを参照します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、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 フォルダに、goodslearn.str ファイルは streams ディレクトリにあります。

データの調査

各レコードには、次のデータが含まれています。

- 種別 :製品タイプ。
- コスト :単価。
- 販促 :販売促進活動に使用した金額の指標。
- 販促前の収益 :販売促進活動前の収益。
- 販促後の収益 :販売促進活動後の収益。

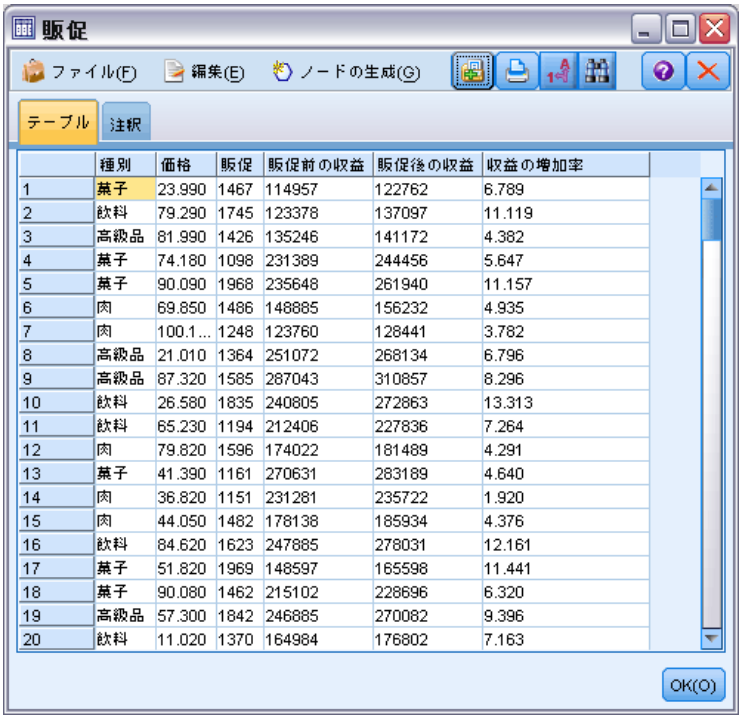
ストリーム goodsplot.str は、データを表示するための単純なストリームです。2 つの収益フィールド（販促前の収益と販促後の収益）は、単純に金額で表されています。しかし、販売促進活動後の収益増加（販売促進活動の効果と考えられる数値）の方が役に立つでしょう。

図 19-1
製品売上額への販売促進活動の効果

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

goodsplot.str には、販売促進活動前の収益と比較したパーセンテージを、[収益の増加率] フィールドを作成し、それをテーブルに表示するノードが含まれています。

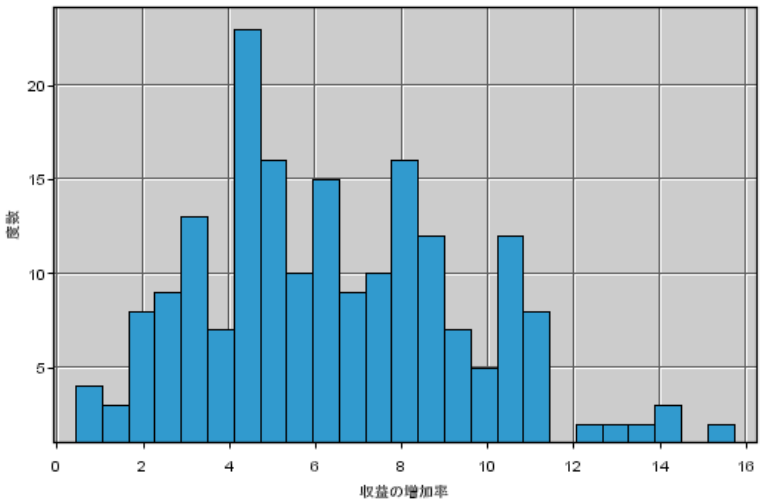
図 19-2
販売促進活動後の収益増加



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

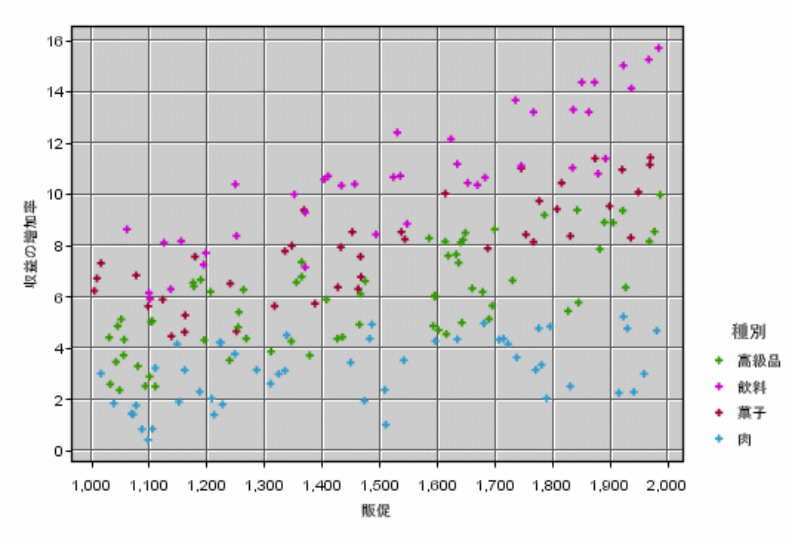
さらに、このストリームは、収益増加のヒストグラムと、販売促進活動にかかったコストと収益増加を比較した散布図を作成し、そこに関連する製品カテゴリをオーバーレイします。

図 19-3
収益増加のヒストグラム



散布図を見ると、各製品のクラスについて、収益増加と販売促進活動コスト増加の間に、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 ネットワークによって、他のフィールドの値から、ある程度正確に収益増加を予測できると考えられます。

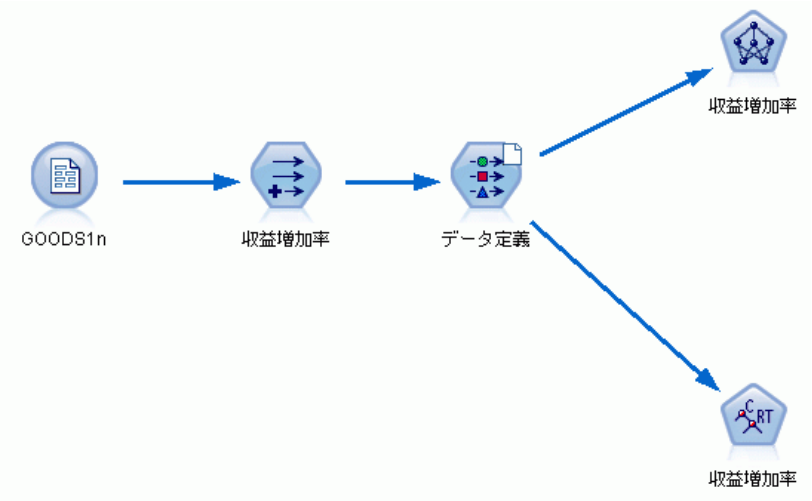
図 19-4
収益増加と販売促進活動コストの比較



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 19-5
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。テストするには、ディシジョン ツリーとニューラル ネットワークをデータ型ノードと精度分析ノードの間に直列に接続し、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（特に予測増加量と実際値の間の線型相関から）、システムが学習を行えば収益増加をかなり正確に予測できることがわかります。

さらに詳細に検証を行う場合は、学習を行ったシステムが比較的大きな誤差を持っている部分に焦点を当てるのがよいでしょう。誤差は、収益増加の予測値と実際値を対比してプロットすれば明らかになります。このグラフでの外れ値は、IBM® SPSS® Modeler のインタラクティブ グラフィックを使用して選択できます。また、それらのプロパティから、データ詳細や学習プロセスを調整して精度を向上させられる場合もあります。

章
20

稼動状況の監視 (ニューラル ネットワーク/C5.0)

この例では、機械のステータス情報の監視と、障害の認識やその予測に関する問題を取り上げていきます。データは架空のデータから作成され、時間の経過とともに測定された多数の連結系列で構成されています。各レコードは、次の情報を記録した、機械のスナップショット レポートになります。

- 時間 :整数。
- 電力量 :整数。
- 温度 :整数。
- 圧力 :通常は 0。瞬間の圧力についての警告がある場合は 1。
- 使用時間 :最後に修理、調整してからの経過時間。
- 状態 :通常は 0。不具合が発生した場合、エラー コード (101、202、または 303) を表示。
- 結果 :この時系列データに現れたエラー コード。エラーがない場合は 0(これらのコードは後になってから使えるようになります)。

この例は、COND1n および COND2n というデータ ファイルを参照する condplot.str および condlearn.str というストリームを使用します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、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 ディレクトリの中にあります。

各時系列データにおいて、平常に稼動している期間のレコードがあり、そのあとに、不具合が起こるまでの期間のレコードがあります。これは次の表のようになります。

Time	Power	温度	圧力	使用時間	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

Time	Power	温度	圧力	使用時間	ステータス	結果
			...			
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
			...			
208	644	251	0	209	0	101
209	640	251	0	209	101	101

次の過程は、データ マイニングのプロジェクトにおいて一般に行われるものです。

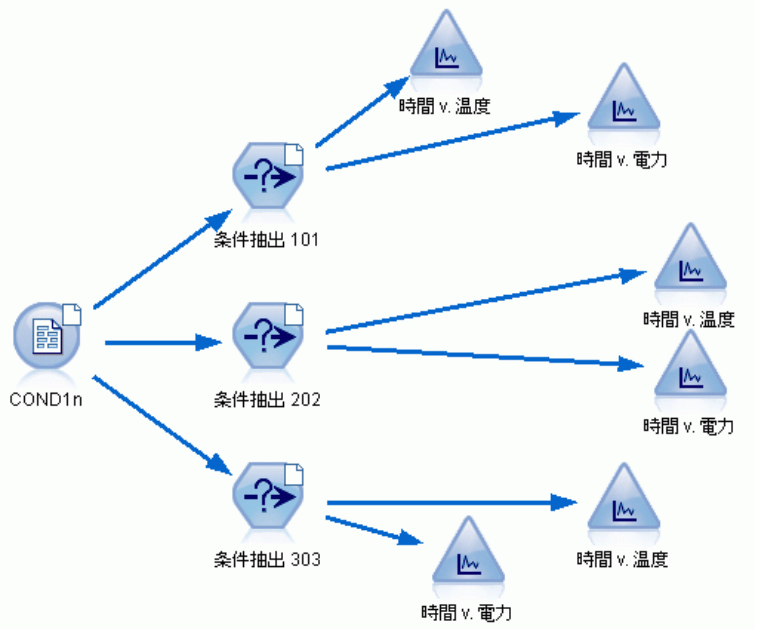
- まずデータを検証して、今ここで対象とする状態を予測したり認識したりするのに有効であると思われる属性を特定します。
- これらの属性を保持するか（すでに含まれている場合）、あるいは必要であればこれらの属性をデータに追加します。
- 結果として作成されたデータを、ルールとニューラル ネットの学習に使用します。
- 学習を行ったシステムを、別のテスト データを使用してテストします。

データの調査

ファイル condplot_j.str は、この過程の最初の部分を示しています。このファイルには、多数の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おいて明らかなパターンが見られる場合、不具合の発生状況を識別したり、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。次のストリームは、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する時系列のグラフをそれぞれ作図しま

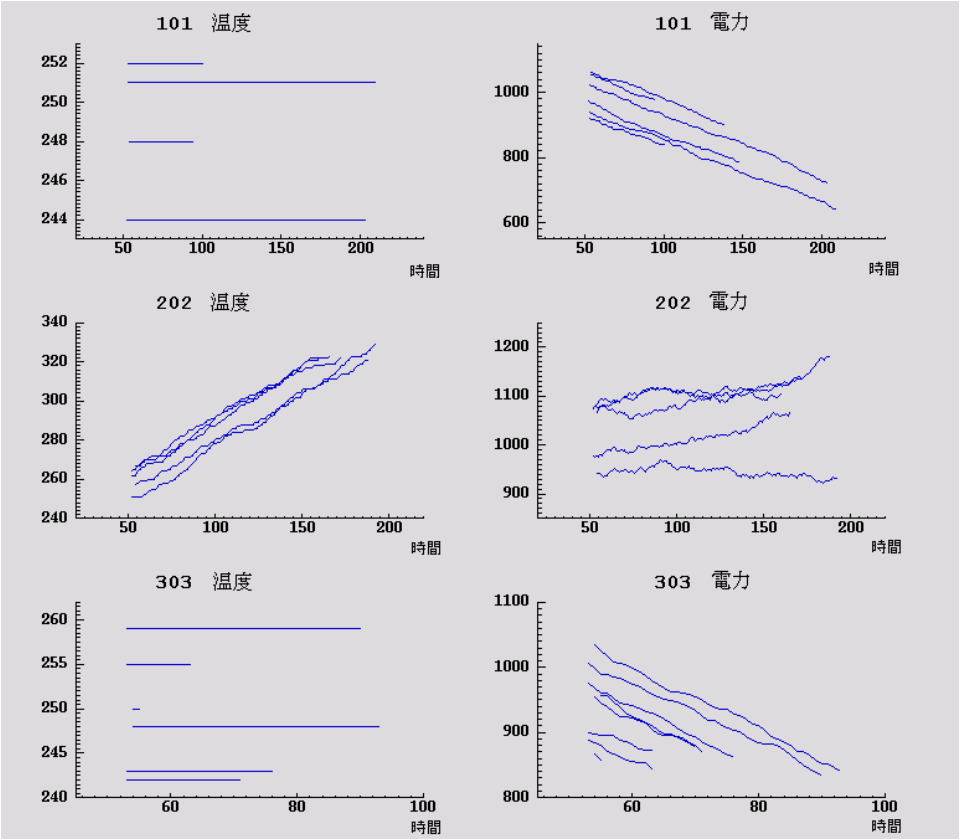
す。つまり合計 6 つのグラフが作成されることになります。条件抽出ノードによって、それぞれのエラー コードに関連するデータが選別されます。

図 20-1
Condplot ストリーム



このストリームの結果を図に示します。

図 20-2
時間の経過に伴う温度と電力量の変化



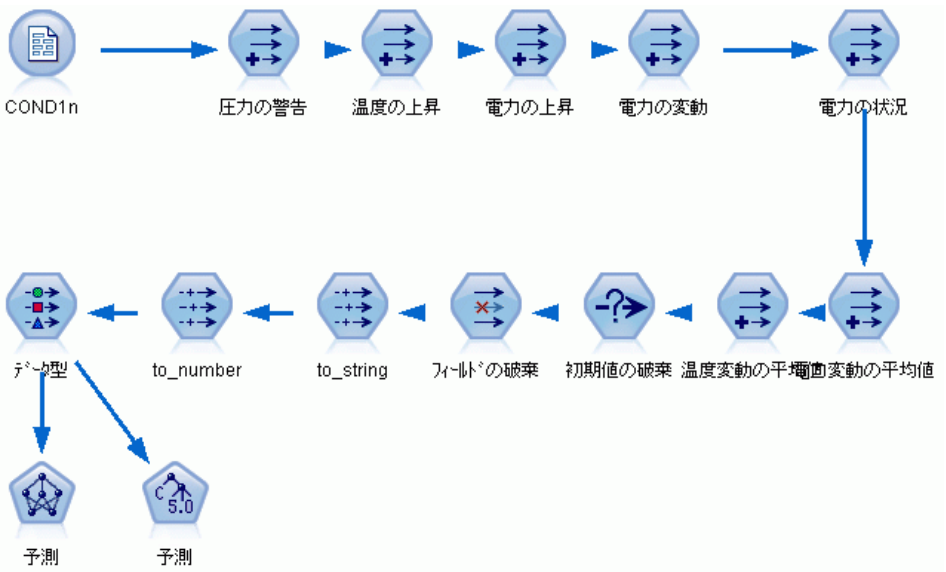
グラフは、202 エラーを 101 や 303 のエラーと明確に識別するパターンを示しています。202 エラーにおいては、時間の経過とともに温度は上昇し、また電力量は細かく変動しています。しかし他のエラーにはこのような動きは見られません。101 エラーと 303 エラーを区別するパターンは、さほど明確ではありません。どちらにおいても温度は一定であり、また電力量は減少しています。しかしながら 303 エラーのほうが電力量の減少が急激なようです。

これらのグラフから、温度や電力量が変化することやその変化の割合、また変動の存在とその程度などが、不具合を予測したり識別したりするのに役に立つことがわかります。そこで、学習システムを使用する前に、これらの属性をデータに追加する必要があります。

Data Preparation

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

図 20-3
Condlearn ストリーム



このストリームは、多数のフィールド作成ノードを使用してモデル作成用のデータを準備します。

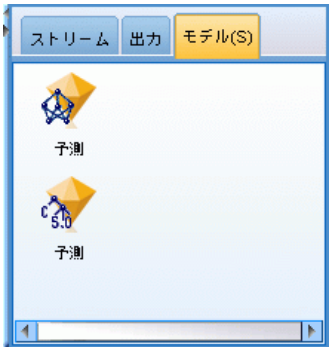
- **可変長ファイル ノード** :データ ファイル COND1n_j を読み込みます。
- **圧力警告のフィールド作成** :瞬間の圧力の警告数のカウント。時間が 0 に戻るとリセットされます。
- **温度上昇のフィールド作成** :@DIFF1 を使用して、温度の瞬間の変化率を計算します。
- **電力量増加のフィールド作成** :@DIFF1 を使用して、電力量の瞬間の変化率を計算します。
- **電力量変動のフィールド作成** :これはフラグであり、1 つ前のレコードから該当するレコードになった際に電力量がその変化の方向を変えたときに真 (true) となります。つまり電力量の最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 つの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** ：レコードを使用時間、状態、結果、圧力の警告、電力量状態、電力量変化、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。それに加えて、結果の測定レベルは**数値型**、圧力の警告は**連続型**、電力量状態を**フラグ型**として定義します。

学習

condlearn_j.str 中のストリームを実行すると、C5.0 ルールとニューラル ノード（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、使用に耐えう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。学習を終えたら、マネージャ ウィンドウの右上の［モデル］タブが点滅して、新しい 2 個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 20-4
モデル ナゲットとモデル マネージャ



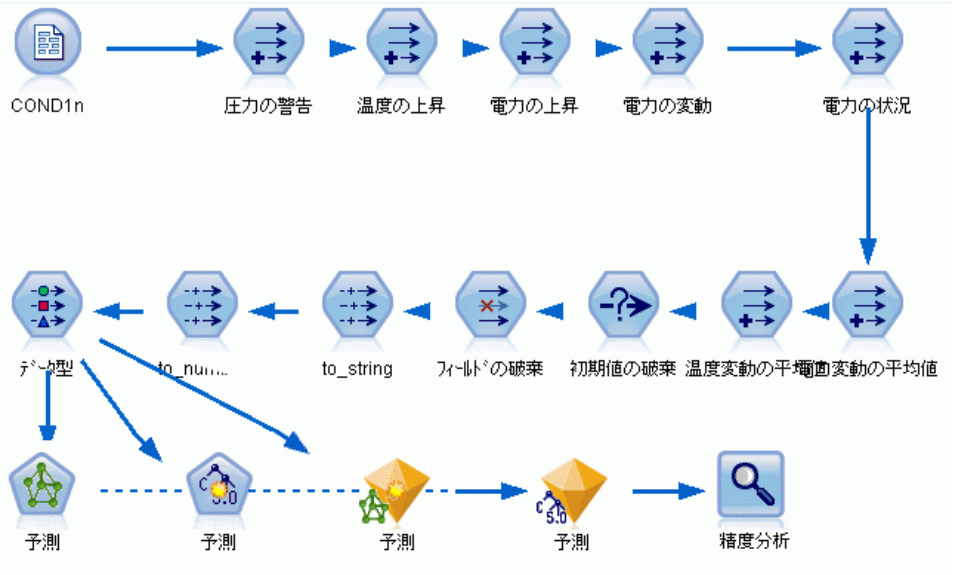
既存のストリームにモデル ナゲットを追加して、システムをテストしたり、モデルの実行結果をエクスポートすることができます。この例では、モデルの結果をテストしていきます。

テスト

モデル ナゲットをストリームに追加します。それらのいずれもデータ型ノードに接続しています。

- 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 ナゲットに接続し、C5.0 ナゲットにs悦属するようにします。
- 精度分析ノードを C5.0 ノードに接続します。
- 次に、COND2n_j ファイル (COND1n_j ではない) を読み込むために、元の入力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 20-5
学習済みのネットワークのテスト



- 精度分析ノードを開いて、[実行] をクリックします。

精度分析ノードを実行すると、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。

章
21

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、数値型フィールドではなくカテゴリフィールドを対象フィールドとします。

たとえば、電気通信プロバイダーがその顧客を、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ライズしているとします。顧客がどのグループに属するかを、人口統計データを使って予測できれば、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_custcat_discriminant.str という名前のストリームを使用します。このストリームは、telco.sav という名前のデータ ファイルを参照します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、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フォーカスします。対象フィールド custcat は、次のように 4 つの顧客グループに対応して 4 つの値を取ります。

値	Label
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

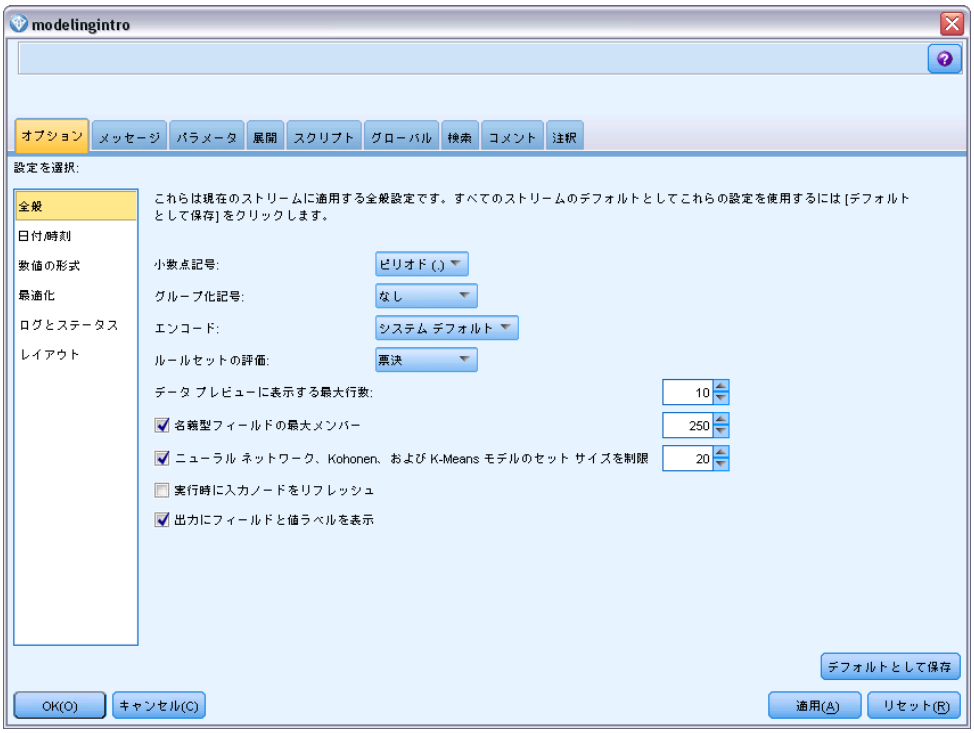
ストリームの作成

- 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を設定します。メニューから次の項目を選択します。

  File > **ストリームのプロパティ...** > **オプション** > **全般**

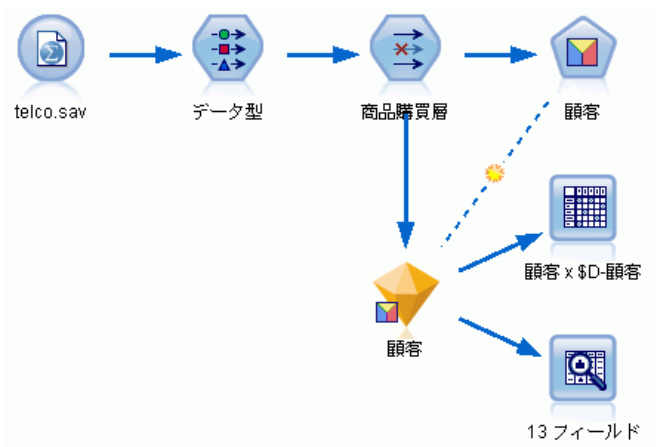
► [出力にフィールド ラベルと値ラベルを表示] を選択して [OK] をクリックします。

図 21-1
ストリームのプロパティ



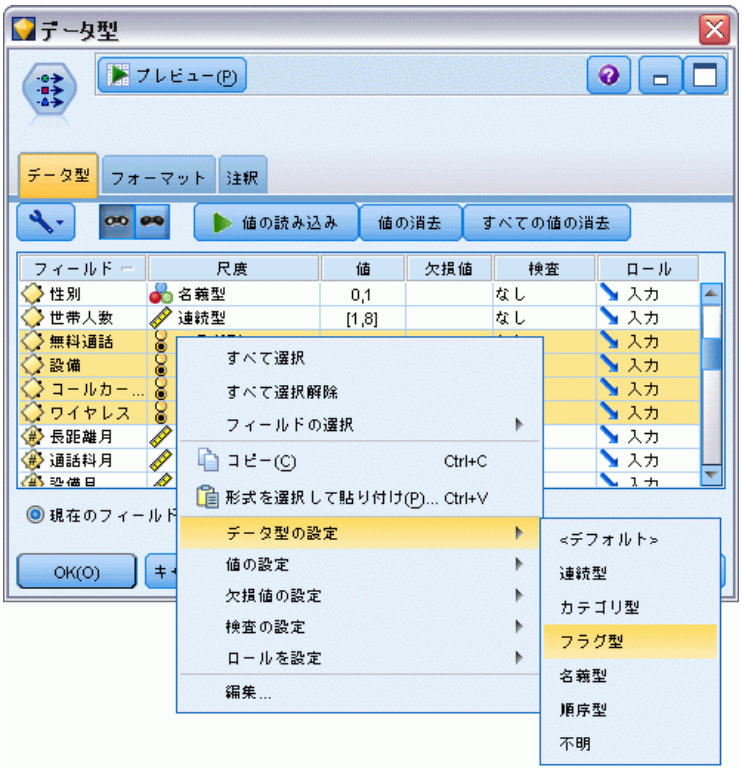
► telco.sav を指定するStatistics ファイル入力ノードを Demos フォルダに追加します。

図 21-2
分類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



- データ型ノードを追加して［**値の読み込み**］をクリックし、すべての測定レベルが正しく設定されているか確認します。たとえば、値が 0 と 1 の多くのフィールドは整数してフラグが付きます。

図 21-3
複数のフィールドの測定レベル設定



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プロパティを変更するには、［値］列のヘッダをクリックしてフィールドを値によってソートし、Shift キーを押したままマウスまたは矢印キーを使って、変更するフィールドをすべて選択します。その後、選択したフィールドの上で右クリックをすると、選択したフィールドの測定レベルまたは属性を変更することができます。

性別はフラグではなく値が 2 つのフィールドとしてより正確に認識されるため、尺度値をは［**名義型**］のままになります。

► 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 21-4
フィールドの役割の設定

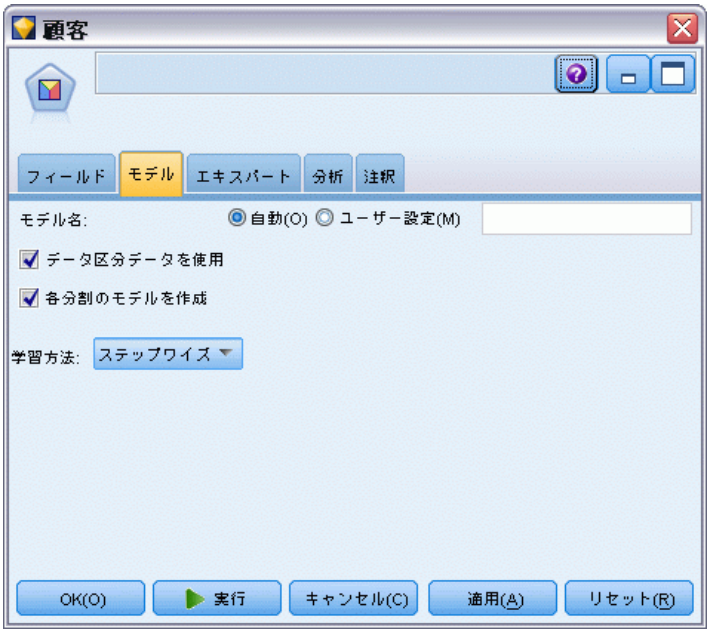
例では人口統計に焦点を当てているため、フィルター ノードを使用して関連フィールドのみを選択します（地域、年齢、 結婚、住所、収入、学歴、職業，退職、性別、居住および custcat）。他のフィールドは、この分析のために除外されることがあります。

図 21-5
人口統計フィールドのフィルタリング

（代わりに、これらのフィールドの役割を除外するのではなく**なし**に変更するか、あるいはモデル作成ノードで使用したいフィールドを選択できます。）

▶ 分類ノードで、［モデル］タブをクリックし、**［ステップワイズ法］**の方法を選択します。

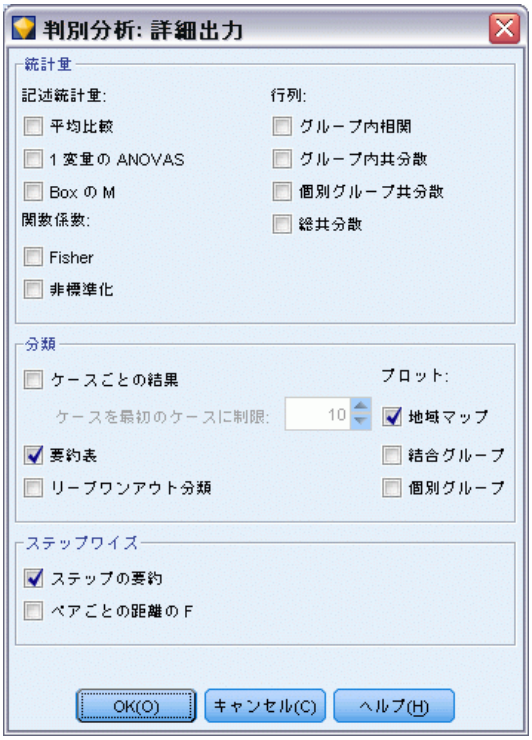
図 21-6
モデル オプションの選択



▶ ［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。

▶ ［詳細出力］ダイアログボックスで［**要約表**］、［**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 21-7
出力オプションの選択

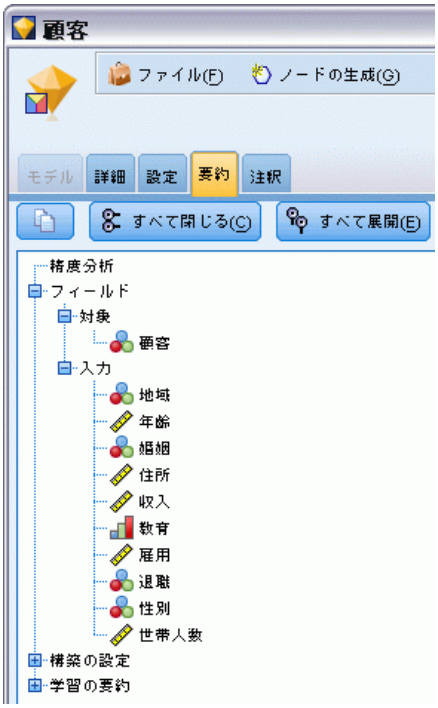


モデルの検証

▶ ［**実行**］をクリックしてモデルを生成します。生成されたモデル ナゲットはストリーム、そして右上の［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、考察のために提出された対象および入力（予測値フィールド）の完全なリストが（他の項目とともに）表示されます。

図 21-8
対象および入力フィールドを表示するモデルの要約



判別分析結果の詳細を参照するには、次のようにします。

- ［詳細設定］タブをクリックします。
- ［モデル］タブ下の「外部ブラウザで起動」をクリックして、Web ブラウザで結果を表示します。

ステップワイズ判別分析

図 21-9
分析に含まれない変数、ステップ 0

ステップ		許容度	最低許容度	入力するための F 値	Wilks のラムダ
0	年齢	1.000	1.000	7.521	.978
	婚姻状況	1.000	1.000	3.500	.990
	現在の住所	1.000	1.000	8.433	.975
	世帯の収入(千単位)	1.000	1.000	6.689	.980
	教育レベル	1.000	1.000	61.454	.844
	現職の在籍年数	1.000	1.000	16.976	.951
	退職	1.000	1.000	3.005	.991
	性別	1.000	1.000	.373	.999
	世帯の同居人数	1.000	1.000	3.976	.988

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場合に役立ちます。 ステップワイズ法は、予測値を何も含んでいないモデルから始めます。 ステップごとに、入力基準（デフォルトは 3.84）を上回る最大の F to Enter の値を持つ予測値がモデルに追加されます。

図 21-10
分析に含まれない変数、ステップ 3

ステップ		許容度	最低許容度	入力するための F 値	Wilks のラムダ
3	年齢	.535	.535	.252	.795
	婚姻状況	.605	.593	1.507	.792
	現在の住所	.776	.771	3.514	.787
	世帯の収入(千単位)	.688	.657	.687	.794
	退職	.917	.880	.353	.795
	性別	.997	.931	.395	.795

最後のステップの分析に含まれない変数はすべて 3.84 より小さな F to Enter の値を持っているため、何も追加されません。

図 21-11
分析に含まれる変数

ステップ		許容度	削除するための F 値	Wilks のラムダ
1	教育レベル	1.000	61.454	
2	教育レベル	.953	59.108	.951
	現職の在籍年数	.953	14.933	.844
3	教育レベル	.951	60.046	.940
	現職の在籍年数	.934	15.824	.834
	世帯の同居人数	.979	4.841	.807

このテーブルは、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、式に含まれる他の独立変数で説明されない変数分散の比率です。非常に低い許容度の変数はモデルに対する情報にほとんど寄与せず、計算上の問題を引き起こすことがあります。

F to Remove の値は、ある変数が現在のモデルから削除される（他の変数は削除されない）場合に何が起るかを説明するのに便利です。変数を入力するための F to Remove は、前のステップの F to Enter（「分析に含まれない変数」のテーブルに表示）と同じです。

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上の利点のみに基づいてモデルを選択するため、**実際的意義**を持たない予測値を選択する場合があります。データを取り扱ったことがあるのならば、また予測値の重要性に期待しているのならば、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べきです。しかし、多数の予測値があってどれから手をつけたらよいかわからない場合は、ステップワイズ分析を実行し、選択したモデルを修正する方が何もしないよりはましです。

モデルの適合性をチェック

図 21-12
固有値

関数	固有値	分散の %	累積 %	正準相関
1	.198(a)	80.2	80.2	.407
2	.048(a)	19.4	99.6	.214
3	.001(a)	.4	100.0	.031

モデルごとに説明される分散のほとんどすべては、最初の 2 種類の判別関数によるものです。 3 種類の関数が自動的に適合しますが、非常に小さな固有値ですから、3 番目の関数は問題なく無視できます。

図 21-13
Wilks のラムダ

関数の検定	Wilks のﾗﾑﾀﾞ	ｶｲ2乗	自由度	有意確率
1 から 3 まで	.796	227.345	9	.000
2 から 3 まで	.953	47.486	4	.000
3	.999	.929	1	.335

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。関数のそれぞれのセットについて、これは表示される関数の手段はグループを通じて等しいという帰無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

図 21-14
構造行列

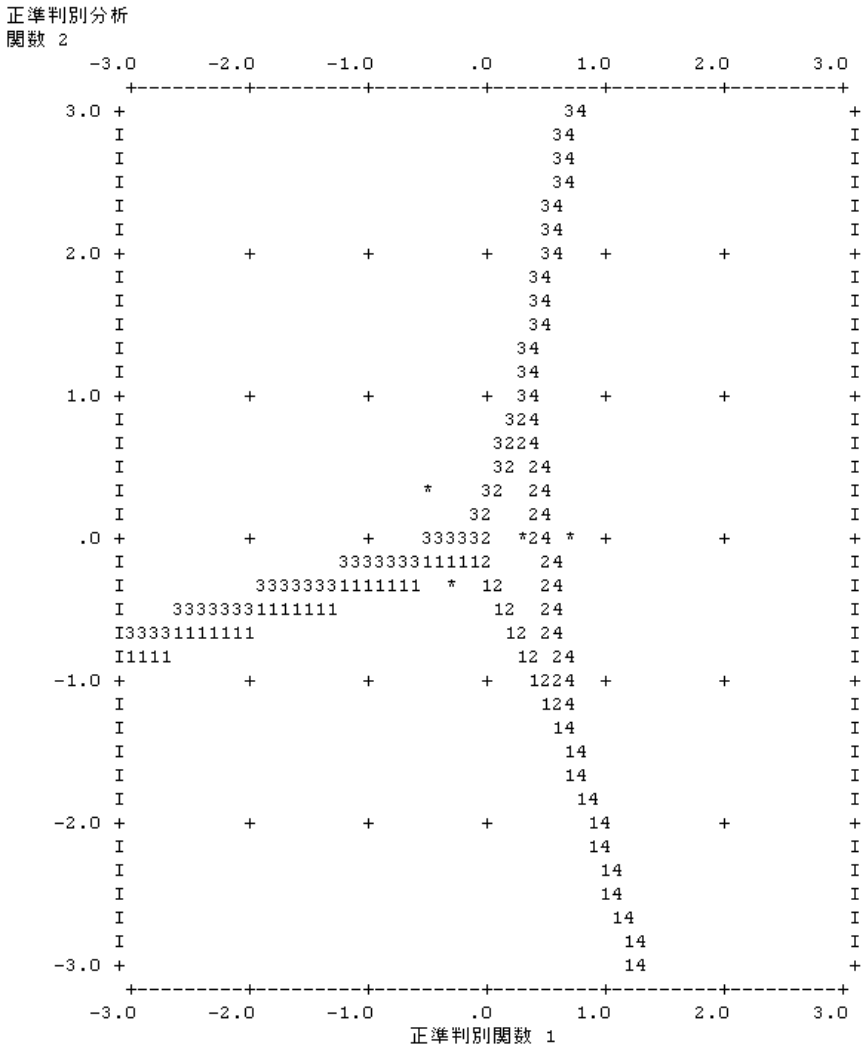
	関数		
	1	2	3
教育レベル	.966(*)	-.090	-.244
現職の在籍年数	-.182	.964(*)	-.193
年齢(a)	-.162	.598(*)	-.285
世帯の収入(千単位)(a)	.109	.514(*)	-.190
現在の住所(a)	-.151	.394(*)	-.214
退職(a)	-.108	.230(*)	-.137
性別(a)	.008	.054(*)	.009
世帯の同居人数	.232	.097	.968(*)
婚姻状況(a)	.132	.134	.600(*)
判別変数と標準化された正準判別関数間のプールされたグループ内相関変数は関数内の相関の絶対サイズにしたがって並べ替えられます。			
*. 各変数と任意の判別関数間の最大絶対相関			
a. この変数は分析に使用されません。			

複数の判別関数がある場合、標準関数の 1 つによりそれぞれの変数の最大絶対相関関係にアスタリスク (*) が付きます。各関数において、印を付けられた変数は相関関係に規模に応じて並べ替えられます。

- 教育水準は第 1 の関数と緊密な相関関係にあり、これはこの関数と最も緊密な相関関係にある変数だけです。
- 性別と引退は他の関数との相関関係はあまり強くはありませんが、勤続年数、年齢、世帯あたりの収入（単位：千)、居住年数、引退、および性別は第 2 の関数と密接な相関関係にあります。その他の変数はこの関数を「安定性」関数として印を付けます。
- 家族の人数と既婚/未婚は、第 3 の判別関数と最も密接な相関関係にありますが、この関数は役に立たないため、これらはほとんど実用にならない予測値です。

地域マップ

図 21-15
領域マップ(T)



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の図示説明が表示されます。 横軸に示されている第 1 の関数は、グループ 4（総合サービス顧客）を他のグループから選り分けます。教育水準は第 1 の関数と強力で明確な相関関係にあるため、これは総合サービス顧客が一般的には最高水準の教育を受けているということを示しています。 第 2 の関数は、グループ 1 と 3（基本サービスと付加サービスの顧客）を選り分けます。付加サービス顧客は基本サービス顧客よりも長い期間働き続けて年長であるという傾向があります。 E-サービス顧客は十分な教育を受けていて適度

な職歴を有しているという傾向があることをマップが示していますが、彼らはその他の顧客からは完全に分離されません。

一般的には、アスタリスク (*) が付けられたグループ重心が領域線に接近すると、すべてのグループの分類がそれほど明確ではないということになります。

最初の 2 種類の判別関数だけ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

図 21-16
分類結果

			予測ｸﾞﾙｰﾌﾟ番号				合計
		顧客カテゴリ	ベーシック サービス	E-サービス	プラス サービス	トータル サービス	
元のﾃﾞｰﾀ	度数	ベーシック サービス	125	11	61	69	266
		E-サービス	49	15	58	95	217
		プラス サービス	102	14	112	53	281
		トータル サービス	40	16	37	143	236
	%	ベーシック サービス	47.0	4.1	22.9	25.9	100.0
		E-サービス	22.6	6.9	26.7	43.8	100.0
		プラス サービス	36.3	5.0	39.9	18.9	100.0
		トータル サービス	16.9	6.8	15.7	60.6	100.0

a. 元のｸﾞﾙｰﾌﾟ化されたｹｰｽのうち 39.5% 個が正しく分類されました。

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。 観測データを前提とすると、「ヌル」モデル（すなわち、予測値のないモデル）はすべての顧客を形式上のグループである付加サービスに分類します。したがって、ヌル モデルは時間の 281/1000 = 28.1% が適切です。 このモデルは、顧客の 11.4% 以上または 39.5% を得ます。特に、このモデルは総合サービス顧客を特定することに卓越しています。ただし、E-サービス顧客を分類することについては不得手です。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かもしれません。

要約

それぞれの顧客からの人口統計情報に基づき、4 種類の定義済みの「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、顧客ベースを分割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、このモデルは E-サービス顧客を分類するのが不得手であるということを分類結果は示しています。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ものしだいでは、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。たとえば、E-サービス顧客を特定することに携わっていなくても、モデルの精度は十分であるかもしれません。これは、E-サービスがほとんど利益を生み出さない目玉商品である場合のことです。たとえば、最高の投資収益が付加サービスまたは総合サービスの顧客によってもたらされる場合、モデルは必要な情報を提供しているのかもしれません。

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。モデルが他のデータをどれだけうまく一般化しているかを評価するには、データ区分ノードを使用して、テストおよび検証の目的でレコードのサブセットを保持します。詳細は、 4 章 データ区分ノード in IBM SPSS Modeler 15 入力ノード、プロセス ノード、出力ノード を参照してください。

IBM® SPSS® Modelerで使用されているモデリング方法の数学的な基礎の説明は、『SPSS Modeler Algorithms Guide』 に一覧されています。これは、インストール ディスクの Documentation ディレクトリにあります。

章
22

区間打ち切り生存データの分析（一般化線型モデル）

区間打ち切りによる生存データを分析する場合（対象となるイベントの正確な時刻は不明であるが一定の期間内に発生したことだけがわかっている場合）、その期間内に発生したイベントの危険性に Cox モデルを適用すると、補数対数-対数回帰モデルが得られます。

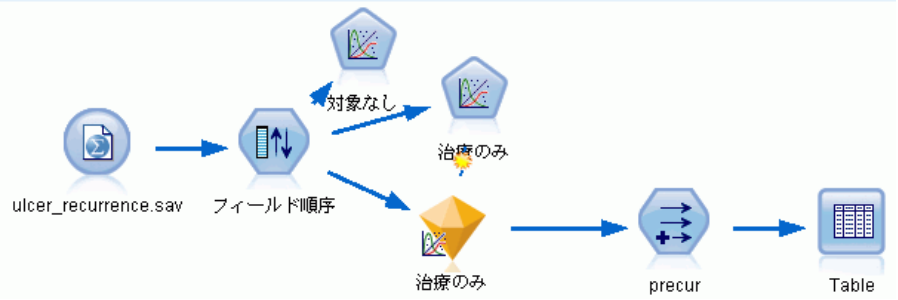
潰瘍の再発防止のための 2 種類の療法の有効性を比較することを目的とした研究による部分情報は、ulcer_recurrence.sav に保存されます。このデータセットは、他の場所で表示および分析されています。一般化線型モデルを使用すると、補数対数-対数回帰モデルの結果を複製できます。

この例では、データ ファイル ulcer_recurrence.sav を参照する ulcer_genlin.str というストリームを使用します。データ ファイルは Demos フォルダにあり、ストリーム ファイルは streams サブフォルダにあります。 詳細は、 1 章 Demos フォルダ in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。

ストリームの作成

- ulcer_recurrence.sav を指定する Statistics ファイル入力ノードを Demos フォルダに追加します。

図 22-1
潰瘍の再発を予測するサンプル ストリーム



▶ 入力ノードの［フィルタ］タブで、id と時間を除外します。

図 22-2
不要なフィールドの除外

▶ 入力ノードの［データ型］タブで、［結果］フィールドの役割を［**対象**］に設定し、その測定レベルを［**フラグ型**］に設定します。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。

► [値の読み込み] をクリックしてデータをインスタンス化します。

図 22-3
フィールドの役割の設定

▶ フィールド順序ノードを追加し、入力の順序として持続時間、治療、および年齢を指定します。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 の結果を複製する手助けになります。

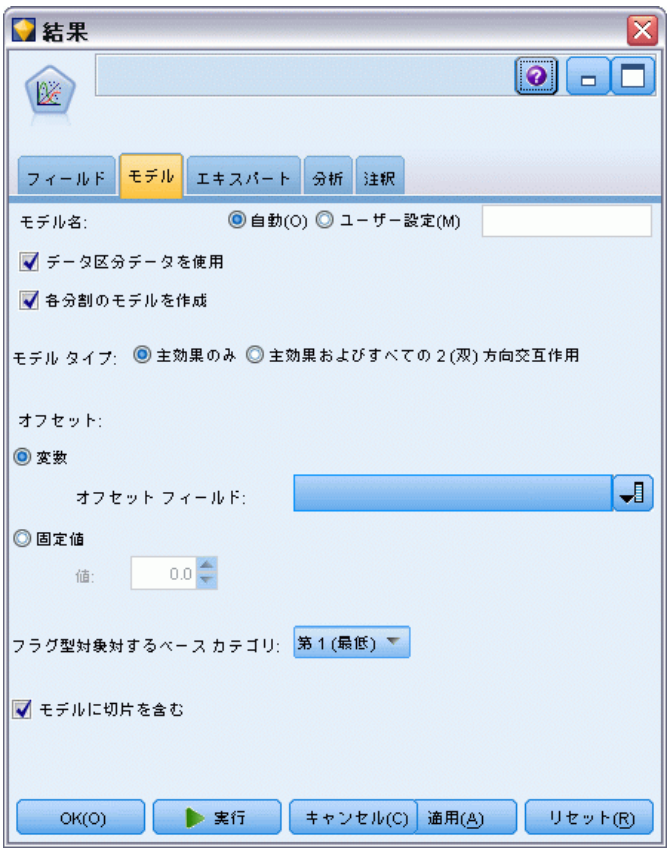
図 22-4
希望どおりにフィールドをモデルに入力するためのフィールドの並べ替え

▶ GenLin ノードを入力ノードに接続します。GenLin ノードで、[**モデル**] タブをクリックします。

▶ 対象の参照カテゴリとして [**第 1（最低）**] を選択します。これは、第 2 のカテゴリが対象となるイベントであり、モデルに対する効果はパラメータ推定の解釈によるということを示します。正の係数を伴う四族型予測値は、予測値の増大により再発の確率が増すことを意味します。より大きな係数

を伴うセット予測値のカテゴリは、名義型の他のカテゴリに関して再発の確率が増すことを意味します。

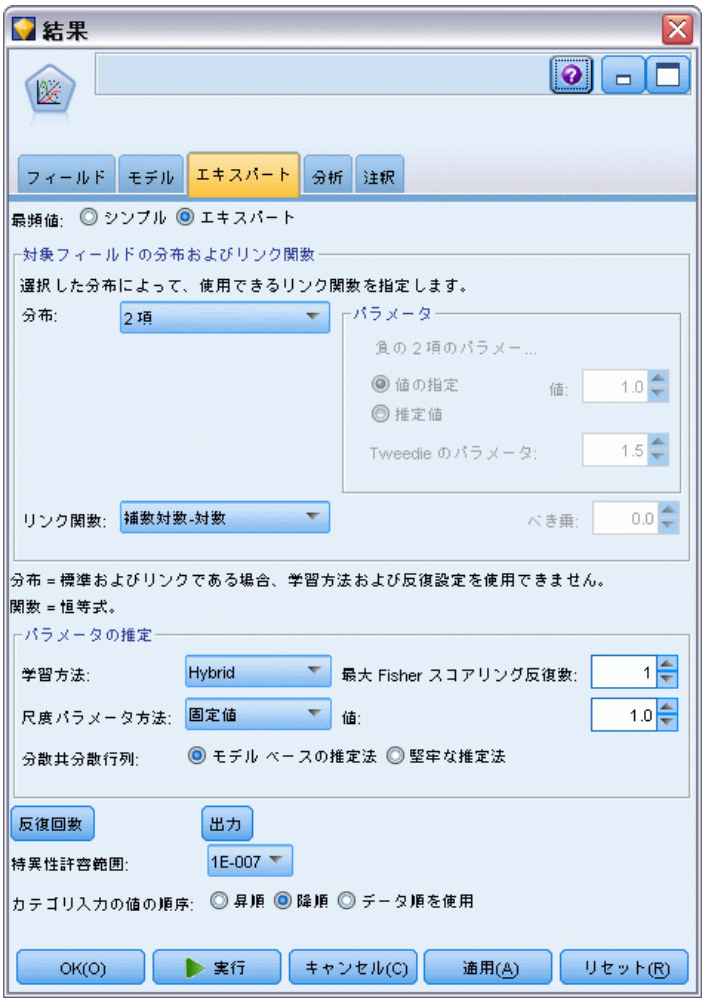
図 22-5
モデル オプションの選択



- [**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。
- 分布として [**二項**] を選択し、リンク関数として [**補数対数-対数**] を選択します。
- 尺度パラメータを推定する方法として [**固定値**] を選択し、1.0 のデフォルト値はそのままにしておきます。

▶ 因子のカテゴリ順序として［降順］を選択します。これは、それぞれの因子の第 1 のカテゴリは参照カテゴリであるということを示します。モデルに関するこの選択の影響は、パラメータ推定値の解釈に反映されます。

図 22-6
エキスパート オプションの選択



▶ ストリームを実行してモデル ナゲットを作成します。ストリーム領域および右上の［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て、［編集］または［ブラウズ］を選択します。

モデル効果の検定

図 22-7
主効果モデルのモデル効果検定

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	.536	1	.464
継期続間	.003	1	.958
治療	.382	1	.537
年齢	.358	1	.550
Dependent Variable: 結果Model: (Intercept), 継期続間, 治療, 年齢			

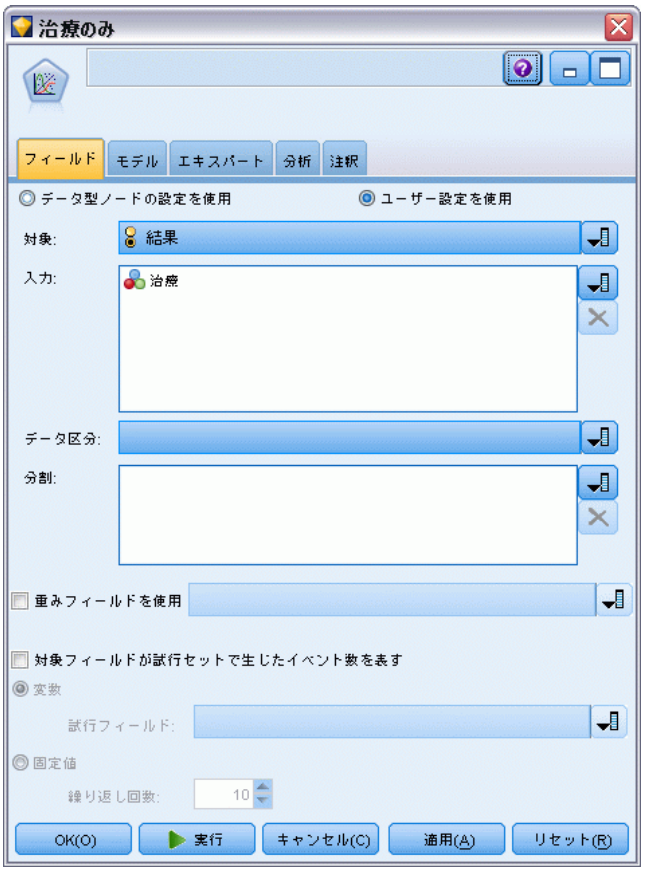
統計的に有意なモデル効果はありません。ただし、治療効果における目立った相違は臨床的関心によるものであるため、モデルの項としての治療のみにより縮小モデルを適合させます。

治療のみのモデルの適合

- GenLin ノードの［フィールド］タブで、**[ユーザー設定を使用]** をクリックします。
- 対象として結果を選択します。

▶ 単独入力として治療を選択します。

図 22-8
フィールド オプションの選択



▶ 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、[**アドバンス**] タブを選択し、一番下までスクロールします。

パラメータ推定値

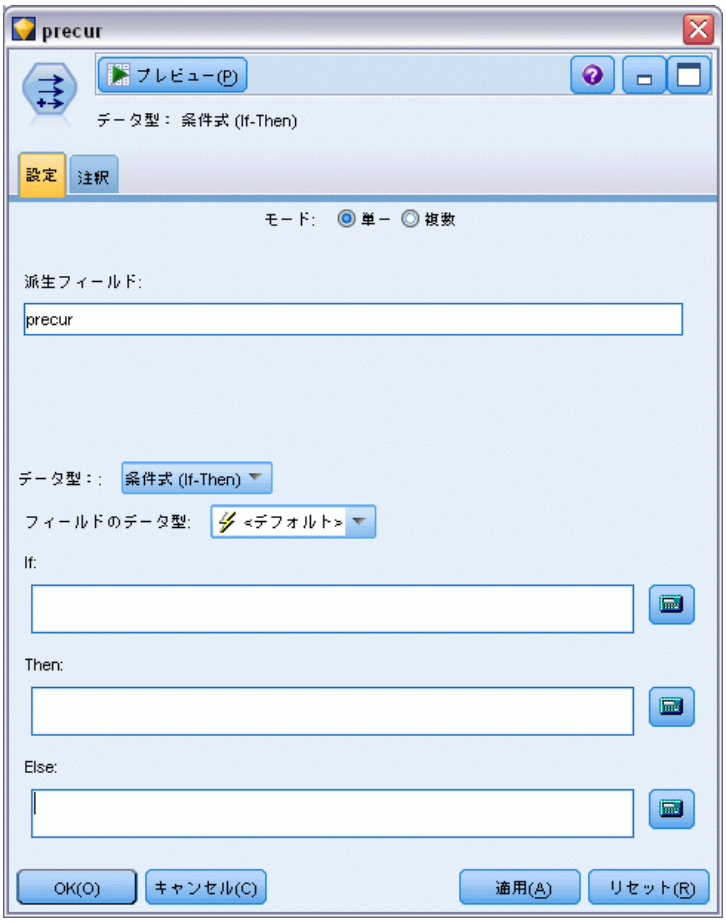
図 22-9
治療のみのモデルに関するパラメータ推定

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
[治療=1]	.378	.6288	-.855	1.610	.361	1	.548
[治療=0]	0(a)	.	.	.	.	.	.
(Scale)	1(b)						
Dependent Variable: 結果Model: (Intercept), 治療							
a. Set to zero because this parameter is redundant.							
b. Fixed at the displayed value.							

治療効果（2 種類の治療レベルの線型予測量、すなわち［treatment=1］の係数）はやはり統計的に有意ではなく、治療 B のパラメータ推定が A よりも大きいために、最初の 12 ヶ月で増大した再発確率に関連付けられるので、治療 A［treatment=0］は B［treatment=1］よりも優れているかもしれないと示唆しているに過ぎません。線型予測量（切片 + 治療効果）は対数（−対数(1−P($recur_{12, t}$)）の推定であり、この場合、P($recur_{12, t}$）は治療 t(=A または B) に関する 12 ヶ月間の再発の確率です。これらの予測確率は、データベースのそれぞれの観測について生成されます。

再発と生存の予測確率

図 22-10
フィールド作成ノードの設定オプション



- 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。
- ［設定］タブで、作成するフィールドとして precur を“ü力します。
- **［条件式］**として作成することを選択します。
- 計算機ボタンをクリックして、If 条件式の式ビルダーを開きます。

図 22-11
フィールド作成ノード :If 条件式の式ビルダー

▶ $G-result フィールドを式に挿入します。

▶ [OK] をクリックします。

作成するフィールド precur は、$G-result が 1 である場合は Then 式の値を取り、0 である場合は Else 式の値を取ります。

図 22-12
フィールド作成ノード :Then 式の式ビルダー

- 計算機ボタンをクリックして、Then 式の式ビルダーを開きます。
- $GP-result フィールドを式に挿入します。
- [OK] をクリックします。

図 22-13
フィールド作成ノード :Else 式の式ビルダー

- 計算機ボタンをクリックして、Else 式の式ビルダーを開きます。
- 式に 1- と入力し、$GP-result フィールドを挿入します。
- [OK] をクリックします。

図 22-14
フィールド作成ノードの設定オプション

▶ 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。

図 22-15
予測された確率(R)

	結果	維期続間	治療	年齢	$G-結果	$GP-結果	precur
1	1	2	1	48	0	0.708	0.292
2	0	1	1	73	0	0.708	0.292
3	0	1	1	54	0	0.708	0.292
4	0	2	1	58	0	0.708	0.292
5	0	1	0	56	0	0.789	0.211
6	0	2	0	49	0	0.789	0.211
7	0	1	1	71	0	0.708	0.292
8	0	1	0	41	0	0.789	0.211
9	0	1	1	23	0	0.708	0.292
10	1	1	1	37	0	0.708	0.292
11	0	1	1	38	0	0.708	0.292
12	0	2	1	76	0	0.708	0.292
13	0	2	0	38	0	0.789	0.211
14	1	1	0	27	0	0.789	0.211
15	1	1	1	47	0	0.708	0.292
16	0	1	0	54	0	0.789	0.211
17	1	1	1	38	0	0.708	0.292
18	1	2	1	27	0	0.708	0.292
19	0	2	0	58	0	0.789	0.211

治療 A に割り当てられた患者が最初の 12 ヶ月に再発を経験するという推定 0.211 の確率があり、治療 B の場合は 0.292 です。1-P(recur$_{12,\ t}$) は 12 ヶ月間の生存者の確率であり、これは生存確率アナリストにとって興味深いことかもしれません。

期間による再発確率のモデル作成

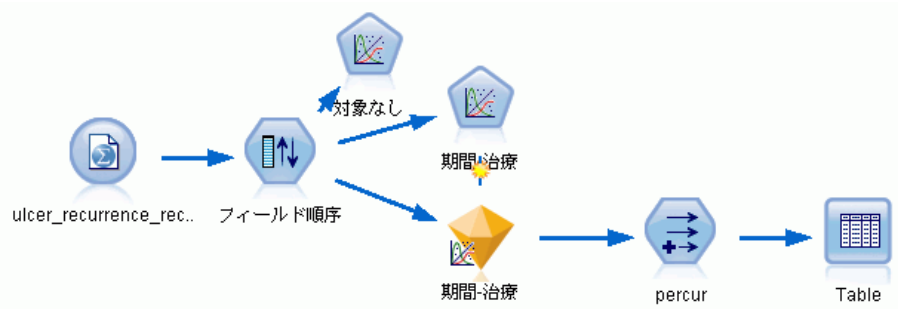
現状でのモデルに対する問題は、最初の調査時に収集した情報が無視されることです。すなわち、多くの患者は最初の 6 ヶ月間に再発を経験しませんでした。「より優れた」モデルは、各期間中にイベントが発生したかどうかを記録する 2 値レスポンスをモデル作成します。このモデルを適合させるには、元のデータセットを再構築する必要があります。このデータセットは ulcer_recurrence_recoded.sav にあります。詳細は、 1 章 Demos フォルダ in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。このファイルには、2 つの追加変数が記載されています。

- 期間は、ケースが最初の調査期間に対応するのかそれとも 2 番目の調査期間に対応するのかを記録します。
- 期間別の結果は、ある特定の期間中にある特定の患者に再発が見られた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期間ごとに 1 つのケースに寄与します。したがって、たとえば、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第 2 の調査期間に関するものです。一方、最初の期間において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 ヶ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寄与します。

▶ ulcer_recurrence_recoded.sav を指定する Statistics ファイル入力ノードを Demos フォルダに追加します。

図 22-16
潰瘍の再発を予測するサンプル ストリーム



▶ 入力ノードの［フィルタ］タブで、id、時間、および結果を除外します。

図 22-17
不要なフィールドの除外

▶ 入力ノードの［データ型］タブで、result2 フィールドの役割を［**対象**］に設定し、その測定レベルを［**フラグ型**］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。

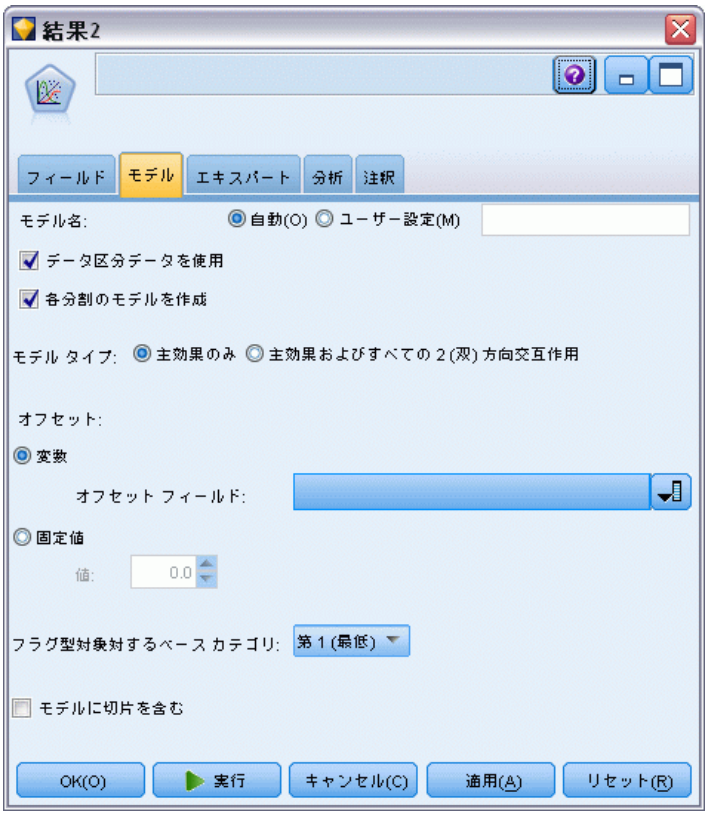
図 22-18
フィールドの役割の設定

▶ フィールド順序ノードを追加し、入力の順序として期間、持続時間、治療、および年齢を指定します。期間を最初の入力にすることにより（切片の項をモデルに含めずに）、期間効果を取り込むように変数一式を適合できます。

図 22-19
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

▶ GenLin ノードで、[モデル] タブをクリックします。

図 22-20
モデル オプションの選択

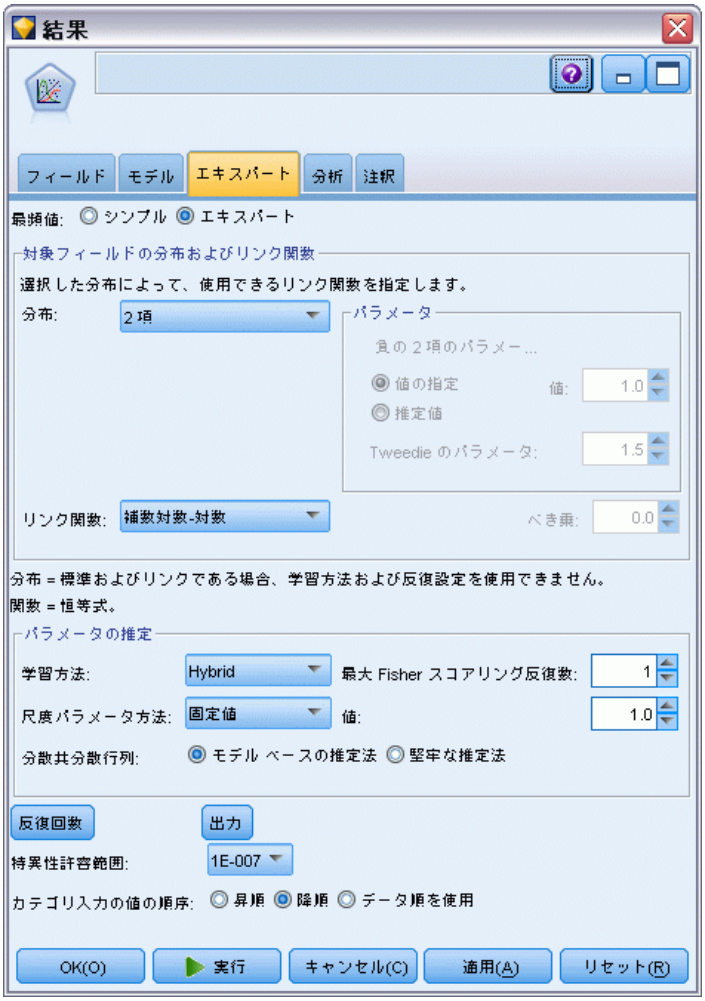


▶ 対象の参照カテゴリとして [第 1 (最低)] を選択します。これは、第 2 のカテゴリが対象となるイベントであり、モデルに対する効果はパラメータ推定の解釈によるということを示します。

▶ 「モデルに切片を含む」の選択を解除します。

▶ [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 22-21
エキスパート オプションの選択



▶ 分布として [二項] を選択し、リンク関数として [補数対数-対数] を選択します。

▶ 尺度パラメータを推定する方法として [固定値] を選択し、1.0 のデフォルト値はそのままにしておきます。

▶ 因子のカテゴリ順序として [降順] を選択します。これは、それぞれの因子の第 1 のカテゴリは参照カテゴリであるということを示します。モデルに関するこの選択の影響は、パラメータ推定値の解釈に反映されます。

▶ ストリームを実行してモデル ナゲットを作成します。ストリーム領域および右上の［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て、[編集] または [ブラウズ] を選択します。

モデル効果の検定

図 22-22
主効果モデルのモデル効果検定

Source	Type III		
	Wald Chi-Square	df	Sig.
期間	.464	1	.496
継続期間	.000	1	.988
治療	.117	1	.732
年齢	.314	1	.575
Dependent Variable: 期間の結果Model: 期間, 継続期間, 治療, 年齢			

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関心によるものであるため、これらのモデルの項のみにより縮小モデルを適合させます。

縮小モデルの適合

▶ GenLin ノードの［フィールド］タブで、[ユーザー設定を使用] をクリックします。

▶ 対象としてresult2 を選択します。

▶ 入力として期間および治療を選択します。

図 22-23
フィールド オプションの選択

▶ ノードを実行して生成されたモデルをブラウズし、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。

パラメータ推定値

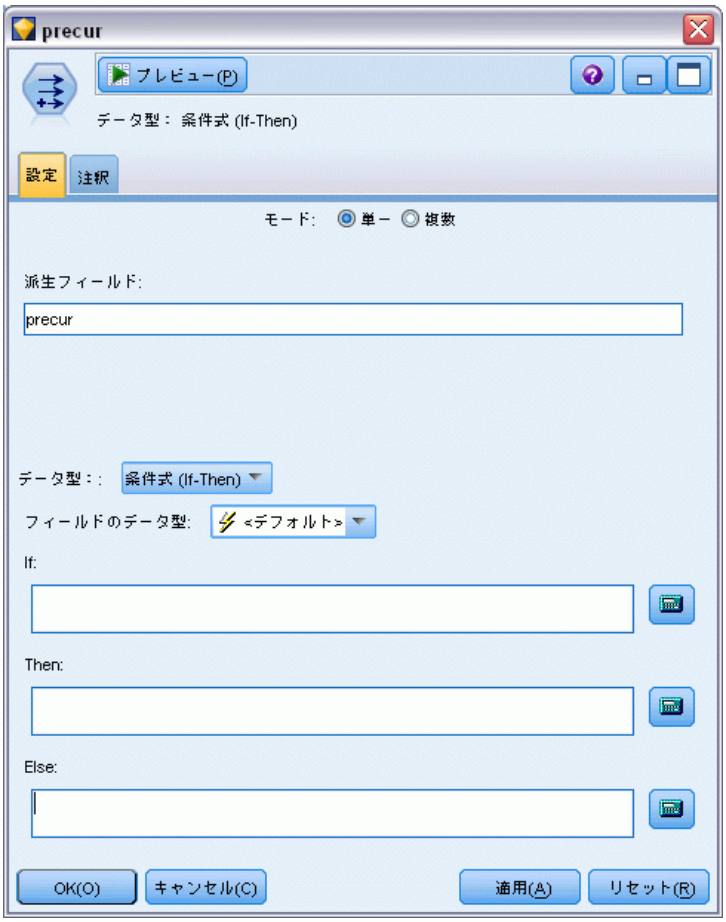
図 22-24
治療のみのモデルに関するパラメータ推定

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
[期間=2]	-1.794	.5792	-2.929	-.659	9.597	1	.002
[期間=1]	-2.206	.5912	-3.365	-1.047	13.926	1	.000
[治療=1]	.195	.6279	-1.035	1.426	.097	1	.756
[治療=0]	0(a)	.	.	.	.	.	.
(Scale)	1(b)						
Dependent Variable: 期間の結果Model: 期間, 治療							
a. Set to zero because this parameter is redundant.							
b. Fixed at the displayed value.							

治療効果はやはり統計的に有意ではなく、治療 B のパラメータ推定が最初の 12 ヶ月で増大した再発確率に関連付けられるため、治療 A は B よりも優れているかもしれないと示唆しているに過ぎません。期間の値は 0 から統計的に著しくかけ離れていますが、これは切片の項が適合していないことによるものです。モデル効果のテストにおいてみられるように、統計効果（[period=1] と [period=2] の線型予測量の値の差）は統計的に有意ではありません。 線型予測量（期間効果 + 治療効果）は対数（−対数(1−P($recur_{p,\ t}$)）の推定であり、この場合、P($recur_{p,\ t}$）は治療 t(=A または B) を仮定した場合の期間 p (=1 または 2、すなわち 6 ヶ月または 12 ヶ月）における再発の確率です。これらの予測確率は、データベースのそれぞれの観測について生成されます。

再発と生存の予測確率

図 22-25
フィールド作成ノードの設定オプション



- 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。
- ［設定］タブで、作成するフィールドとして precur を“ü力します。
- **［条件式］**として作成することを選択します。
- 計算機ボタンをクリックして、If 条件式の式ビルダーを開きます。

図 22-26
フィールド作成ノード :If 条件式の式ビルダー

▶ $G-result2 フィールドを式に挿入します。

▶ [OK] をクリックします。

作成するフィールド precur は、$G-result2 が 1 である場合は Then 式の値を取り、0 である場合は Else 式の値を取ります。

図 22-27
フィールド作成ノード :Then 式の式ビルダー

- 計算機ボタンをクリックして、Then 式の式ビルダーを開きます。
- $GP-result2 フィールドを式に挿入します。
- [OK] をクリックします。

図 22-28
フィールド作成ノード :Else 式の式ビルダー

- ▶ 計算機ボタンをクリックして、Else 式の式ビルダーを開きます。
- ▶ 式に 1- と入力し、$GP-result2 フィールドを挿入します。
- ▶ [OK] をクリックします。

図 22-29
フィールド作成ノードの設定オプション

▶ 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。

図 22-30
予測された確率(R)

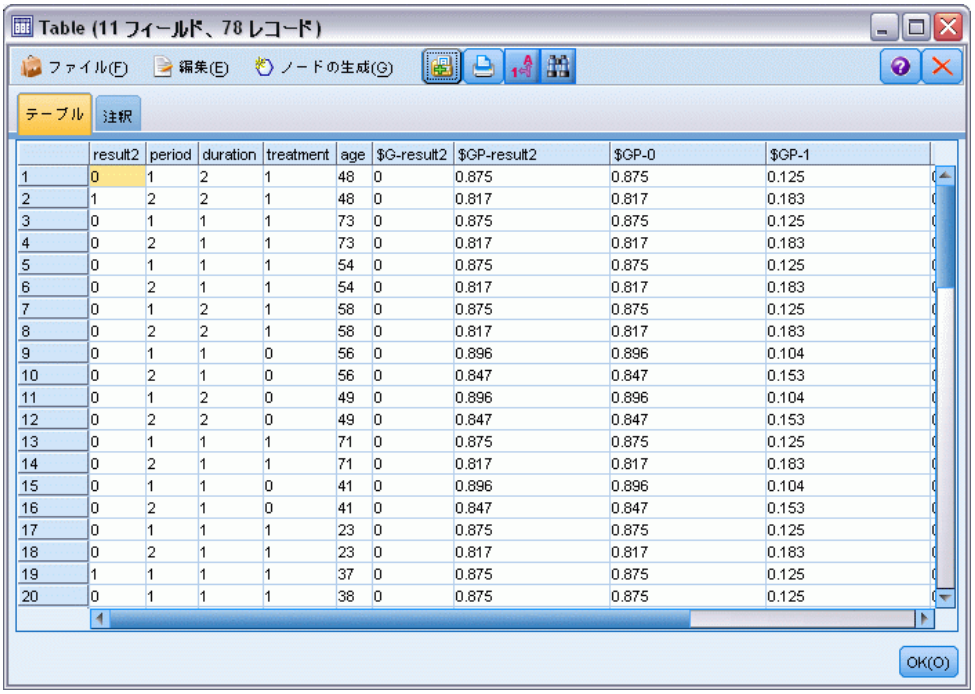


Table (11 フィールド、78 レコード)

ファイル(F) 編集(E) ノードの生成(G)

テーブル 注釈

	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

OK(O)

推定再発確率は以下のように要約できます。

治療	6 ヶ月	12 ヶ月
A	0.104	0.153
B	0.125	0.183

これらのことから、12 ヶ月の生存確率は、$1-(P(recur_{1,t}) + P(recur_{2,t}) \times (1-P(recur_{1,t})))$ と表されます。それぞれの治療については、次のようになります。

A:1 − (0.104 + 0.153*0.896) = 0.759

B:1 − (0.125 + 0.183*0.875) = 0.715

これは、よりよい治療として A に対する非統計的に有意なサポートを示しています。

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ついて一連の補数対数-対数回帰モデルを適合させました。治療 A の選択に関しては何らかのサポートがあるのに対し、統計的に有意な結果を得るには大規模な研究が必要です。しかし、既存のデータを調査するさらなる手段がいくつかあります。

- 交互作用効果でモデルを再適合させることは、特に期間と治療グループの間では価値のあること,©もしれません。

IBM® SPSS® Modeler で使用されているモデル作成方法の数学的な基礎の説明は、『SPSS Modeler Algorithms Guide』 に一覧されています。

章

23

船舶損傷率の分析のためのポアソン回帰の使用（一般化線型モデル）

一般化線型モデルは、カウントデータの分析についてポアソン回帰を適合させるのに使用できます。たとえば、他の場所で表示および分析されるデータセットは波によってこうむる貨物船の損害に関するものです。予測値が与えられると、ポア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。

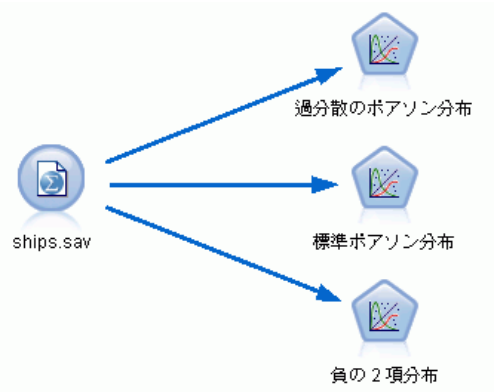
この例では、データ ファイル ships.sav を参照するストリーム ships_genlin.str を使用します。データ ファイルは Demos フォルダにあり、ストリーム ファイルは streams サブフォルダにあります。詳細は、 1 章 Demos フォルダ in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。

集計契約月数が船舶タイプにより変動するため、生のセルのカウントのモデル作成はこの状況では誤解を招くことがあります。リスクの「負担」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、ポアソン回帰は従属変数の対数が予測では線型であると想定します。したがって、一般化線型モデルを使用してポアソン回帰を事故率に適合させるには、集計契約月数の対数を使用する必要があります。

「過分散」ポアソン回帰の適合

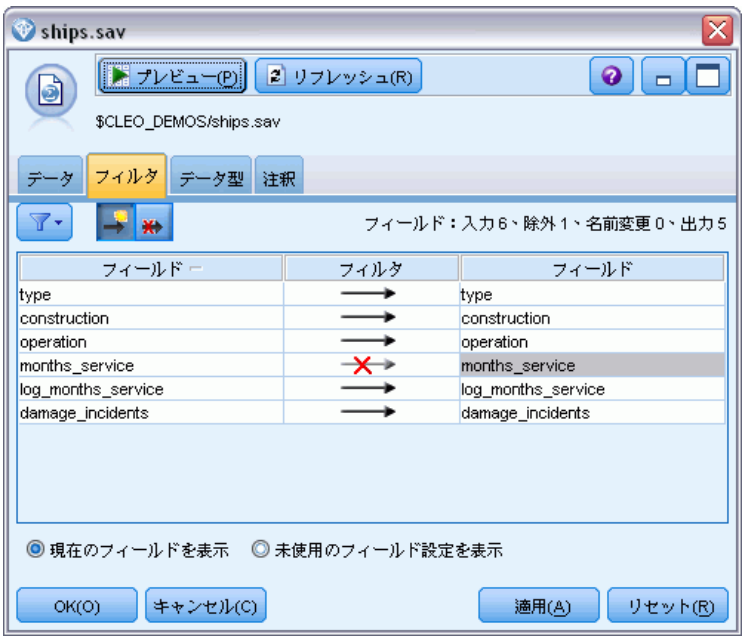
▶ ships.sav を指定するStatistics ファイル入力ノードを Demos フォルダに追加します。

図 23-1
損傷率を分析するためのサンプル ストリーム



▶ 入力ノードの［フィルタ］タブで、フィールド months_service を除外します。この変数の対数変換値は log_months_service に含まれており、分析で使用します。

図 23-2
不要なフィールドのフィルタリング



（代わりに、［タイプ］タブでこのフィールドの役割を除外するのではなく**なし**に変更するか、あるいはモデル作成ノードで使用したいフィールドを選択できます。）

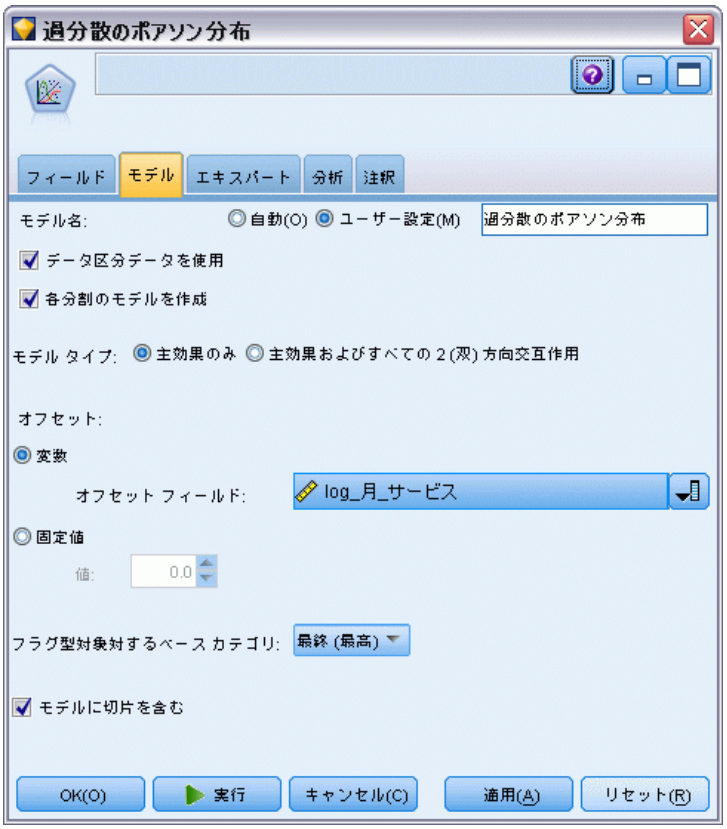
- 入力ノードの［タイプ］タブで、damage_incidents フィールドの役割を**対象**に設定します。その他のフィールドの役割はすべて**入力**に設定します。
- **[値の読み込み]** をクリックしてデータをインスタンス化します。

図 23-3
フィールドの役割の設定

- Genlin ノードを入力ノードに接続します。Genlin ノードで、**[モデル]** タブをクリックします。

▶ オフセット変数として log_months_service を選択します。

図 23-4
モデル オプションの選択



▶ [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 23-5
エキスパート オプションの選択

▶ レスポンスの分布として [ポアソン] を選択し、リンク関数として [対数] を選択します。

▶ 尺度パラメータの推定方法として [Pearson のカイ 2 乗] を選択します。ポアソン回帰では尺度パラメータを通常は 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と有意水準を得ます。

▶ 因子のカテゴリ順序として [降順] を選択します。これは、それぞれの因子の第 1 のカテゴリは参照カテゴリであるということを示します。モデルに関するこの選択の影響は、パラメータ推定値の解釈に反映されます。

▶ [実行] をクリックしてモデル ナゲットを生成します。生成されたモデル ナゲットはストリーム領域、そして右上の［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て、[編集] または [ブラウズ] を選択し、[アドバンス] タブをクリックします。

適合度統計

図 23-6
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(b,c)	-68.281		
Adjusted Log Likelihood(d)	-40.379		
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: 損傷件数Model: (Intercept), 種類, 建造, 作業, offset = log_月_サービス			
a. Information criteria are in small-is-better form.			
b. The full log likelihood function is displayed and used in computing information criteria.			
c. The log likelihood is based on a scale parameter fixed at 1.			
d. The adjusted log likelihood is based on an estimated scale parameter and is used in the model fitting omnibus test.			

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測定値を示しています。さらに、逸脱とPearson のカイ 2 乗統計の値/自由度は、尺度パラメータの対応する推定を示します。これらの値は、ポア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

図 23-7
オムニバス検定

Likelihood Ratio Chi-Square	df	Sig.
107.633	8	.000
Dependent Variable: 損傷件数Model: (Intercept), 種類, 建造, 作業, offset = log_月_サービス		
a. Compares the fitted model against the intercept-only model.		

オムニバス検定は、現行モデル対ヌルモデル（この場合は切片）の尤度比カイ 2 乗検定です。0.05 未満の有意確率は、現在のモデルがヌル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

図 23-8
モデル効果の検定

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	3616.491	1	.000
種類	26.067	4	.000
建造	29.157	3	.000
作業	10.567	1	.001
Dependent Variable: 損傷件数Model: (Intercept), 種類, 建造, 作業, offset = log_月_サービス			

モデルのそれぞれの項が効果の有無をテストされます。有意確率が 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。

パラメータ推定値

図 23-9
パラメータ推定値(M)

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-6.406	.2174	-6.832	-5.980	867.891	1	.000
[種類=5]	.326	.2359	-.137	.788	1.905	1	.168
[種類=4]	-.076	.2906	-.645	.494	.068	1	.794
[種類=3]	-.687	.3290	-1.332	-.042	4.364	1	.037
[種類=2]	-.543	.1776	-.891	-.195	9.361	1	.002
[種類=1]	0(a)	.	.	.	.	.	.
[建造=75]	.453	.2332	-.004	.910	3.782	1	.052
[建造=70]	.818	.1698	.486	1.151	23.239	1	.000
[建造=65]	.697	.1496	.404	.990	21.704	1	.000
[建造=60]	0(a)	.	.	.	.	.	.
[作業=75]	.384	.1183	.153	.616	10.567	1	.001
[作業=60]	0(a)	.	.	.	.	.	.
(Scale)	1(b)						
Dependent Variable: 損傷件数Model: (Intercept), 種類, 建造, 作業, offset = log_月_サービス							
a. Set to zero because this parameter is redundant.							
b. Fixed at the displayed value.							

パラメータ推定値テーブルは、各予測変数の効果をまとめたものです。リンク関数の特性のためにこのモデルにおける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はモデルにおける予測値の効果に対する重要な洞察を示します。

- 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。正の係数による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、より大きな係数による因子レベルは、より多くの損傷発生を示します。因子レベルの係数の符号は、参照カテゴリに対する因子レベルの効果によって異なります。

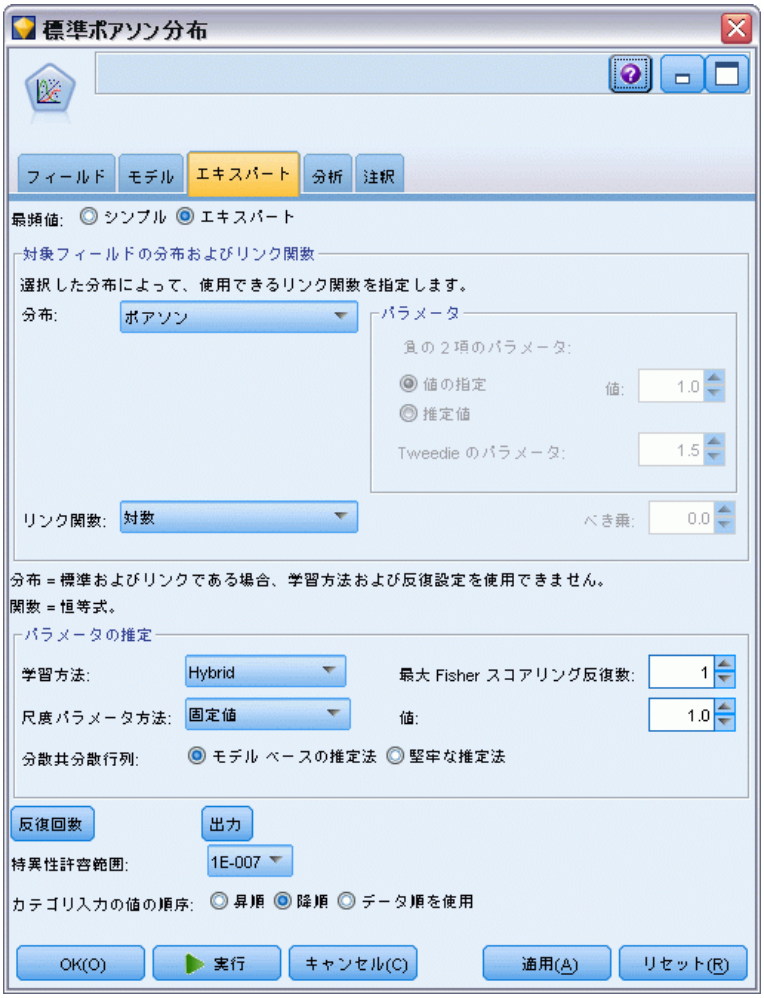
パラメータ推定値に基づいて次のように解釈できます。

- 船舶タイプ B [type=2] は、タイプ A [type=1] の参照カテゴリよりも統計的に著しく（0.019の p 値）低い損傷率（-0.543の推定係数）になります。タイプ C [type=3] は実際には B よりも低い推定パラメータを持っていますが、C の推定における変動性はこの効果をあいまいにします。因子レベル間のあらゆる関係については、推定境界平均を参照してください。
- 1965-69 年 [construction=65] および 1970-74 年 [construction=70] に建造された船舶は、1960-64 年の参照カテゴリ [construction=60] に建造された船舶よりも統計的に著しく（<0.001 の p 値）高い損傷率（それぞれ 0.697 と 0.818 の推定係数）になります。因子レベル間のあらゆる関係については、推定境界平均を参照してください。
- 1975-79 年 [operation=75] に運行されていた船舶は、1960-1974 年 [operation=60] に運行されていた船舶よりも統計的に著しく（0.012の p 値）高い損傷率（0.384の推定係数）になります。

代替モデルの適合

「過分散」ポアソン回帰による 1 つの問題は、「標準」ポアソン回帰に対してそれをテスト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 1 つの正式な提案テストは、その他のすべての同じ設定について「標準」ポアソン回帰と負の二項回帰の間で尤度比テストを実行するためのものです。ポアソン回帰に過分散がない場合、統計値 -2×（ポアソン モデルの対数尤度 - 負の二項回帰の対数尤度）は、0 では半分の確率質量の混合分布を示し、自由度 1 のカイ 2 乗分布ではその残りを示すはずです。

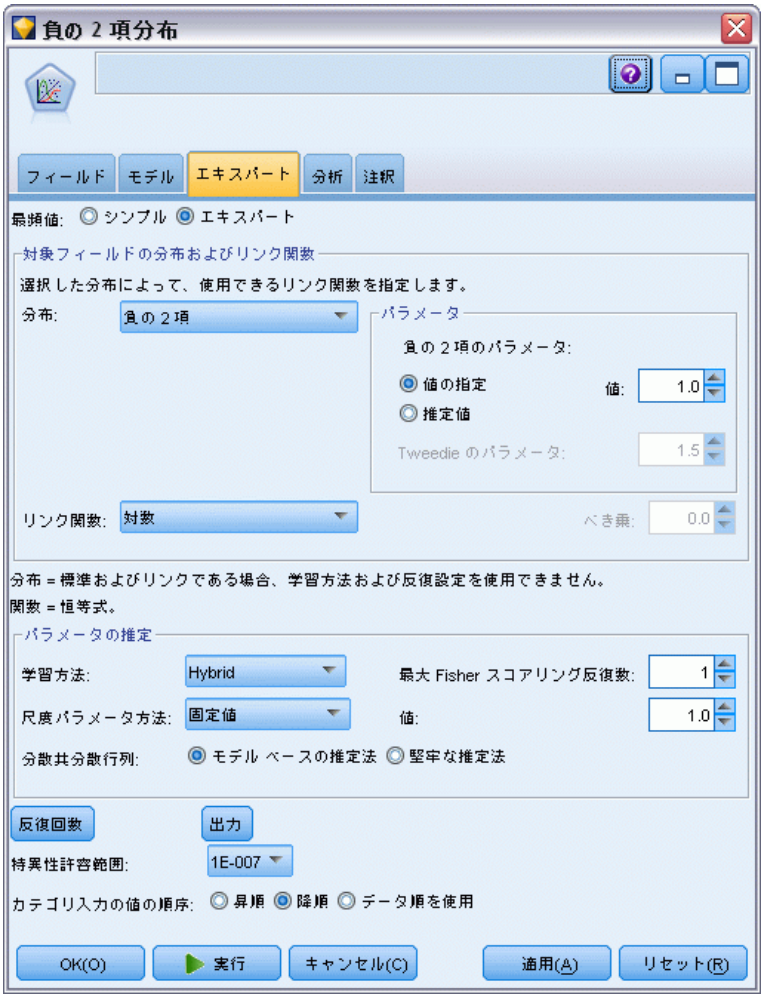
図 23-10
[エキスパート] タブ



「標準」ポアソン回帰を適合させるには、Genlin ノードを開いてコピーアンド ペーストし、入力ノードに接続して新しいノードを開き、**[エキスパート]** タブをクリックします。

- 尺度パラメータの推定方法として **[固定値]** を選択します。この値のデフォルトは 1 です。

図 23-11
[エキスパート] タブ



- 負の二項回帰を適合させるには、Genlin ノードを開いてコピー アンド ペーストし、入力ノードに接続して新しいノードを開き、**[エキスパート]** タブをクリックします。
- 分布として **[負の二項]** を選択します。補助パラメータについては、1 のデフォルト値のままにしておきます。
- 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［アドバンス］タブをブラウズします。

適合度統計

図 23-12
標準ポアソン回帰の適合度統計

	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(b)	-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: 損傷件数Model: (Intercept), 種類, 建造, 作業, offset＝log_月_サービス			
a. Information criteria are in small-is-better form.			
b. The full log likelihood function is displayed and used in computing information criteria.			

標準ポアソン回帰について示される対数尤度は -68.281 です。これと負の二項回帰を比較します。

図 23-13
負の二項回帰の適合度統計

	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(b)	-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: 損傷件数Model: (Intercept), 種類, 建造, 作業, offset＝log_月_サービス			
a. Information criteria are in small-is-better form.			
b. The full log likelihood function is displayed and used in computing information criteria.			

負の二項回帰について示される対数尤度は -83.725 です。これは実際にはポアソン回帰の対数尤度よりも「小さく」、この負の二項回帰はポアソン回帰よりも優れているわけではないということを（尤度比テストを行う必要なしに）示しています。

ただし、負の二項分布の補助パラメータについて選択した 1 の値は、このデータセットには最適ではない場合があります。過分散をテストする別の方法は、0 に相当する補助パラメータで負の二項モデルを適合させ、［エキスパート］タブの［出力］ダイアログ ボックスでラグランジュ乗数テス

トを要求することです。テストが有意でない場合、このデータセットにとって過分散は問題になりません。

要約

一般化線型モデルを使用して、カウント データについて 3 種類のモデルを適合させました。負の二項回帰は、ポア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。過分散ポアソン回帰は標準ポアソンモデルの妥当な代替案を示しているように思われますが、それを選択するための正式なテストはありません。

IBM® SPSS® Modeler で使用されているモデル作成方法の数学的な基礎の説明は、『SPSS Modeler Algorithms Guide』 に一覧されています。

章
24

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）

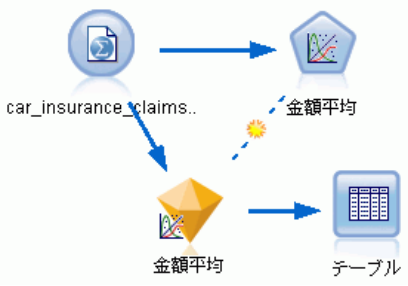
一般化線型モデルは、正の範囲のデータの分析の場合にガンマ回帰を適合させるのに使用できます。たとえば、他の場所で表示および分析されるデータセットは自動車の損害請求に関するものです。逆リンク関数を使用して従属変数の平均を予測値の線型結合に関連付けて、ガンマ分布を示すように平均的な請求額をモデル化することができます。平均請求額を計算するのに使用する請求回数の変化を説明するには、尺度の重みとして請求回数を指定します。

この例では、car-insurance_genlin.str というストリームを使用します。これは、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 フォルダにあり、ストリーム ファイルは streams サブフォルダにあります。詳細は、 1 章 Demos フォルダ in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。

ストリームの作成

- car_insurance_claims.sav を指定する Statistics ファイルの入力ノードを Demos フォルダに追加します。

図 24-1
自動車保険金請求を予測するサンプル ストリーム



- 入力ノードの［タイプ］タブで、claimamt 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。

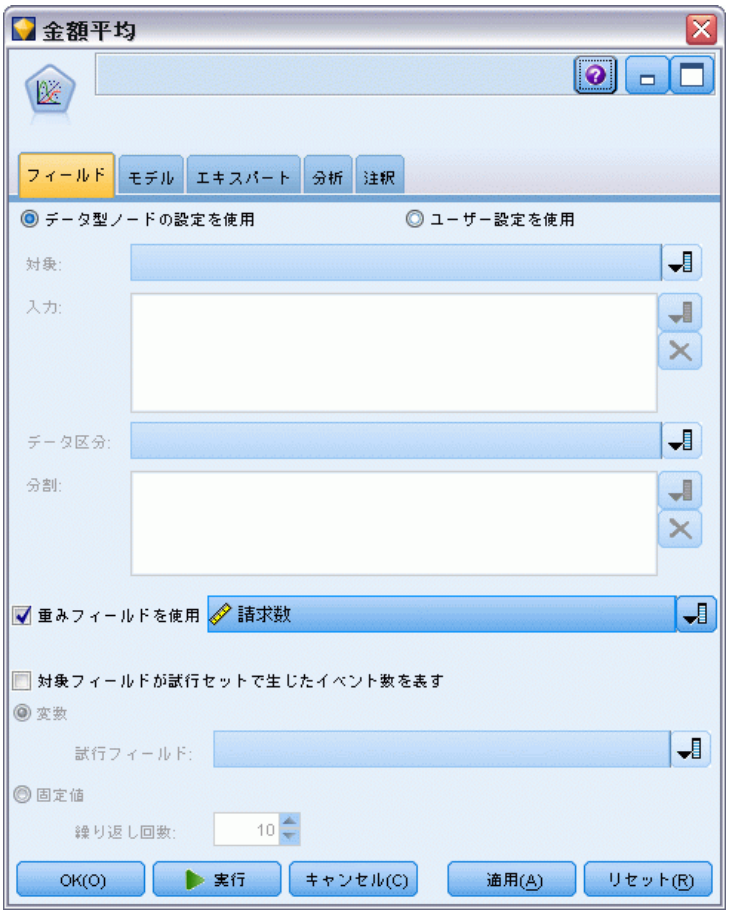
▶ [値の読み込み] をクリックしてデータをインスタンス化します。

図 24-2
フィールドの役割の設定

▶ Genlin ノードを入力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。

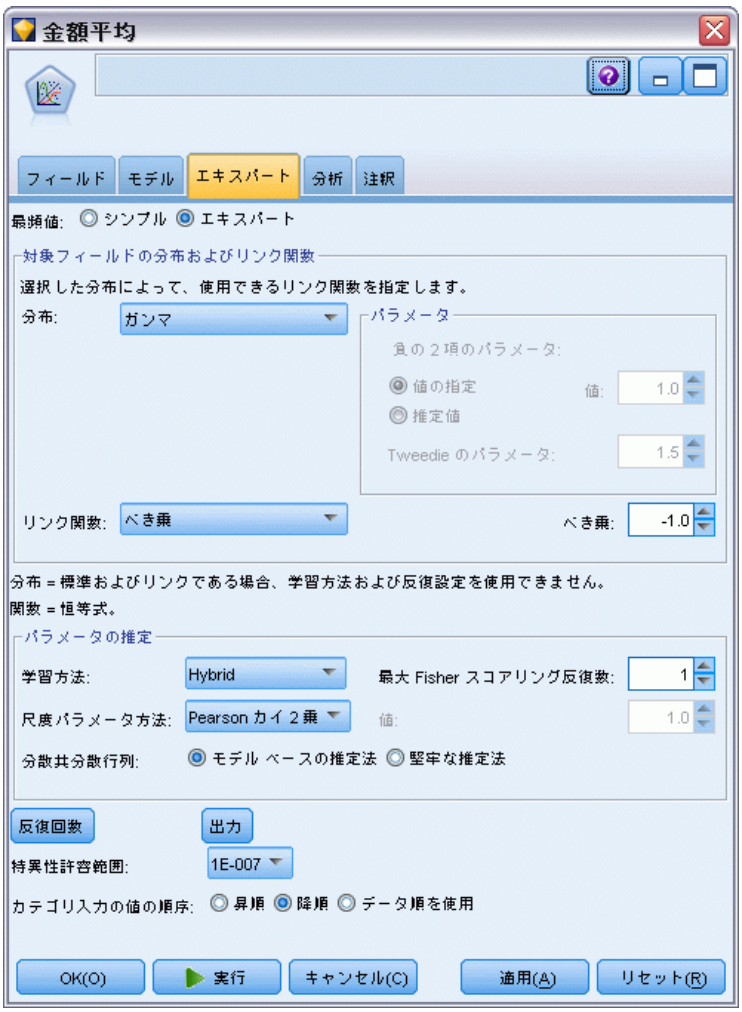
▶ 尺度重みフィールドとして nclaims を選択します。

図 24-3
フィールド オプションの選択



▶ ［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 24-4
エキスパート オプションの選択



▶ レスポンス分布として [**ガンマ**] を選択します。

▶ リンク関数として [**べき**] を選択し、べき関数の指数として -1.0 を入力します。これは逆リンクです。

▶ 尺度パラメータの推定方法として [**Pearson のカイ 2 乗**] を選択します。これは McCullagh と Nelder が使用する方法ですから、結果を反復するには、ここではそれに従います。

- 因子のカテゴリ順序として [降順] を選択します。これは、それぞれの因子の第 1 のカテゴリは参照カテゴリであるということを示します。モデルに関するこの選択の影響は、パラメータ推定値の解釈に反映されます。

- [実行] をクリックしてモデル ナゲットを生成します。生成されたモデル ナゲットはストリーム領域、そして右上の［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て、[編集] または [ブラウズ] を選択し、［アドバンス］を選択します。

パラメータ推定値

図 24-5
パラメータ推定値(M)

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.003	.0004	.003	.004	66.593	1	.000
[年齢=8]	.001	.0004	.000	.002	4.898	1	.027
[年齢=7]	.001	.0004	.000	.002	5.046	1	.025
[年齢=6]	.001	.0004	.000	.002	5.740	1	.017
[年齢=5]	.001	.0004	.001	.002	10.682	1	.001
[年齢=4]	.000	.0004	.000	.001	1.268	1	.260
[年齢=3]	.000	.0004	.000	.001	.720	1	.396
[年齢=2]	.000	.0004	-.001	.001	.054	1	.816
[年齢=1]	0(a)	.	.	.	.	.	.
[車グループ=4]	-.001	.0002	-.002	-.001	61.883	1	.000
[車グループ=3]	-.001	.0002	-.001	.000	13.039	1	.000
[車グループ=2]	3.77E-005	.0002	.000	.000	.050	1	.823
[車グループ=1]	0(a)	.	.	.	.	.	.
[車年齢=4]	.004	.0004	.003	.005	88.175	1	.000
[車年齢=3]	.002	.0002	.001	.002	53.013	1	.000
[車年齢=2]	.000	.0001	.000	.001	13.191	1	.000
[車年齢=1]	0(a)	.	.	.	.	.	.
(Scale)	1.209(b)						

Dependent Variable: 請求の平均金額Model: (Intercept), 年齢, 車グループ, 車年齢

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

オムニバス検定およびモデル効果検定（表示されていない）は、そのモデルがヌル モデルよりも優れていること、および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。パラメータ推定値テーブルは因子レベルおよび尺度パラメータについて McCullagh と Nelder が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が得られます。一般的には、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。しかし、対数リンクは指数が0の場合のべきリンクの特殊なケースですから、対数リンクを持つモデルとべき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 で使用されているモデル作成方法の数学的な基礎の説明は、『SPSS Modeler Algorithms Guide』 に一覧されています。

章
25

細胞サンプルの分類（SVM）

Support Vector Machine（SVM）は特に広範なデータセット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は多数の予測値があるもので、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）の分野で遭遇する可能性があります。

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ます。元のデータの分析では、良性と悪性のサンプルの間で、多数の特性が大きく異なることがわかりました。研究者は、他の患者から採取したサンプルのこれら細胞の特性の値を使用できる SVM モデルを開発し、サンプルが良性または悪性かを早期に特定できるようにしたいと考えています。

この例では、streams サブフォルダの下の Demos フォルダ内にある svm_cancer.str という名前のストリームを使用します。データ ファイルは、cell_samples.data です。 詳細は、 1 章 Demos フォルダ in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。

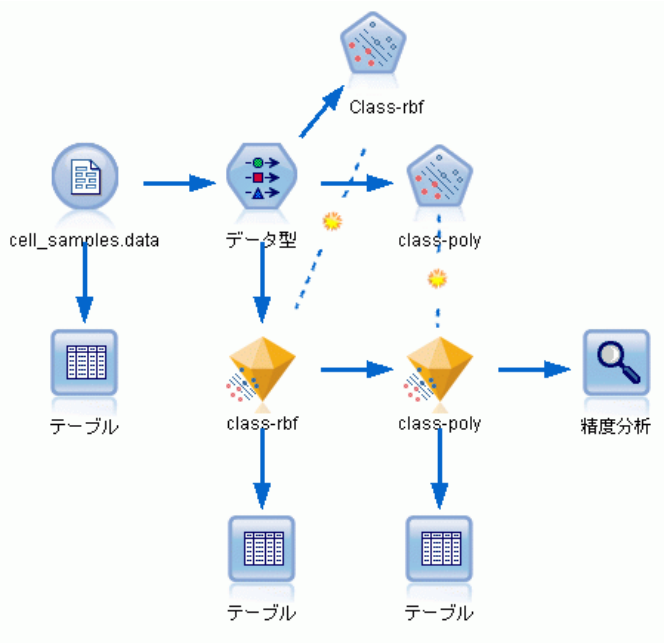
例は UCI マシン学習リポジトリ(アスンシオン および ニューマン, 2007)から公表されているデータセットに基づいています。データセットは何百ものヒト細胞サンプル レコードで構成されており、それぞれに一連の細胞特性の値が含まれています。各レコードのフィールドは次のとおりです。

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

この例の目的のために、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 25-1
SVM モデリングを表示するサンプル ストリーム



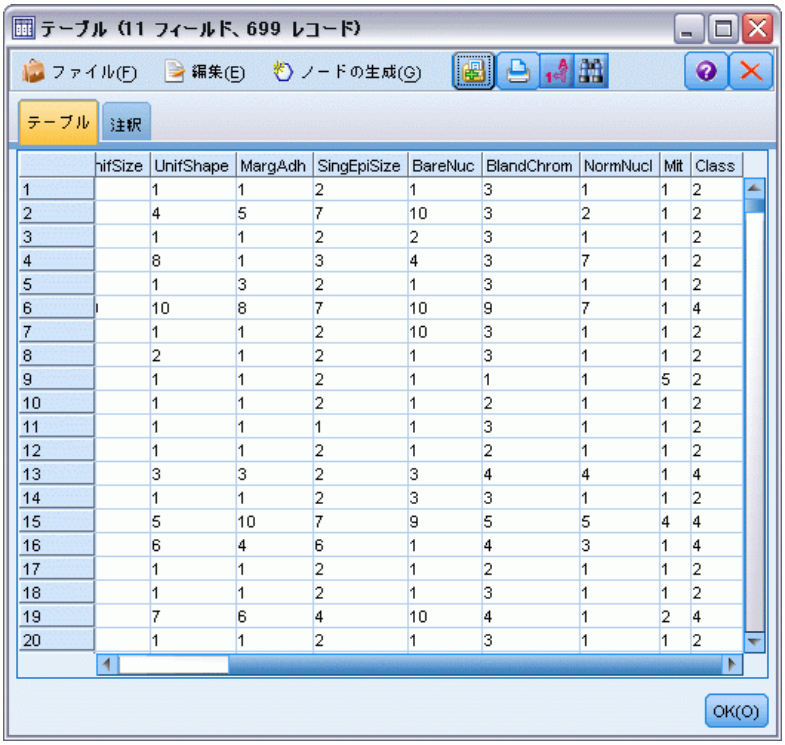
- 新規のストリームを作成し、IBM® SPSS® Modeler インストールの Demos フォルダにある cell_samples.data を示す可変長ファイル入力ノードを追加します。

  ソース ファイルのデータを見てみましょう。

- テーブル ノードをストリームに追加します。

- テーブルノードを可変長ファイル ノードに接続してストリームを実行します。

図 25-2
SVM のソース・データ



	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

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 フィールドには別の医療処置で確認された診断が含まれており、サンプルが良性 （値 = 2） または悪性 （値 = 4） であるかを示しています。

図 25-3
データ型ノードの設定

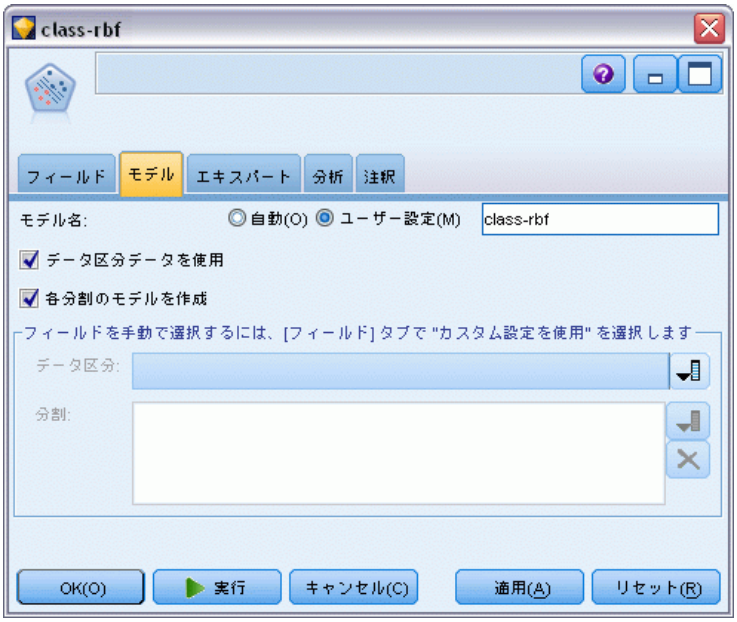
- データ型ノードを追加し、可変長ファイル ノードへ接続します。
- データ型ノードを開きます。

  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、測定レベルを変更してこれを反映する必要があります。

- Class フィールド（リストの最後）の **[尺度]** 列で、**[連続型]** 値をクリックして **[フラグ]** に変更します。
- **[値の読み込み]** をクリックします。
- **[役割]** 列で、モデルの予測値や対象として使用されないため、[ID]（患者の ID）の役割を **[なし]** に設定します。
- 対象の Class の役割を **[対象]** に設定し、その他のフィールド（予測値）の方向はすべて **[入力]** にしておきます。
- [OK] をクリックします。

  SVM ノードはその処理の実行にカーネル関数の選択を提供します。指定されたデータセットにどの関数を実行するのがベストかを知るのは容易でないため、順番に様々な関数を選択して結果を比較します。デフォルトの RBF（Radial Basis Function）から始めます。

図 25-4
モデル タブ設定



- ［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。
- データ型ノードを開きます。**[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 を入力します。

図 25-5
デフォルトの [エキスパート] タブの設定

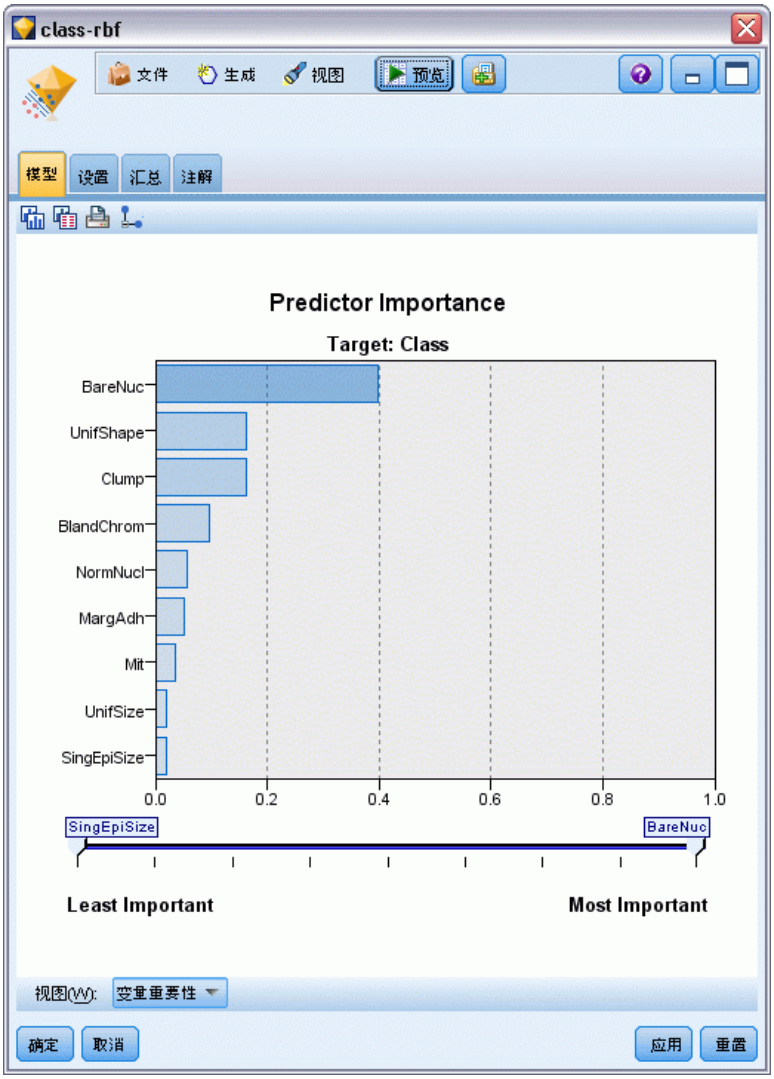
▶ [エキスパート] タブで、[モード] を [エキスパート] に設定しますが、すべてのデフォルト設定はそのままにします。デフォルトでは、[カーネル タイプ] は [RBF] に設定されます。簡易モードではオプションはすべてグレイアウトされています。

図 25-6
タブ設定の分析

- [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。
- [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 ナゲットが追加されます。
- ストリーム中のモデル ナゲットをダブル クリックします。

データの調査

図 25-7
予測重要度グラフ

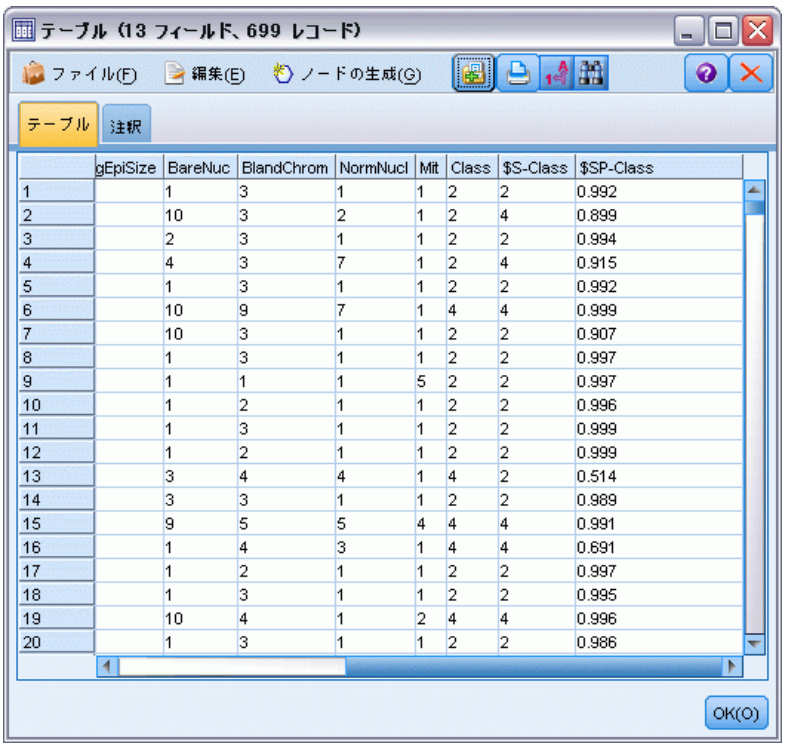


[モデル] タブで、予測重要度グラフは予測での様々なフィールドの相対効果を示しています。これにより、 BareNuc が最大効果を持っているのが一目瞭然で、また UnifShape および Clumpもかなりの効果があることがわかります。

- [OK] をクリックします。
- テーブルノードを class-rbf デル ナゲットに接続します。

▶ テーブルノードを開いて、[実行] をクリックします。

図 25-8
予測値および確信値に追加するフィールド



	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

▶ モデルは 2 つの新しいフィールドを作成しています。右側に出力されたテーブルをスクロールして見てください。

新しいフィールド名	説明
$S-Class	モデルで予測された 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）が適度に高いことがわかります。

ただし、顕著な例外がいくつかあります。たとえば 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いです。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でも（たとえば、2 行目と 4 行目）このモデルには間違った予測がたくさんあることがわかります。

異なる関数タイプを選んで、これを改善できるか試してみましょう。

異なる関数を試す

図 25-9
モデルの新しい名前の設定

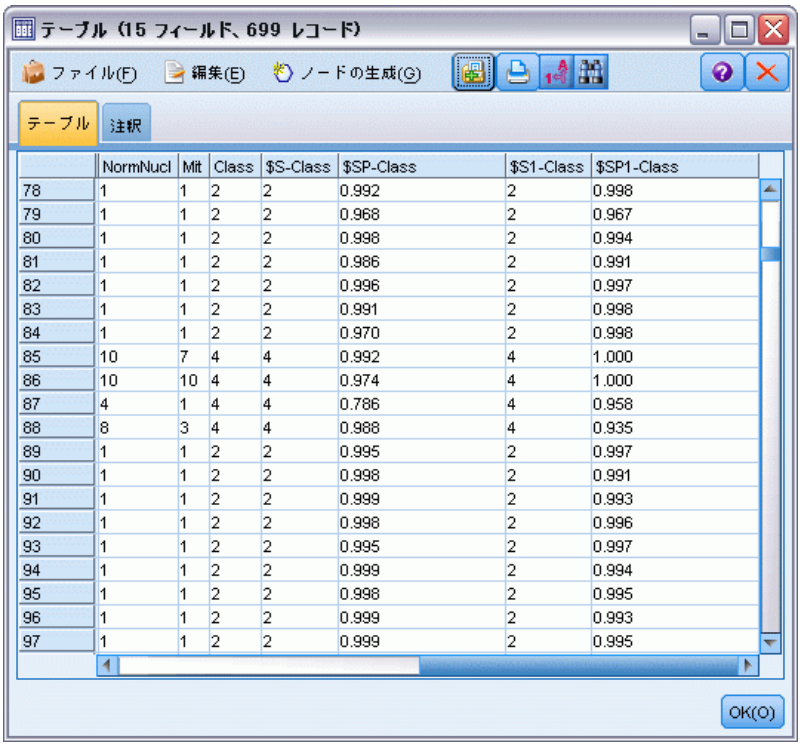
- テーブル出力ウィンドウを閉じます。
- 2 番目のモデル作成ノードをデータ型ノードに接続します。
- 新しい SVM ノードを開きます。
- **[モデル]** タブで［カスタム］をクリックし、モデル名に class-poly を入力します。

図 25-10
多項式の [エキスパート] タブ設定

- [**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。
- **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］パレットにモデル ナゲットが追加されます。
- class-rbf モデル ナゲットを class-poly 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）。
- テーブルノードを class-poly モデル ナゲットに接続します。
- 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

図 25-11
多項式関数に追加されたフィールド



	NormNucl	Mit	Class	$S-Class	$SP-Class	$S1-Class	$SP1-Class
78	1	1	2	2	0.992	2	0.998
79	1	1	2	2	0.968	2	0.967
80	1	1	2	2	0.998	2	0.994
81	1	1	2	2	0.986	2	0.991
82	1	1	2	2	0.996	2	0.997
83	1	1	2	2	0.991	2	0.998
84	1	1	2	2	0.970	2	0.998
85	10	7	4	4	0.992	4	1.000
86	10	10	4	4	0.974	4	1.000
87	4	1	4	4	0.786	4	0.958
88	8	3	4	4	0.988	4	0.935
89	1	1	2	2	0.995	2	0.997
90	1	1	2	2	0.998	2	0.991
91	1	1	2	2	0.999	2	0.993
92	1	1	2	2	0.998	2	0.996
93	1	1	2	2	0.995	2	0.997
94	1	1	2	2	0.999	2	0.994
95	1	1	2	2	0.998	2	0.995
96	1	1	2	2	0.999	2	0.993
97	1	1	2	2	0.999	2	0.995

- 右側に出力されたテーブルをスクロールして、新しく追加されたフィールドを見ます。

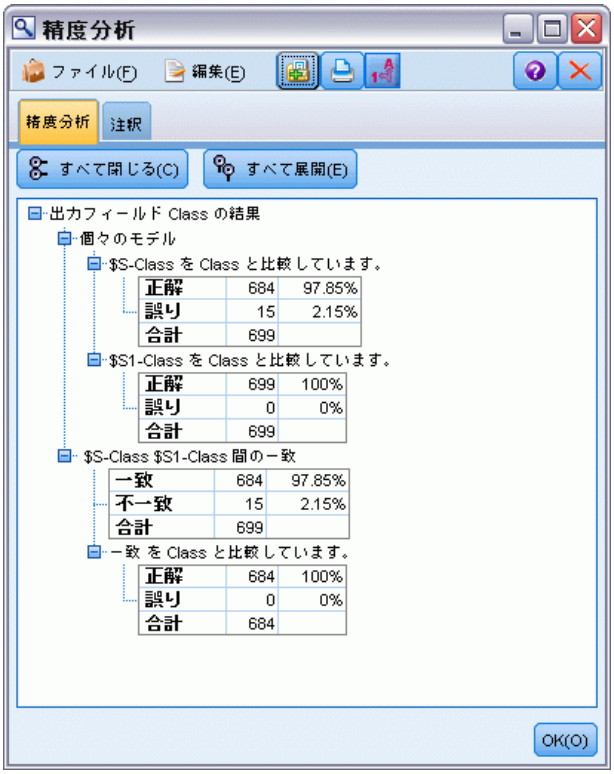
  多項式関数タイプのための生成されたフィールドは、$S1-Class および $SP1-Class という名前です。

  多項式の結果が改善されたように見えます。傾向スコアの大半が 0.995 かそれ以上、これは非常に有望です。

- モデルでの改善を確認するため、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 25-12
精度分析ノード



分析ノードを使ったこの技術で、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。分析モードからの出力で、RBF 関数がケースの 97.85% を正確に予測したことがわかり、これは非常に良好です。ただし、この出力で多項式関数がどの単一ケースでも診察を正確に予測したことがわかります。実際は 100% の精度はなかなか見られませんが、モデルが特定のアプリケーション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と同様には機能しません。しかし別のデータセットでは結果が異なる可能性があります。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 カーネル関数の様々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様々な結果を出すか、そしてモデルごとの改善をどのように判定できるかを学びました。

章
26

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

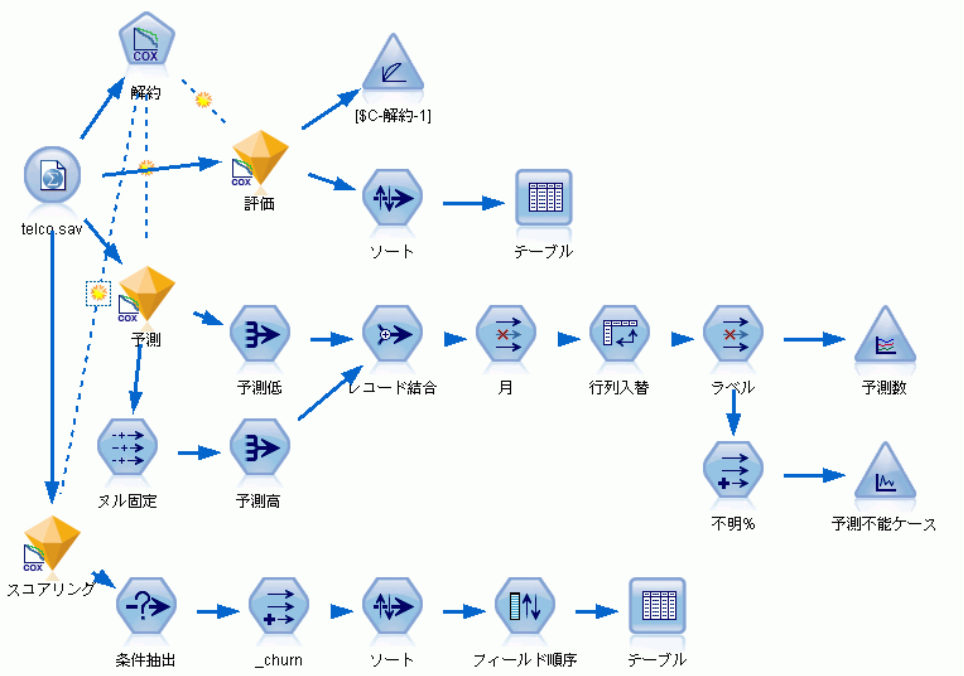
顧客離れを減らすという目標の一環として、ある通信会社は「解約するまでの期間」のモデル作成に注目し、解約して他社のサービスに切り替える顧客と関連する要因を特定します。そのために、顧客の無作為サンプルが選択され、顧客としての期間（アクティブな顧客かどうかに関係なく）やその他のさまざまなフィールドがデータベースから取り出されます。

この例では、ストリーム telco_coxreg.str を使用します。このストリームは、データ ファイル telco.sav を参照します。データ ファイルは Demos フォルダにあり、ストリーム ファイルは streams サブフォルダにあります。 詳細は、 1 章 Demos フォルダ in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。

適切なモデルの構築

► telco.sav を指定するStatistics ファイル入力ノードを Demos フォルダに追加します。

図 26-1
解約するまでの期間を分析するためのサンプル ストリーム



▶ 入力ノードの［フィルタ］タブで、region、income、longten から wireten まで、および loglong から logwire までを除外します。

図 26-2
不必要なフィールドのフィルタリング

（代わりに、［データ型］タブでこれらのフィールドの役割を除外するのではなく［なし］に変更するか、あるいはモデル作成ノードで使用したいフィールドを選択することもできます。）

▶ 入力ノードの［データ型］タブで、churn フィールドの役割を［対象］に設定し、その測定レベルを［フラグ型］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。

► [値の読み込み] をクリックしてデータをインスタンス化します。

図 26-3
フィールドの役割の設定

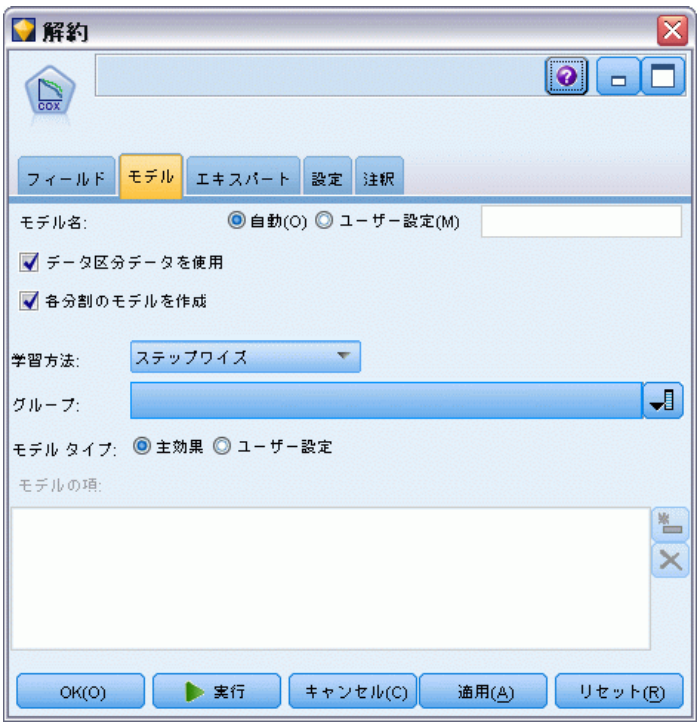
▶ Cox ノードを入力ノードに接続します。[フィールド] タブで生存時間変数として tenure を選択します。

図 26-4
フィールド オプションの選択

▶ [モデル] タブをクリックします。

▶ 変数選択手法として［**ステップワイズ法**］を選択します。

図 26-5
モデル オプションの選択



▶ ［**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。

▶ [出力] をクリックします。

図 26-6
詳細出力オプションの選択

▶ 生成するプロットとして [生存関数] と [ハザード関数] を選択し、[OK] をクリックします。

▶ [実行] をクリックしてモデル ナゲットを生成します。生成されたモデル ナゲットはストリーム、そして右上の [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。まず、[詳細出力] タブを調べます。

打ち切りケース

図 26-7
ケース処理要約(S)

		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: サービスの継続期間(月単位)			

ステータス変数は、与えられた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います。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。処理したケースの要約には、726 個のケースが打ち切られたことが示されています。これらは、解約していない顧客です。

カテゴリ変数のコード化

図 26-8
カテゴリ変数のコード化

		Frequency	(1)(s)	(2)	(3)	(4)
婚姻(t)	0=未婚	505	1			
	1=既婚	495	0			
教育(t)	1=高校未満	204	1	0	0	0
	2=高校卒	287	0	1	0	0
	3=短大卒	209	0	0	1	0
	4=4年生大学卒	234	0	0	0	1
	5=大学院卒	66	0	0	0	0
退職(t)	0=いいえ	953	1			
	1=はい	47	0			
性別(t)	0=男性	483	1			
	1=女性	517	0			
無料通話(t)	0=いいえ	526	1			
	1=はい	474	0			
設備(t)	0=いいえ	614	1			
	1=はい	386	0			
コールカード(t)	0=いいえ	322	1			
	1=はい	678	0			
ワイヤレス(t)	0=いいえ	704	1			
	1=はい	296	0			
複数ライン(t)	0=いいえ	525	1			
	1=はい	475	0			
ボイス(t)	0=いいえ	696	1			
	1=はい	304	0			
ポケベル(t)	0=いいえ	739	1			
	1=はい	261	0			
インターネット(t)	0=いいえ	632	1			
	1=はい	368	0			
コールID(t)	0=いいえ	519	1			
	1=はい	481	0			
キャッチ(t)	0=いいえ	515	1			
	1=はい	485	0			
転送(t)	0=いいえ	507	1			
	1=はい	493	0			
スリーホン(t)	0=いいえ	498	1			
	1=はい	502	0			
電子請求(t)	0=いいえ	629	1			
	1=はい	371	0			
顧客(t)	1=ベーシック サービス	266	1	0	0	
	2=E-サービス	217	0	1	0	
	3=プラス サービス	281	0	0	1	
	4=トータル サービス	236	0	0	0	

カテゴリ変数のコード化は、カテゴリ共変量の回帰係数、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、参照カテゴリは「最後の」カテゴリです。したがって、たとえば、既婚の顧客はデータ ファイル内で 1 の変数値を持っていても、回帰の目的では 0 とコード化されます。

変数選択

図 26-9
オムニバス検定

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(c)	3392.536	162.303	1	.000	133.828	1	.000	133.828	1	.000
2(d)	2904.101	253.576	2	.000	488.436	1	.000	622.263	2	.000
3(e)	2854.720	331.799	3	.000	49.381	1	.000	671.644	3	.000
4(f)	2802.945	510.853	4	.000	51.775	1	.000	723.419	4	.000
5(g)	2778.003	528.033	5	.000	24.942	1	.000	748.361	5	.000
6(h)	2779.439	489.249	4	.000	1.436	1	.231	746.925	4	.000
7(i)	2764.575	512.693	5	.000	14.863	1	.000	761.789	5	.000
8(j)	2755.972	522.582	6	.000	8.603	1	.003	770.392	6	.000
9(k)	2748.905	526.103	7	.000	7.067	1	.008	777.459	7	.000
10(l)	2742.366	572.931	8	.000	6.539	1	.011	783.998	8	.000
11(m)	2736.517	591.096	9	.000	5.849	1	.016	789.847	9	.000
12(n)	2731.530	596.063	10	.000	4.987	1	.026	794.834	10	.000
13(o)	2727.484	596.203	11	.000	4.046	1	.044	798.880	11	.000

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 3526.364
b. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)
c. Variable(s) Entered at Step Number 1: コールカード
d. Variable(s) Entered at Step Number 2: 長距離計
e. Variable(s) Entered at Step Number 3: 設備
f. Variable(s) Entered at Step Number 4: 設備計
g. Variable(s) Entered at Step Number 5: 雇用
h. Variable Removed at Step Number 6: コールカード
i. Variable(s) Entered at Step Number 7: ボイス
j. Variable(s) Entered at Step Number 8: 住所
k. Variable(s) Entered at Step Number 9: 電子請求
l. Variable(s) Entered at Step Number 10: コールカード
m. Variable(s) Entered at Step Number 11: 複数ライン
n. Variable(s) Entered at Step Number 12: コールID
o. Variable(s) Entered at Step Number 13: スリーホン

モデル構築プロセスは、変数増加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のモデルの−2 対数尤度の間の差です。 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、その投入は有意義です。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、その除外は有意義です。12 ステップの場合、12 個の変数がモデルに追加されます。

図 26-10
式内の変数 (ステップ 12 のみ)

		B	SE	Wald	df	Sig.	Exp(B)
Step 12	住所	-.026	.009	8.007	1	.005	.974
	雇用	-.041	.010	16.803	1	.000	.960
	設備	-1.930	.241	64.046	1	.000	.145
	コールカード	.523	.152	11.775	1	.001	1.687
	長距離計	-.006	.001	124.874	1	.000	.994
	設備計	-.001	.000	58.559	1	.000	.999
	複数ライン	.316	.134	5.517	1	.019	1.371
	ボイス	-.488	.149	10.737	1	.001	.614
	コールID	-.330	.147	5.035	1	.025	.719
	電子請求	-.439	.151	8.392	1	.004	.645

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) を見てみましょう。これは、予測値のハザードにおける変化をユニットの増分で解釈します。

- address の Exp(B) の値は、解約ハザードが、顧客が同じ場所に居住する年ごとに 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 年間居住した顧客の解約ハザードは、100%−(100%×0.966^5)=15.88% 減少しています。
- callcard の Exp(B) の値は、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が、同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ています。カテゴリ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) の値は、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が、同サービスに加入している顧客より 0.697 倍大きいことを意味します。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多く解約していることを示しており、いささか心配です。

図 26-11
モデル内にない変数（ステップ 12 のみ）

		Score	df	Sig.
Step 12	年齢	.557	1	.456
	婚姻	.280	1	.597
	教育	5.022	4	.285
	教育(1)	.535	1	.465
	教育(2)	.034	1	.854
	教育(3)	1.519	1	.218
	教育(4)	4.327	1	.038
	退職	.003	1	.959
	性別	.363	1	.547
	世帯人数	2.274	1	.132
	無料通話	.310	1	.578
	ワイヤレス	.907	1	.341
	通話料計	2.511	1	.113
	カード計	1.976	1	.160
	ワイヤレス計	2.175	1	.140
	ポケベル	.019	1	.890
	インターネット	2.009	1	.156
	キャッチ	1.899	1	.168
	転送	.001	1	.978
	スリーホン	4.022	1	.045
	ln収入	2.293	1	.130
	顧客	4.578	3	.205
	顧客(1)	3.077	1	.079
	顧客(2)	.125	1	.724
	顧客(3)	.016	1	.899

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 以上です。ただし、tollfree と cardmon の有意水準は、0.05 より小さくはありませんが、かなり近いです。これらについて、さらに調べていくと面白いでしょう。

共変量平均

図 26-12
共変量平均

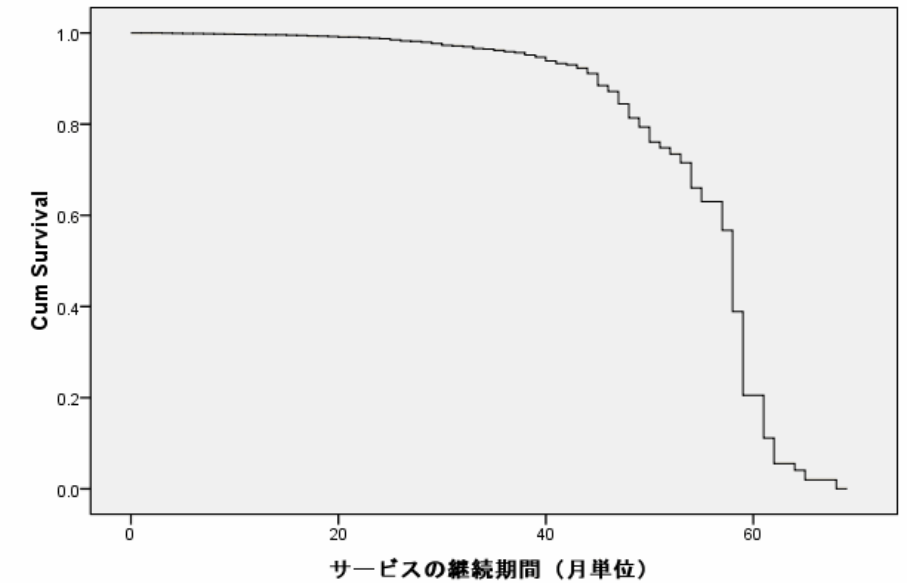
	Mean
年齢	41.684
婚姻	.505
住所	11.551
教育(1)	.204
教育(2)	.287
教育(3)	.209
教育(4)	.234
雇用	10.987
退職	.953
性別	.483
世帯人数	2.331
無料通話	.526
設備	.614
コールカード	.322
ワイヤレス	.704
長距離計	574.050
通話料計	551.259
設備計	465.633
カード計	605.774
ワイヤレス計	442.737
複数ライン	.525
ボイス	.696
ポケベル	.739
インターネット	.632
コールID	.519
キャッチ	.515
転送	.507
スリーホン	.498
電子請求	.629
ln収入	3.957
顧客(1)	.266
顧客(2)	.217
顧客(3)	.281

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、平均値のために構築される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、カテゴリ予測値の指標変数の平均を見る場合、「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の場合でも、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るには、[プロット] ダイアログ ボックスで、生存曲線がプロットされる部分の共変量の値を変更します。特定のケースの生存曲線を見るには、[詳細出力] ダイアログ ボックスの [プロット] グループで、生存曲線がプロットされる部分の共変量の値を変更します。

生存曲線

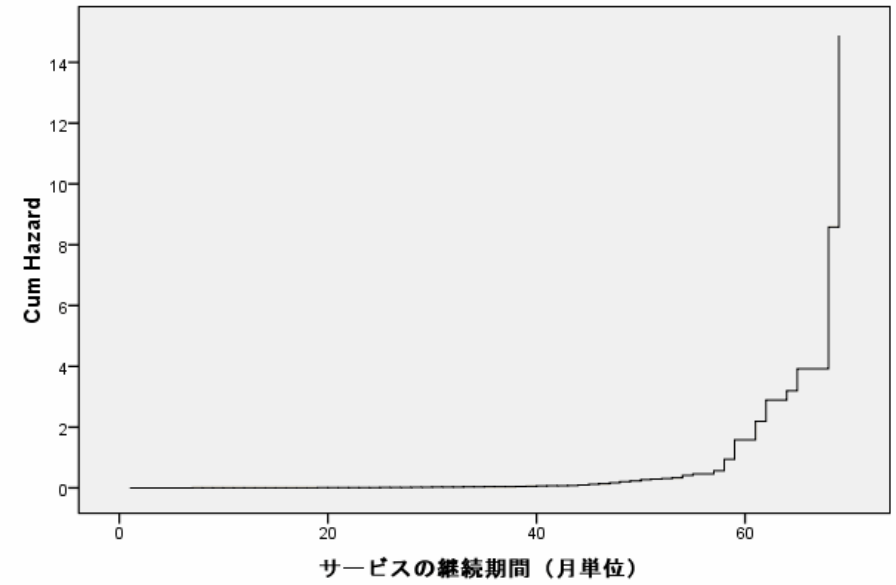
図 26-13
「平均」の顧客の生存曲線



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たものです。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、生存曲線の任意のポイントは、「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 か月を過ぎると、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 26-14
「平均」の顧客のハザード曲線



基本的な生存曲線は、モデルが予測した、「平均」の顧客が解約する累積的な可能性を視覚的に表示したものです。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価である累積ハザードを示します。 55 か月を過ぎると、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

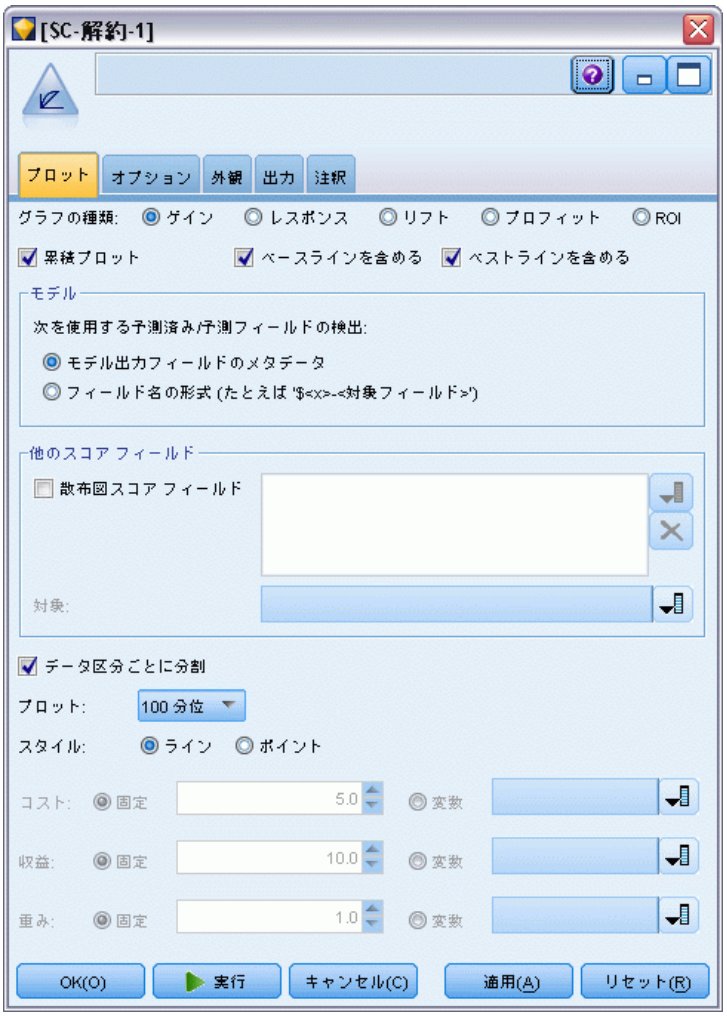
ステップワイズ選択法は、モデルに「統計的に有意な」予測値だけが含まれることを確実にしますが、そのモデルが実際に対象を予測するのに最適であることを保証するものではありません。そのためには、スコアリングされたレコードを分析する必要があります。

図 26-15
Cox ナゲットの [設定] タブ

- モデル ナゲットをキャンバスに配置し、それを入力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。
- **[時間フィールド]** を選択し、tenure を指定します。各レコードは、その保有期間の長さでスコアリングされます。
- **[すべての確率を追加]** を選択します。

これにより、0.5 を顧客が解約するかどうかの閾値として使用するスコアが作成され、解約の傾向が 0.5 以上であれば、それらの顧客は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、異なる閾値を使えば、もっと理想的な結果が得られる可能性があります。閾値の選択を考慮する 1 つの方法は、評価ノードを使用することです。

図 26-16
評価ノードの [プロット] タブ

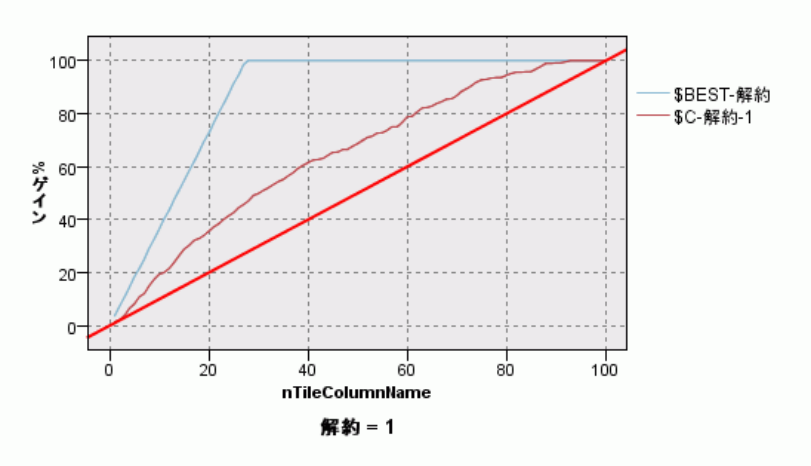


- 評価ノードをモデル ナゲットに接続します。[プロット] タブで、[**ベスト ラインを含める**] を選択します。
- [**オプション**] タブをクリックします。

図 26-17
評価ノードの [オプション] タブ

- [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。
- [実行] をクリックします。

図 26-18
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目標にすることで、特定のカテゴリ「ゲイン」のケースの総数のパーセントを示します。たとえば、曲線の 1 つのポイントは（10%, 15%）です。これは、モデル内のデータセットをスコアリングし、すべてのケースを解約の予測傾向によってソートする場合、上位 10% に、カテゴリ 1（解約者）をとるケース全体の約 15% が含まれてい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。

対角線が「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無作為にレコードの 20% を選択すると、カテゴリ 1 をとる全レコードの約 20% を「取得する」と予測できます。曲線がベースラインより上になるほど、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、目標のゲインに対応するパーセントを選択し、パーセントを適切な分割値にマッピングすることで、分類の分割を選択できます。

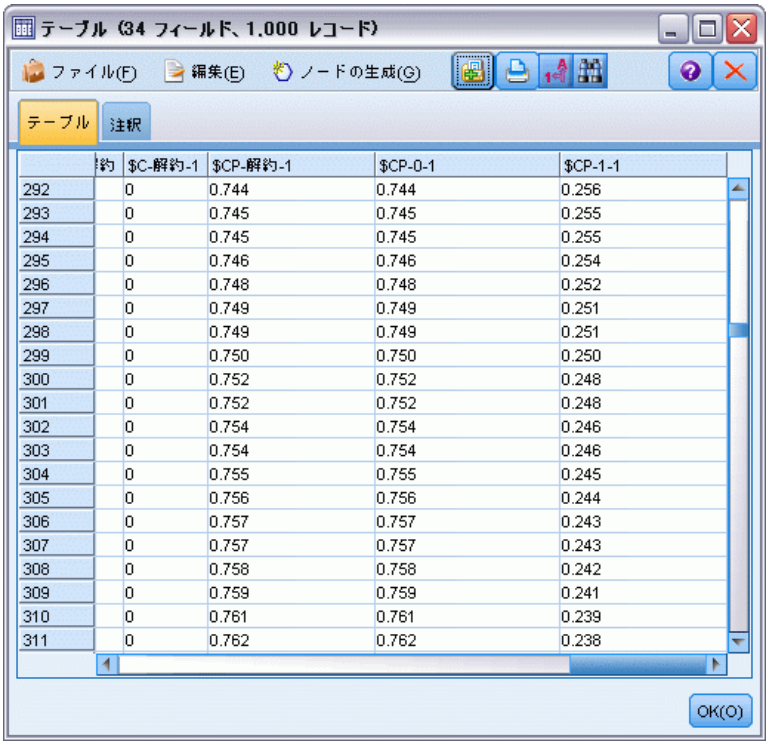
何が「望ましい」ゲインを構成するかは、タイプ I およびタイプ II の誤りにかかるコストに依存します。解約者を非解約者と分類することによるコストは何か（タイプ I）？非解約者を解約者と分類することによるコストは何か（タイプ II）？です。顧客の固定客化が主な考慮事項であれば、タイプ I の誤差を小さくします。累積ゲイン グラフでは、予測された傾向が 1 である上位 60% の顧客に対する顧客サービスを向上させることに対応するでしょう。これにより解約の可能性のある顧客の 79.2% を獲得できますが、それにかかる時間とリソースを、新規顧客の獲得のほうに費やすこともできます。 現在の顧客ベースを維持するためコストを下げることが最優先であれば、タイプ II の誤差を小さくします。グラフでは、上位 20% の顧客サービスを向上させることに対応するでしょう。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 26-19
ソート ノードの [設定] タブ

- たとえば、45.6% が望まれるゲインと決定したとします。これは、レコードの上位 30% をとります。適切な分類の閾値を見つけるには、ソート ノードをモデル ナゲットに接続します。
- [設定] タブで、$CP-1-1 によって降順にソートするよう選択し、[OK] をクリックします。

図 26-20
テーブル



- テーブル ノードをこのソート ノードに接続します。
- テーブルノードを開いて、[実行] をクリックします。

出力を下にスクロールすると、$CP-1-1 の値が 300 番目のレコードで 0.248 だということがわかります。0.248 を分類の閾値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% であり、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

モデルに満足したら、今後 2 年間にわたって維持できると予測される、データセット内の顧客数を追跡します。ヌル値、つまり、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内の生存時間の範囲を超えている顧客は、大変興味深い検討対象です。それらを扱う 1 つの方法は、ヌル値をすでに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。

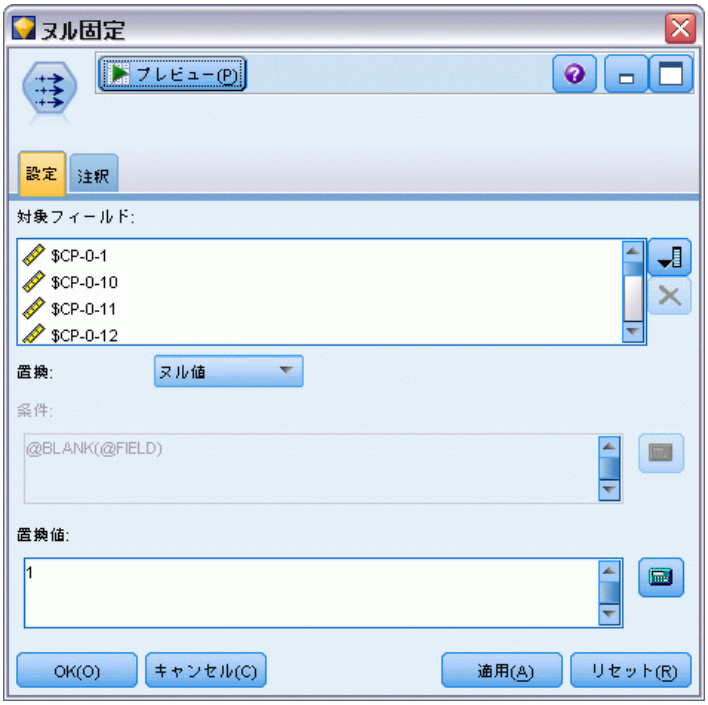
図 26-21
Cox ナゲットの [設定] タブ

- [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、新しいナゲットを入力ノードに接続します。
- ナゲットを［設定］タブに開きます。
- **[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」**を指定し、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 か月間、毎月スコアリングされます。
- 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。スコアリングのアルゴリズムは、各顧客の会社の顧客としての時間の長さを考慮します。
- **[すべての確率を追加]** を選択します。

図 26-22
レコード集計ノードの [設定] タブ

- レコード集計ノードをモデル ナゲットに接続します。[設定] タブで、デフォルトのモードとしての [**平均**] を選択解除します。
- 集計するフィールドとして、$CP-0-1 から $CP-0-24 まで、フォーム $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。
- [**フィールドにレコード度数を含める**] を選択解除します。
- [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 26-23
置換ノードの [設定] タブ



- 先ほど集計レコード ノードを接続した Cox 回帰ノード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、フォーム $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。
- [ヌル値] を値 1 と置換することを選択します。
- [OK] をクリックします。

図 26-24
レコード集計ノードの [設定] タブ

- レコード集計ノードを置換ノードに接続します。[設定] タブで、デフォルトのモードとしての [**平均**] を選択解除します。
- 集計するフィールドとして、$CP-0-1 から $CP-0-24 まで、フォーム $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。
- [**フィールドにレコード度数を含める**] を選択解除します。
- [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 26-25
フィルタ ノードの［設定］タブ

- 1 つのレコード追加ノードを 2 つのレコード集計ノードに接続します。その後、フィルタ ノードをレコード追加ノードに接続します。
- フィルタ ノードの［設定］タブで、フィールドの名前を 1 から 24 に変更します。行列入替ノードを使用すると、これらのフィールド名は、下流のグラフの x 軸の値になります。

図 26-26
行列入替ノードの [設定] タブ

- 行列入替ノードをフィルタ ノードに接続します。
- 新しいフィールドの数として「2」を入力します。

図 26-27
フィルタ ノードの [フィルタ] タブ

- フィルタ ノードを行列入替ノードに接続します。
- ［フィルタ］ノードの［設定］タブで、ID を月に変更し、フィールド 1 を推定値の下限とし、フィールド 2 を推定値の上限とします。

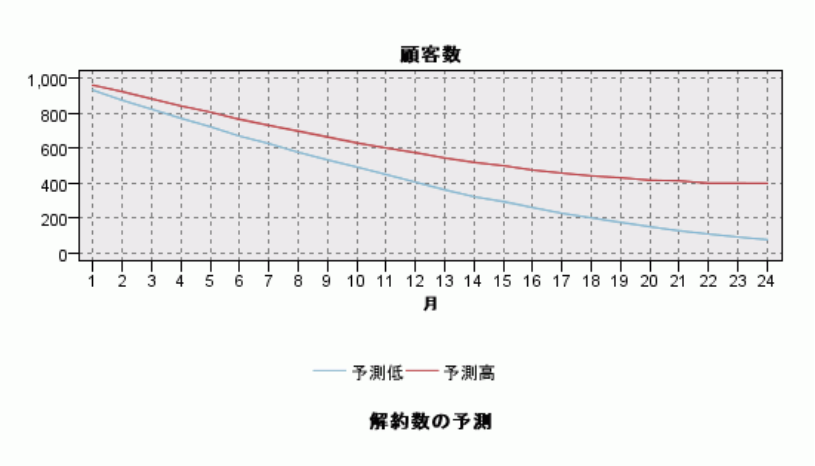
図 26-28
線グラフ ノードの [プロット] タブ

- 線グラフ ノードをフィルタ ノードに接続します。
- [プロット] タブで、X フィールドとして 月 を選択し、Y フィールドとして推定値の下限と推定値の上限 を選択します。

図 26-29
線グラフ ノードの [外観] タブ

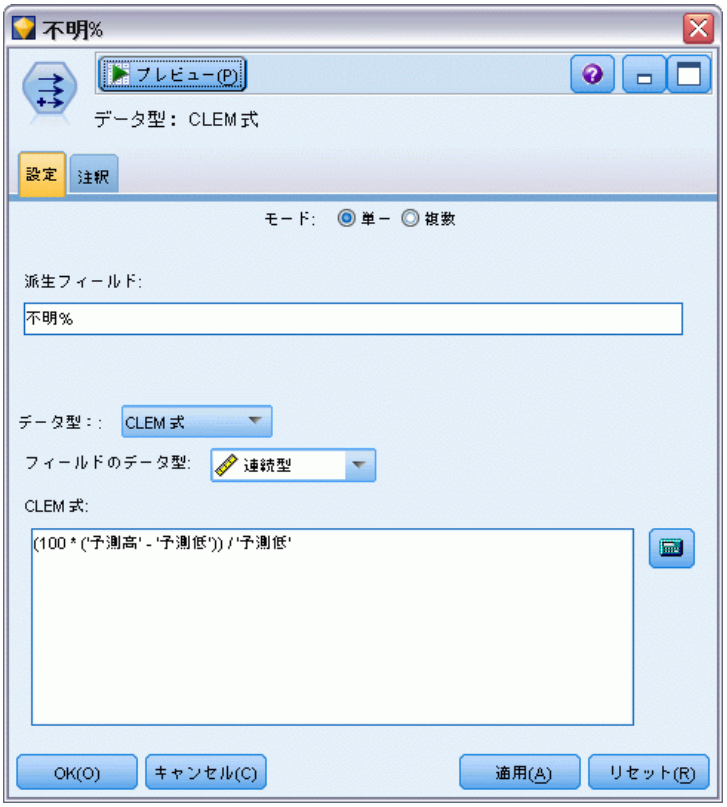
- ▶ [外観] タブをクリックします。
- ▶ タイトルとして、**「顧客数」**と入力します。
- ▶ 解説として、**「固定客となる顧客数を推定」**と入力します。
- ▶ **[実行]** をクリックします。

図 26-30
固定客となる顧客数を推定する線グラフ



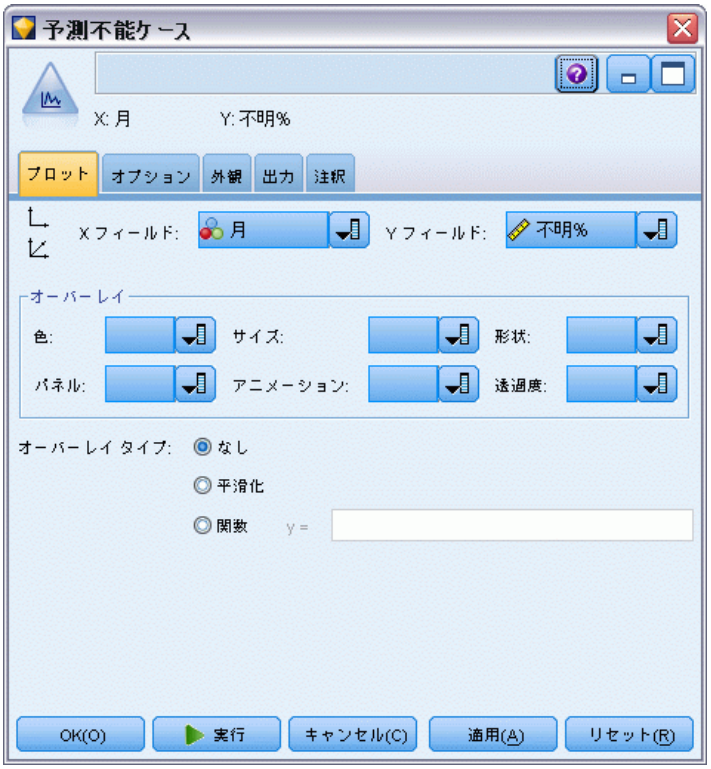
固定客となる顧客数の推定値の上限と下限がグラフに表示されます。2 本の線の間の違いはヌルとスコアリングされた顧客数です。そのため、そのステータスはかなり不確実です。時間が経てば、これらの顧客数は増加します。12 か月後にはデータセット内の元の顧客の 601 ～ 735 人が固定客であると期待されますが、24 か月後には 288 ～ 597 人に減少します。

図 26-31
フィールド作成ノード :[設定] タブ



- 固定客数の推定がどれほど不確実であるかを見る別の方法は、フィールド作成ノードをフィルタ ノードに接続する方法です。
- フィールド作成ノードの［設定］タブで、作成フィールドとして 「不明%」 と入力します。
- フィールドのデータ型として **[連続型]** を選択します。
- 式として (100 * ('Upper Estimate' - 'Lower Estimate')) / 'Lower Estimate' を入力します。不明 % は、「疑わしい」顧客数を推定値の下限のパーセンテージで表した数値です。
- [OK] をクリックします。

図 26-32
散布図ノードの [プロット] タブ

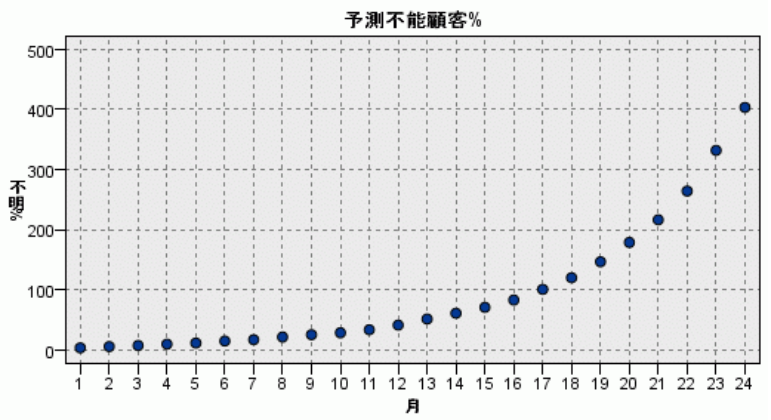


- 散布図ノードをフィールド作成ノードに接続します。
- [プロット] タブで、X フィールドとして 月 を選択し、Y フィールドとして不明 % を選択します。
- [外観] タブをクリックします。

図 26-33
散布図ノードの [外観] タブ

- 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。
- ノードを実行します。

図 26-34
予測不可能な顧客のプロット

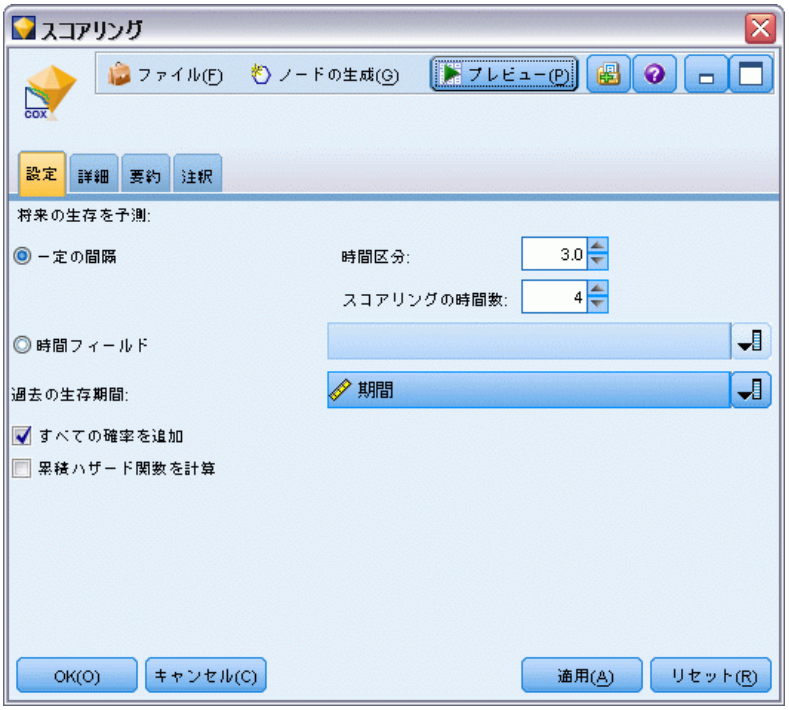


最初の 1 年間は予測不可能な顧客のパーセンテージは線形速度で増加しますが、増加率は 2 年目に急増し、23 か月までに、ヌル値を持った顧客の数が、固定客になると予測される顧客の数を上回ります。

スコアリング

モデルに満足したら、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 26-35
Cox 回帰ナゲット[設定] タブ



- 3 番目のモデル ナゲットを入力ノードに接続し、モデル ナゲットを開きます。
- [**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**」を指定し、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、各レコードは以降 4 四半期の間、スコアリングされます。
- 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。スコアリングのアルゴリズムは、各顧客の会社の顧客としての時間の長さを考慮します。
- [**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は、テーブルに表示する際のレコードのソートをより簡単にします。

図 26-36
条件抽出ノードの [設定] タブ

- 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、すでに解約している顧客が結果のテーブルから削除されます。

図 26-37
フィールド作成ノード :[設定] タブ

- 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。
- $CP-1-1 から $CP-1-4 まで、フォーム $CP-1-n のフィールドを作成するように選択し、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。
- フィールドを［**条件式** (If-Then)］として作成するよう選択します。
- 測定レベルに［**フラグ型**］を選択します。
- [If] 条件式として、「@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の閾値であることを思い出してください。
- [Then] 式として「1」を入力します。
- [Else] 式として「0」を入力します。

▶ [OK] をクリックします。

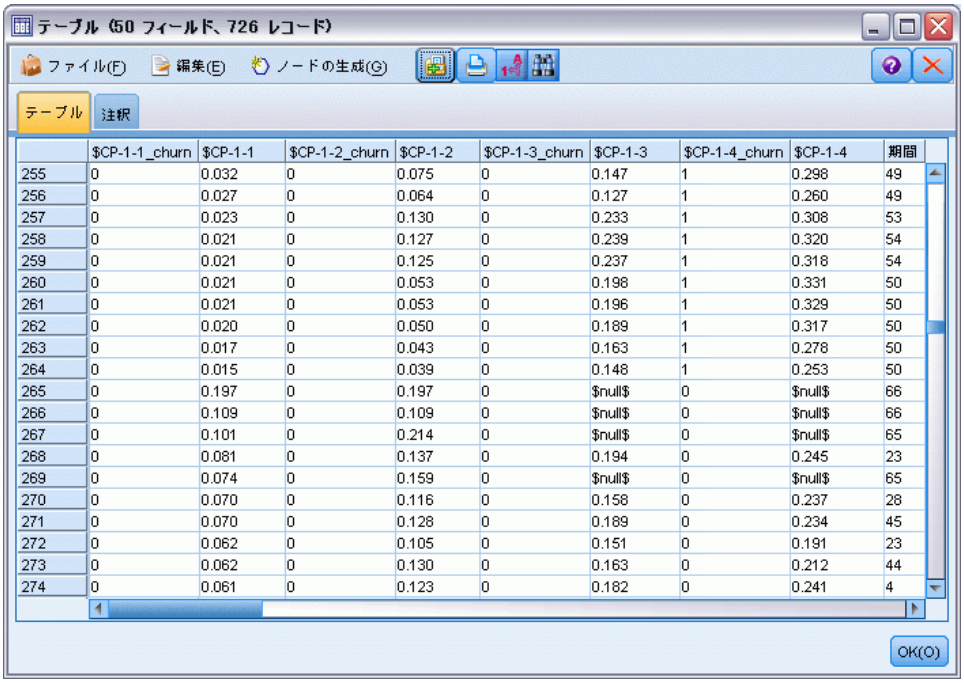
図 26-38
ソート ノードの [設定] タブ

▶ 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 までを、すべて降順でソートするように選択します。解約すると予測される顧客は上位に表示されます。

図 26-39
フィールド順序ノードの [順序] タブ

- フィールド順序ノードをソート 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するように選択します。これは、ただ結果のテーブルを見やすくするためなので、オプションです。図のように表示させるには、ボタンを使ってフィールドを移動する必要があります。

図 26-40
ユーザー設定のスコアが表示されたテーブル


テーブル (50 フィールド、726 レコード)

ファイル(F) 編集(E) ノードの生成(G)

テーブル 注釈

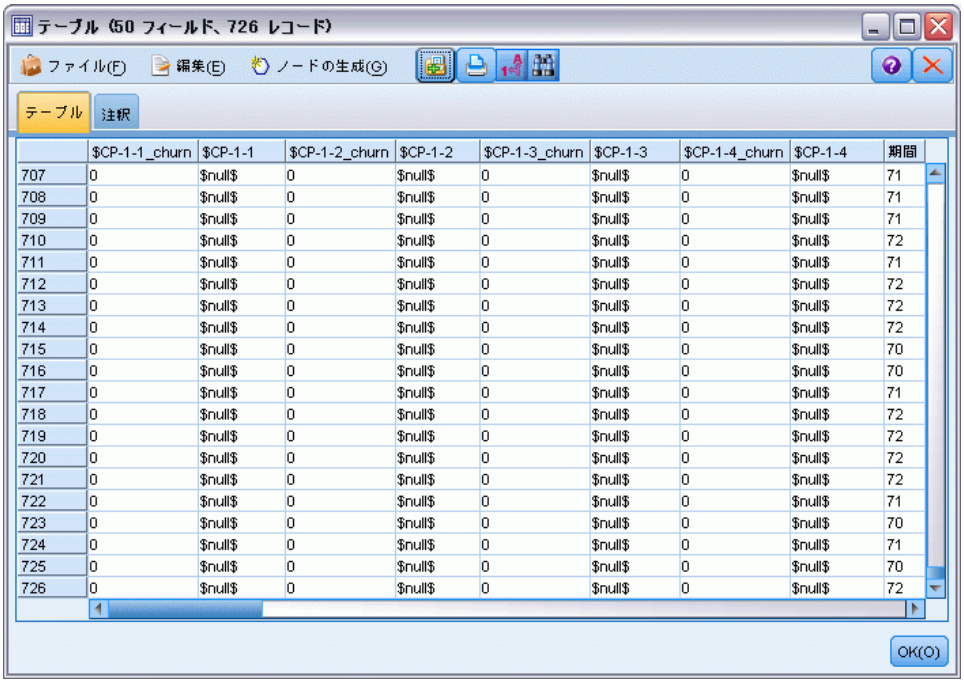
	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	期間
255	0	0.032	0	0.075	0	0.147	1	0.298	49
256	0	0.027	0	0.064	0	0.127	1	0.260	49
257	0	0.023	0	0.130	0	0.233	1	0.308	53
258	0	0.021	0	0.127	0	0.239	1	0.320	54
259	0	0.021	0	0.125	0	0.237	1	0.318	54
260	0	0.021	0	0.053	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

OK(O)

▶ テーブル ノードをフィールド順序ノードに接続して実行します。

年の終わりまでに 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。与えられた 2 人の顧客について見ると、第 1 四半期に解約の傾向がより強かった顧客が、後の四半期で解約の傾向がより強いとは限りません。たとえば、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ヶ月のハザード関数のかたちに起因するようです。たとえば、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ようです。ただし、他社に切り替えなかった場合、そのような顧客の保有期間はより長くなる可能性があります。顧客をソートし直して、解約すると考えられる顧客を別のかたちで表示することもできます。

図 26-41
ヌル値を持った顧客を示すテーブル



	$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	期間
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

テーブルの一番下に、予測値がヌル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内の生存時間の範囲を超えている顧客です。

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、来年中に解約すると予測される個々の顧客を識別しました。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かもしれないことに注意してください。理想的には、少なくとも、変数増加法を使用して得られたこのモデルを、変数減少法を使用して作成されたモデルと比較するべきです。

IBM® SPSS® Modeler で使用されているモデル作成方法の数学的な基礎の説明は、『SPSS Modeler Algorithms Guide』 に一覧されています。

章
27

マーケット バスケットの分析（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、さらに関連付けられた購入者の個人データを扱います。この個人データは、ロイヤルティ カード スキーマから取得されることもあります。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。

この例は、次のようなデータ マイニングの 2 つのフェーズを表します。

- 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 ルール算出によって、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 : このアプリケーションは予測のための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。

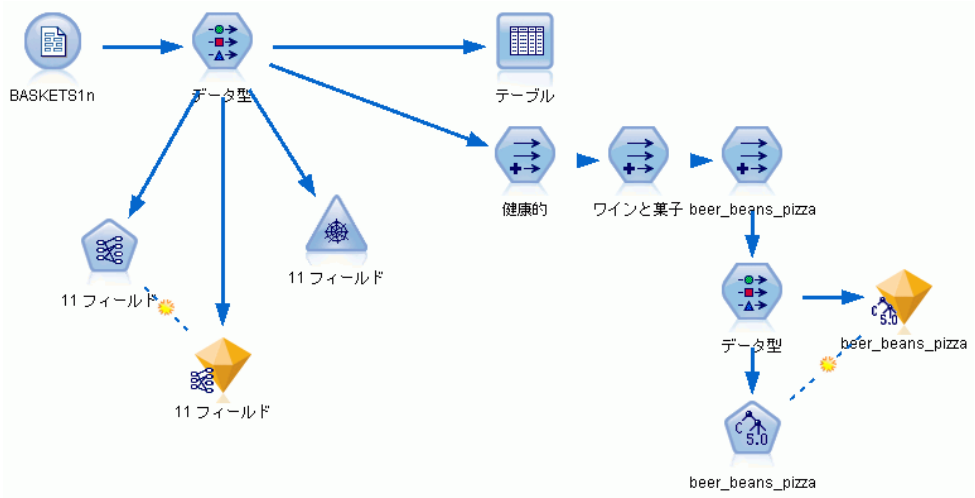
この例では、baskrule というの名前のストリームを使用します。これらのストリームは BASKETS1n という名前のデータ ファイルを参照します。これらのデータ ファイルは、IBM® SPSS® Modeler のインストール ディレクトリ中の Demos ディレクトリにあります。これは、Windows の［スタート］メニューの IBM® SPSS® Modeler プログラム グループからアクセスできます。baskrule ファイルは、streams ディレクトリにあります。

データのアクセス

可変長ノードを使って、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［カード ID］の測定レベルを［データ型不明］に設定します（各ロイヤルティ カード ID は、データセット中に 1 回しか発生せず、モデル作成には不要なため）。フィールド ［性別］ の測定レベルとして［名

義型］を選択します（Apriori モデリング アルゴリズムが ［性別］ をフラグ型として処理しないようにするため）。

図 27-1
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 のフィールドには、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：

- カード ID ：このバスケットを購入する顧客のロイヤルティ カードの ID です。
- 購入価格 ：バスケットの合計購入金額です。
- 支払方法 ：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：

- 性別
- 家の所有 ：カード所有者が家屋を所有しているかどうか。
- 収入
- 年齢

バスケットの内容 – 製品カテゴリの存在を示すフラグ：

- 果物/野菜
- 肉
- 日用雑貨

- 缶詰野菜
- 缶詰肉
- 冷凍食品
- ビール
- wine（ワイン）
- 清涼飲料
- 魚
- 菓子

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。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。データ型ノードを編集してすべての製品カテゴリの役割を［両方］に設定し、他のすべての役割を［なし］に設定してください(([両方］はそのフィールドが結果のモデルの入力と出力のどちらにもな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れば、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 27-2
モデリング用フィールドの選択

モデリング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 ノードを実行します。結果は、マネージャ 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。このモデルには、アソシエーション ルールが含まれており、コンテキスト メニューの [**ブラウズ**] を選択して表示することができます。

図 27-3
アソシエーション ルール

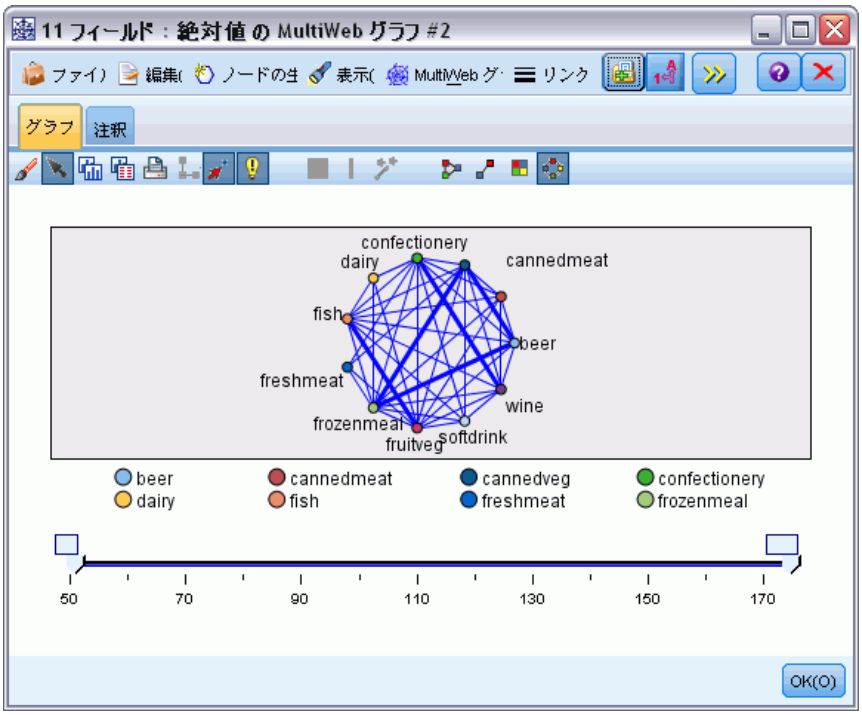
これらのルールは、冷凍食品、缶詰野菜、およびビールのさまざまな関連性を示します。次のような双方向のアソシエーション ルールが存在します。

frozenmeal -> beer
beer -> frozenmeal

これは、MultiWeb グラフ表示（双方向の連関だけを表示します）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。

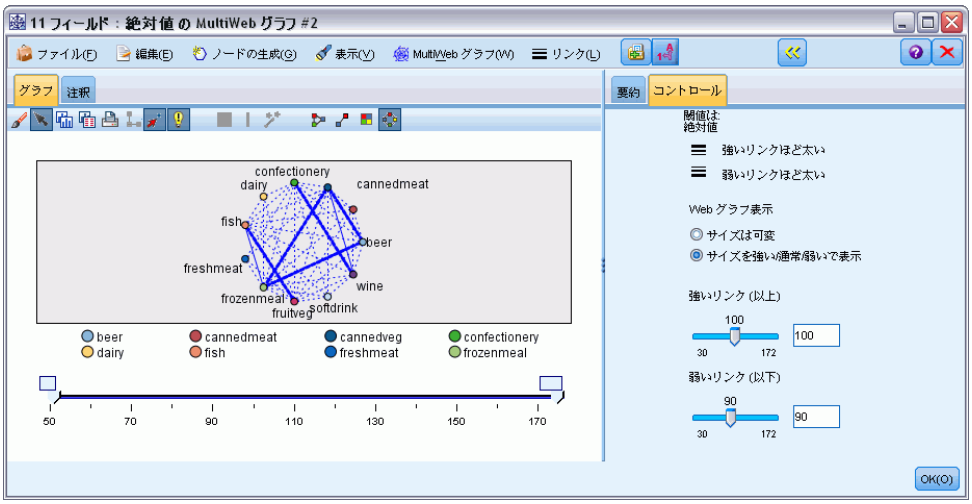
Web グラフ ノードをデータ型ノードに接続して、Web グラフ ノードを編集し、すべてのバスケットの内容フィールドを選択します。次に、[**フラグは真の値のみ**] オプションを選択して、Web グラフ ノードを実行します。

図 27-4
製品の連関の MultiWeb グラフ表示



大半の製品カテゴリ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 27-5
制限された MultiWeb グラフ表示



- ▶ 弱いリンクと強いリンクを指定するには、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 ボックスに、Web 出力の概要およびコントロールが表示されます。
- ▶ **[サイズを強い/通常/弱いで表示]** を選択します。
- ▶ 弱いリンクを 90 以上に設定します。
- ▶ 強いリンクを 100 以上に設定します。

この表示されるグラフ内では、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」と呼びます）
- 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」）

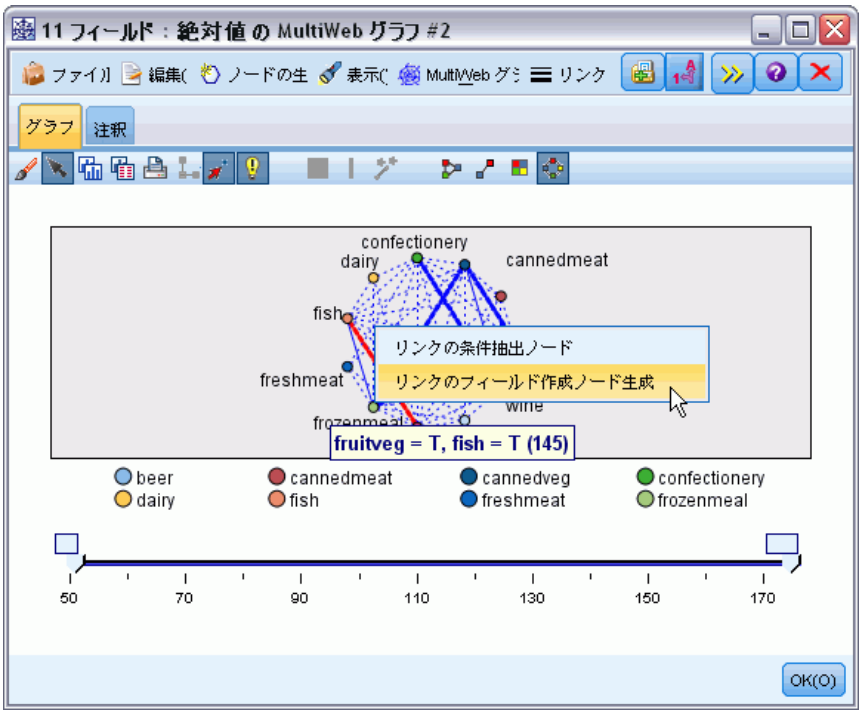
顧客グループのプロファイル作成

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ってタグを付け、さらにルール算出（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。

初めに、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 グラフを使って自動的に生成することができます。マウスの右ボタンを使用して、[果物/野菜] と [魚] の間のリンクをクリックして強調表示し、右クリックして **[リンクのフィールド作成ノードを作成]** を選択します。

図 27-6
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健康的に変更します。ワインから菓子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結果にワインと菓子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。それから、シフト キーを押しながらマウスの左ボタンをクリックすることで、[缶詰野菜]、[ビール]、および [冷凍食品] の 3 角形の中の 3 つのリンクすべてを選択します（編集モードではなくインタラクティブ モードになっていることを確認します）。それから、MultiWeb グラフ表示メニューから、次のオプションを選択します。

ノードの生成 > [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、購入価格、支払方法、性別、家の所有、年収、および年齢のフィールドの役割を［入力］に設定し、関連する顧客のグループ（たとえばbeer_beans_pizza）のフィールドの役割を［対象］に設定します。それ以外のすべてのフィールドの役割を［なし］に設定します。C5.0 ノードを接続し、［出力形式］を **［ルールセット］** に設定して、そのノードを実行します。結果のモデル（ビール_豆_ピザ用）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
```

このコンテキスト内で C5.0 の代わりに GRI を使用すると、より広い範囲の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こともできます。

要約

この例は、IBM® SPSS® Modeler を使用して、モデル作成（Apriori を使用）とビジュアル化（MultiWeb 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、たとえばこのように顧客をグループ分けし、特別セールの対象にしてダイレクト メールの返信率を向上させたり、地域的な顧客ベースの需要に合わせて支店の製品在庫の範囲を調整したりできます。

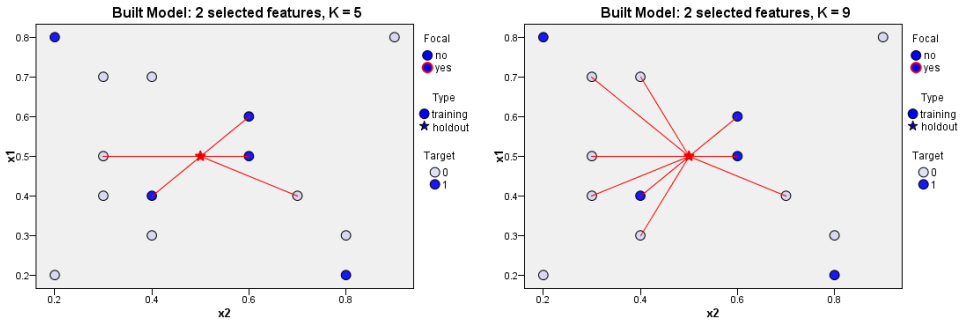
章
28
新しい自動車製品の評価 (KNN)

最近隣分析は、そのほかのケースに対する類似性に基づいてケースを分類する方法です。マシン学習で、保存されたパターン、またはケースへに完全に一致する必要なくデータのパターンを認識する方法として開発されました。同様のケースはお互いに近く、異なるケースはお互いに離れています。つまり、2 つのケース間の距離は、それらの非類似度の尺度です。

お互いに近いケースは「近隣」と呼ばれます。新しいケース（ホールドアウト）が表示されている場合、モデルのケースからの距離が計算されます。最も類似した分類「最近隣」が集計され、新しいケースが、最大数の最近隣を含むカテゴリに投入されます。

検証する最近隣の数を指定できます。この値は k となります。図は、新しいケースが 2 つの異なる値の k を使用してどのように分類されるかを示します。k = 5 の場合、最近隣の大部分はカテゴリ 1 に属するため、新しいケースはカテゴリ 1 にあります。ただし k = 9 の場合、最近隣の大部分はカテゴリ 0 に属するため、新しいケースはカテゴリ 0 にあります。

図 28-1
分類時に k を変更した場合の効果



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、最近隣の平均または中央の対象値を使用して、新しいケースの予測値を取得します。

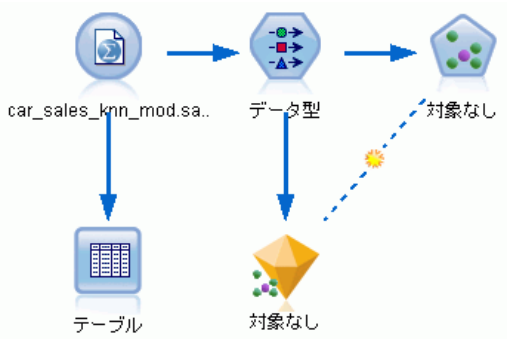
自動車メーカーが、2 つの新しい自動車、乗用車およびトラックのプロトタイプを開発しています。新しいモデルを導入する前に、メーカーは市場にある既存の自動車でどれが最もプロトタイプに近いのか、つまりどの自動車が「最近隣」なのか、そしてどのモデルが競争相手となるのかを判断する必要があります。

メーカーは様々なカテゴリ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、Classification_Module サブフォルダの下の Demos フォルダ内にある streams という名前のストリームを使用します。データ ファイルは、car_sales_knn_mod.sav です。 詳細は、 1 章 Demos フォルダ in IBM SPSS Modeler 15 ユーザー ガイド を参照してください。

ストリームの作成

図 28-2
KNN モデル作成のサンプル ストリーム

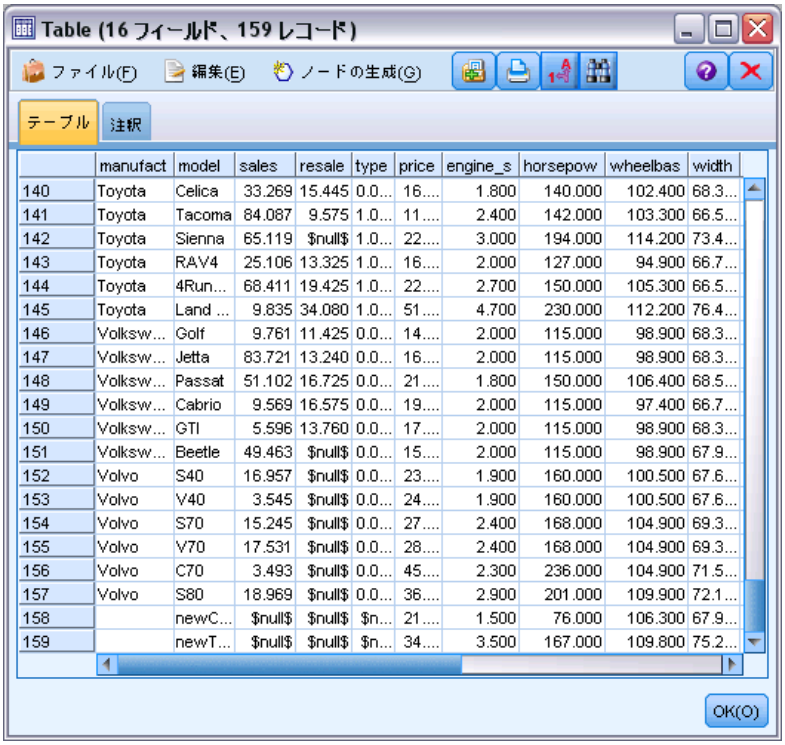


新規のストリームを作成し、IBM® SPSS® Modeler インストールの Demos フォルダにある car_sales_knn_mod.sav を示す Statistics ファイル入力ノードを追加します。

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。

- テーブル ノードを Statistics ファイル入力ノードに接続します。
- テーブルノードを開いて、[実行] をクリックします。

図 28-3
乗用車およびトラックのソース データ



	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...

2 つのプロトタイプの詳細、newCar および newTruck がファイルの最後に追加されます。

メーカーが「トラック」の分類を使用しているソース データ（種類 列の 1 の値）が、漠然と自動車以外の種類の車両を意味していることがわかります。

最近隣を特定する場合、2 つのプロトタイプをホールドアウトとして指定できるようにするには、最後の列のデータ区分が必要です。このように、これらのデータが、検討に入れる市場の残りの部分であるため、計算には影響は与えません。2 つのホールドアウト レコードのデータ区分の値を 1 に設定し、このフィールドの他のすべてのレコードは 0 に設定すると、最近隣を計算するケースである中心レコードを設定する場合に後でこのフィールドを設定できます。

後で参照するため、テーブル出力ウィンドウは開いたままにします。

図 28-4
データ型ノードの設定

- タイプ ノードをストリームに追加します。
- データ型ノードを Statistics ファイル入力ノードに接続します。
- データ型ノードを開きます。

  フィールド mpg から price のみを比較するため、これらのフィールドすべての役割を［**入力**］に設定します。
- その他すべてのフィールド（manufact から type、および lnsales）の役割を［**なし**］に設定します。
- 最後のフィールド、データ区分の測定レベルを［**フラグ型**］に設定します。役割をかならず［**入力**］に設定してください。
- ［**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。
- ［OK］をクリックします。

図 28-5
最近隣の特定を選択

- KNN ノードをデータ型ノードに接続します。
- 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であるため、今回は対象フィールドは予測しません。

- [目的] タブで、[最近隣のみを識別] を選択します。
- [設定] タブをクリックします。

図 28-6
データ区分フィールドを使用して重要レコードを特定

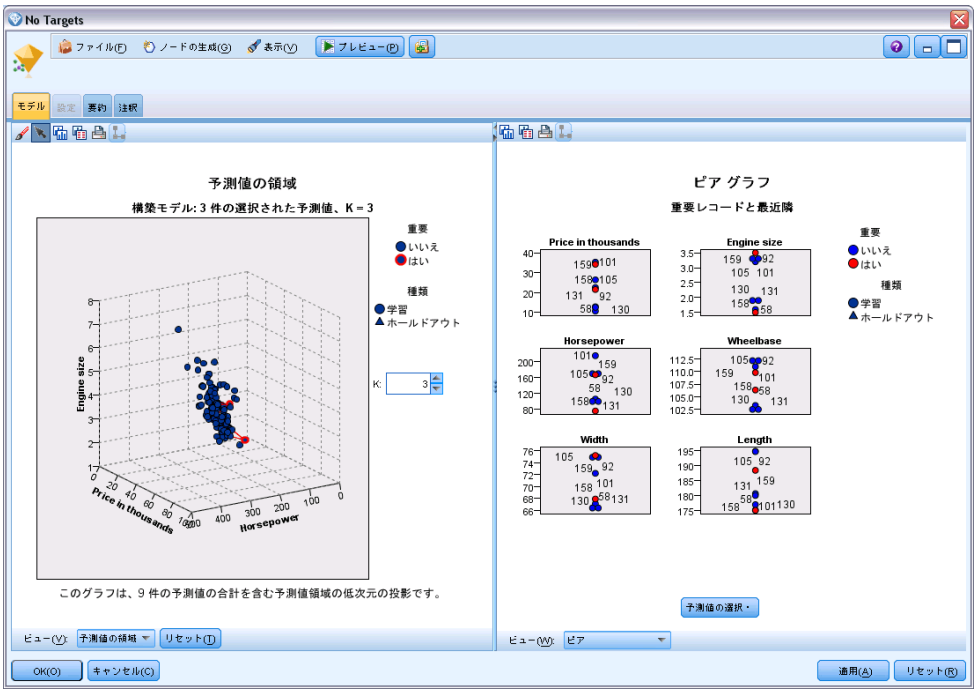
[データ区分] フィールドを使用して、重要レコード（最近隣を特定するレコード）を特定できます。フラグ型フィールドを使用して、このフィールドの値が 1 に設定されたレコードが重要レコードとなっていることを確認します。

このフィールドの値が 1 に設定されているレコードは newCar および newTruck となっているため、これらが重要レコードとなります。

- [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 ボックスをオンにします。
- 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。
- [実行] ボタンをクリックします。

出力の検証

図 28-7
モデル ビューア 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ます。いずれかのナゲットを開くとモデル ビューアが表示されます。モデル ビューアには次の 2 つのウィンドウがあります。

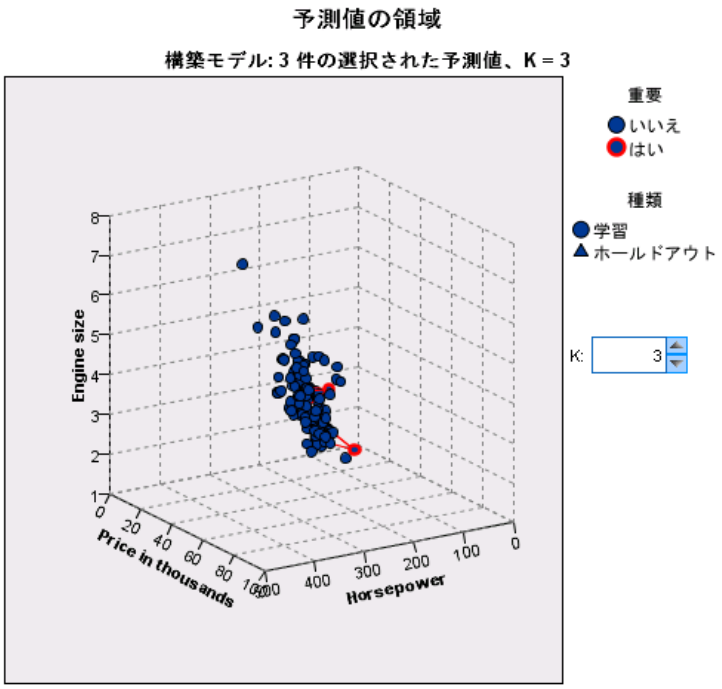
- 1 つめのパネルはメイン ビューと呼ばれ、モデルの概要が表示されます。最近隣モデルのメイン ビューは、**予測領域**と呼ばれます。
- 2 つめのパネルには、次の 2 種類のビューのいずれかが表示されます。

  モデルの詳細を表示するが、モデル自体に焦点を当てていない補助的モデル ビュー。

  メイン ビューの一部について掘り下げた場合、モデルのある特徴についての詳細を示すリンク ビュー。

予測領域

図 28-8
予測領域のグラフ



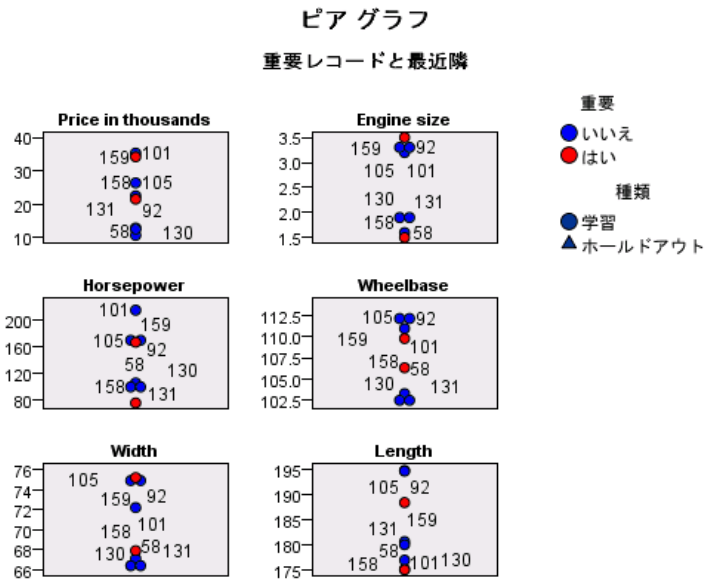
予測領域のグラフは、3 つの特徴（実際はソース データの最初の 3 つの入力フィールド）のデータ ポイントをプロットするインタラクティブな 3 次元のグラフで、価格、エンジンのサイズ、馬力を示します。

2 つの中心レコードは赤く強調表示され、それらと k の最近隣を線で繋ぎます。

グラフをクリック アンド ドラッグし、予測領域のポイントの分布をよりよく表示するために、グラフを回転させることができます。デフォルトの表示に戻すには、[**リセット**] をクリックします。

ピアグラフ

図 28-9
ピアグラフ



デフォルトの補助ビューはピアグラフです。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予測領域で選択した 2 つの中心レコードと k の最近隣が強調表示されます。

自動車は、入力データのレコード番号で表示されます。ここで、自動車を特定するためにテーブル ノードの出力が必要です。

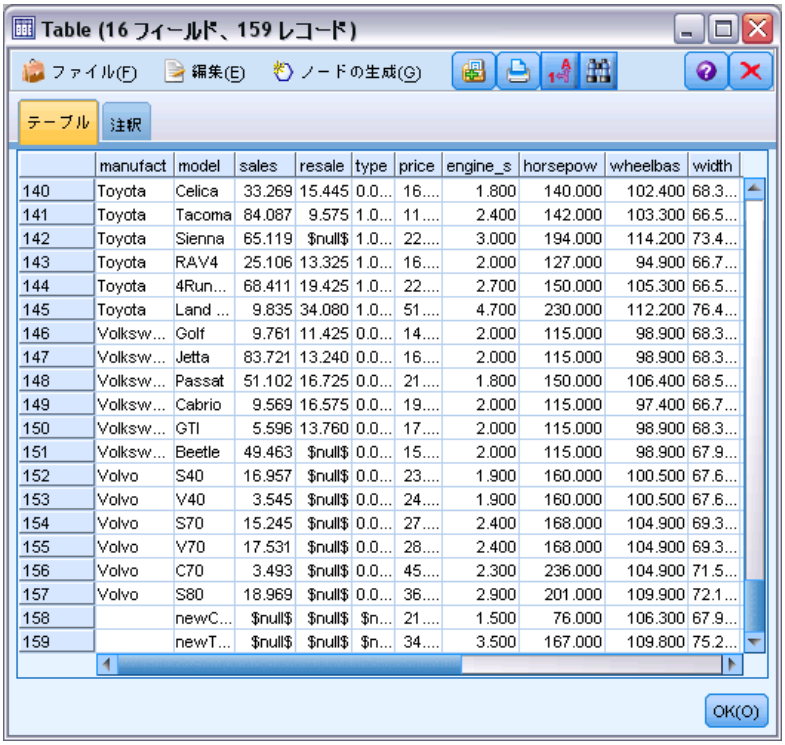
テーブル ノード出力を使用できる場合、次の手順を実行します。

- IBM® SPSS® Modeler メイン ウィンドウの右上にあるマネージャ パネルの [出力] タブをクリックします。
- 投入 [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い場合、次の手順を実行します。

- SPSS Modeler のメイン ウィンドウで、テーブル ノードを開きます。
- [実行] をクリックします。

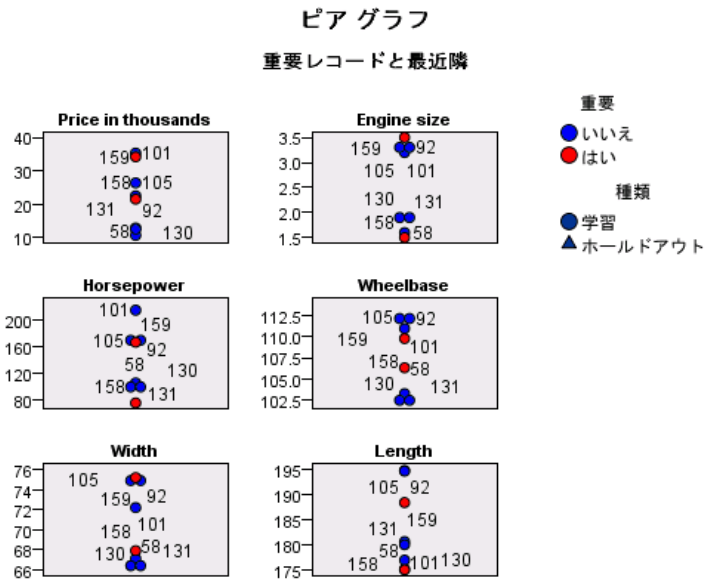
図 28-10
レコード番号によるレコードの特定



	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...

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わかります。

図 28-11
ペアグラフの特徴の比較



ペアグラフで、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わかります。

6 つの特徴それぞれについて、特定の点の上でマウスを移動し、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。

newCar および newTruck のどの自動車が最近隣なのでしょうか？

ペアグラフは、若干混雑して見づらいため、より単純な表示に変更しましょう。

▶ ペアグラフの一番下の [**ビュー**] ドロップダウン リスト（[**ペア**] と表示されている投入）をクリックします。

▶ [**隣接および距離の表**] を選択します。

近隣および距離の表

図 28-12
近隣および距離の表

k 個の最近隣と距離

初期の重要レコードに表示

重要レコード	最近隣			最短距離		
	1	2	3	1	2	3
158	131	130	58	0.979	0.990	1.011
159	105	92	101	0.580	0.634	0.644

2 つのプロトタイプが市場で最も近い 3 つのモデルが表示されます。

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。

3 つの車すべてが中型のサルーン カーが特に燃料効率に優れているため、newCar に収まっているのは驚くべきことではありません。

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。

従来の意味では必ずしもトラックではありませんが、自動車と分類されていない車両となります。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、またこの種類の最も重い車両となります。ただし、燃料効率は最も近い競合他社の製品より優れており、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。

付録

A

注意事項

この情報は、世界各国で提供される製品およびサービス向けに作成されています。

IBM はこのドキュメントで説明する製品、サービス、機能は他の国では提供していない場合があります。現在お住まいの地域で利用可能な製品、サービス、および、情報については、お近くの IBM の担当者にお問い合わせください。IBM 製品、プログラム、またはサービスに対する参照は、IBM 製品、プログラム、またはサービスのみが使用することができることを説明したり意味するものではありません。IBM の知的所有権を侵害しない機能的に同等の製品、プログラム、またはサービスを代わりに使用することができます。ただし、IBM 以外の製品、プログラム、またはサービスの動作を評価および確認するのはユーザーの責任によるものです。

IBMは、本ドキュメントに記載されている内容に関し、特許または特許出願中の可能性があります。本ドキュメントの提供によって、これらの特許に関するいかなる権利も使用者に付与するものではありません。ライセンスのお問い合わせは、書面にて、下記住所に送ることができます。

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

2 バイト文字セット (DBCS) 情報についてのライセンスに関するお問い合わせは、お住まいの国の IBM Intellectual Property Department に連絡するか、書面にて下記宛先にお送りください。

神奈川県大和市下鶴間1623番14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

以下の条項はは、イギリスまたはこのような条項が法律に反する他の国では適用されません。 International Business Machines は、明示的または黙示的に関わらず、第三者の権利の侵害しない、商品性または特定の目的に対する適合性の暗黙の保証を含むがこれに限定されない、いかなる保証なく、本出版物を「そのまま」提供します一部の州では、特定の取引の明示的または暗示的な保証の免責を許可していないため、この文が適用されない場合があります。

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。情報については変更が定期的に行われます。これらの変更は本書の新版に追加されます。IBM は、本書に記載されている製品およびプログラムについて、事前の告知なくいつでも改善および変更を行う場合があります。

IBM 以外の Web サイトに対するこの情報内のすべての参照は、便宜上提供されているものであり、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。これらの Web サイトの資料はこの IBM 製品の資料に含まれるものではなく、これらの Web サイトの使用はお客様の責任によるものとします。

IBM はお客様に対する一切の義務を負うことなく、自ら適切と考える方法で、情報を使用また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取得者が (i) 別途作成されたプログラムと他のプログラム (本プログラムを含む) との間の情報交換および (ii) 交換された情報の相互利用を目的とした本プログラムに関する情報の所有を希望する場合、下記住所にお問い合わせください。

IBM Software Group, Attention:Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

上記のような情報は、該当する条項および条件に従い、有料で利用できるものとします。

本ドキュメントに記載されている許可されたプログラムおよびそのプログラムに使用できるすべてのライセンス認証された資料は、IBM Customer Agreement、IBM International Program License Agreement、および当社とかわした同等の契約の条件に基づき、IBM によって提供されます。

ここに記載されているパフォーマンスデータは、すべて管理環境下で確認されたものです。そのため、他の操作環境で得られた結果は大きく異なる可能性があります。開発レベルのシステムで測定が行われている場合があり、これらの測定値は一般に利用可能なシステムと同じであることを保証するものではありません。また、測定値が推定値である可能性があり、実際の結果は異なる場合があります。本ドキュメントのユーザーは、特定の環境に適したデータを検証する必要があります。

IBM 以外の製品に関する情報は、それらの製品の供給業者、公開済みの発表、または公開で使用できるソースから取得しています。IBM は、それらの製品のテストは行っておらず、IBM 以外の製品に関連する性能、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業者に通知する必要があります。

IBM の将来の方向性または意向に関する記述については、予告なく変更または取り消すことがあり、目的や目標のみを示すものです。

この情報には、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。できる限り詳細に説明するため、例には、個人、企業、ブランド、製品などの名前が使用されています。これらの名称はすべて架空のものであり、実際の企業で使用される名称および住所とは一切関係ありません。

この情報をソフトコピーでご覧になっている場合は、写真やカラーのイラストが表示されない場合があります。

商標

IBM、IBM ロゴ、および ibm.com、SPSS は、世界の多くの国で登録された IBM Corporation の商標です。IBM の商標の現在のリストは、http://www.ibm.com/legal/copytrade.shtml を参照してください。

Intel、Intel のロゴ、Intel Inside、Intel Inside のロゴ、Intel Centrino、Intel Centrino のロゴ、Celeron、Intel Xeon、Intel SpeedStep、Itanium、および Pentium は、米国およびその他の国の Intel Corporation または関連会社の商標または登録商標です。

Linux は、米国およびその他の国における Linus Torvalds の登録商標です。

Microsoft、Windows、Windows NT、および Windows のロゴは、米国およびその他の国における Microsoft 社の商標です。

UNIX は、米国およびその他の国における The Open Group の登録商標です。

Java およびすべての Java ベースの商標およびロゴは、米国およびその他の国の Sun Microsystems, Inc. の商標です。

その他の製品名およびサービス名等は、IBM または他の会社の商標です。

参考文献

アスンシオン, A., および D. ニューマン. 2007. "UCI マシン学習リポジトリ." Available at http://mlearn.ics.uci.edu/MLRepository.html.

索引