IBM SPSS Modeler 18.0 — podręcznik zastosowań

Uwaga

Przed skorzystaniem z niniejszych informacji oraz produktu, którego one dotyczą, należy zapoznać się z informacjami zamieszczonymi w sekcji “Uwagi” na stronie 355.

Informacje o produkcie

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

Spis treści

Rozdział 1. O programie IBM SPSS Modeler

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

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

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

Produkty IBM SPSS Modeler

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

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

IBM SPSS Modeler

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

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

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

IBM SPSS Modeler Server

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

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

IBM SPSS Modeler Administration Console

Oprogramowanie Modeler Administration Console to aplikacja graficzna służąca do obsługi wielu opcji konfiguracji SPSS Modeler Server, które można dostosować również za pomocą pliku opcji. W ramach aplikacji dostępny jest interfejs konsoli użytkownika pozwalający na monitorowanie i konfigurowanie instalacji SPSS Modeler Server. Interfejs jest dostępny bez dodatkowych opłat dla aktualnych użytkowników SPSS Modeler Server. Aplikację można zainstalować tylko na komputerach z systemem Windows, jednak administrować można serwerem zainstalowanym na dowolnej obsługiwanej platformie.

IBM SPSS Modeler Batch

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

IBM SPSS Modeler Solution Publisher

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

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

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

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

Wydania programu IBM SPSS Modeler

Dostępne są następujące wydania oprogramowania SPSS Modeler.

SPSS Modeler Professional

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

SPSS Modeler Premium

Oprogramowanie SPSS Modeler Premium wymaga oddzielnej licencji. Dzięki temu rozwiązaniu oprogramowanie SPSS Modeler Professional może obsługiwać wyspecjalizowane dane, np. używane z technikami Entity Analytics lub analizami sieci społecznościowych, oraz nieustrukturyzowane dane tekstowe. Oprogramowanie SPSS Modeler Premium składa się z następujących komponentów.

IBM SPSS Modeler Entity Analytics dodaje dodatkowy wymiar do analiz predykcyjnych wykonywanych przy użyciu produktu IBM SPSS Modeler. Analizy predykcyjne stanowią próbę przewidzenia przyszłych zachowań na podstawie danych z przeszłości, natomiast techniki Entity Analytics służą przede wszystkim poprawie spójności danych bieżących poprzez rozwiązywanie konfliktów dotyczących tożsamości w samych rekordach. Tożsamość może dotyczyć osoby, organizacji, obiektu lub dowolnej innej jednostki, względem której może istnieć niejednoznaczność. Ustalenie tożsamości może być istotne w przypadku wielu dziedzin, takich jak zarządzanie relacjami z klientami, wykrywanie oszustw, pranie pieniędzy oraz bezpieczeństwo w skali krajowej i międzynarodowej.

Oprogramowanie **IBM SPSS Modeler Social Network Analysis** przetwarza informacje o relacjach w zmienne, które charakteryzują zachowania społeczne pojedynczych osób i grup. Korzystając z danych opisujących relacje w sieciach społecznościowych, oprogramowanie IBM SPSS Modeler Social Network Analysis identyfikuje liderów społecznych, którzy wpływają na zachowanie innych w sieci. Ponadto umożliwia ustalenie ludzi, którzy są pod największym wpływem innych uczestników sieci. Łącząc te wyniki z innymi środkami, można stworzyć obszerne profile osób, na których będą bazowały modele predykcyjne użytkownika. Modele zawierające te informacje społeczne będą dawały lepsze wyniki niż modele, które ich nie zawierają.

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

Dokumentacja IBM SPSS Modeler

Dokumentacja elektroniczna jest dostępna w programie SPSS Modeler z poziomu menu Pomoc. Zasoby te obejmują informacje na temat oprogramowania SPSS Modeler, SPSS Modeler Server i Podręcznik zastosowań (nazywany również Samouczkiem) oraz inne materiały pomocnicze.

Pełna dokumentacja dla każdego produktu (obejmująca instrukcje instalacji) jest dostępna w formacie PDF w osobnym skompresowanym folderze jako część pobieranego produktu. Dokumenty PDF można również pobrać z Internetu pod adresem http://www.ibm.com/support/docview.wss?uid=swg27046871.

Dokumentacja w obydwu formatach jest również dostępna w Centrum Wiedzy SPSS Modeler pod adresem http://www-01.ibm.com/support/knowledgecenter/SS3RA7_18.1.0.

Dokumentacja SPSS Modeler Professional

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

- **IBM SPSS Modeler — podręcznik użytkownika.** Ogólne wprowadzenie do obsługi oprogramowania SPSS Modeler, obejmuje opisy procedur tworzenia strumieni danych, obsługi braków danych, tworzenia wyrażeń CLEM, pracy z projektami i raportami, dane na temat przygotowywania strumieni z przeznaczeniem do wdrażania w produkcie IBM SPSS Collaboration and Deployment Services lub IBM SPSS Modeler Advantage.
- **IBM SPSS Modeler — węzły źródłowe, procesowe i wyników.** Opisy wszystkich węzłów używanych do odczytywania, przetwarzania i tworzenia wynikowych postaci danych w różnych formatach. Czyli wszystkich węzłów poza węzłami modelowania.

- **IBM SPSS Modeler — węzły modelowania.** Opisy wszystkich węzłów używanych do tworzenia modeli eksploracji danych. Oprogramowanie IBM SPSS Modeler umożliwia korzystanie z wielu metod modelowania opartych na sztucznej inteligencji, uczeniu maszynowym i statystykach.
- **IBM SPSS Modeler Algorithms Guide.** Opisy podstaw matematycznych dotyczących metod modelowania stosowanych w oprogramowaniu IBM SPSS Modeler. Ten podręcznik jest dostępny tylko w formacie PDF.
- **IBM SPSS Modeler — podręcznik zastosowań.** Przykłady zawarte w niniejszym przewodniku stanowią skrócone informacje związane z konkretnymi metodami i technikami modelowania. Wersja online tego podręcznika jest również dostępna z poziomu menu Pomoc. Więcej informacji można znaleźć w temacie “Przykłady aplikacji”.
- **IBM SPSS Modeler — podręcznik tworzenia skryptów w języku Python i automatyzacji.** Informacje na temat automatyzacji działania systemu za pomocą skryptów Python wraz z właściwościami służącymi do obsługi węzłów i strumieni.
- **IBM SPSS Modeler — podręcznik wdrażania.** Informacje na temat uruchamiania strumieni IBM SPSS Modeler w postaci krokowych operacji wykonywanych podczas przetwarzania zadań w oprogramowaniu IBM SPSS Collaboration and Deployment Services Deployment Manager.
- **IBM SPSS Modeler CLEF Developer's Guide.** Z oprogramowaniem CLEF można zintegrować inne programy pozwalające na przetwarzanie danych lub obsługę algorytmów modelujących w postaci węzłów w IBM SPSS Modeler.
- **IBM SPSS Modeler — podręcznik eksploracji w bazie danych.** Informacje na temat wydajnego wykorzystywania bazy danych w celu zwiększenia wydajności i zakresu funkcji analitycznych za pomocą algorytmów innych firm.
- **IBM SPSS Modeler Server — podręcznik administracji i wydajności.** Informacje na temat konfiguracji i funkcji administracyjnych w oprogramowaniu IBM SPSS Modeler Server.
- **IBM SPSS Modeler Administration Console — podręcznik użytkownika.** Informacje na temat instalowania i używania interfejsu użytkownika konsoli w celu monitorowania i konfigurowania oprogramowania IBM SPSS Modeler Server. Konsola działa jako wtyczka do aplikacji Deployment Manager.
- **IBM SPSS Modeler CRISP-DM — podręcznik.** Szczegółowy podręcznik metodologii CRISP-DM w kontekście eksploracji danych za pomocą oprogramowania SPSS Modeler.
- **IBM SPSS Modeler Batch — podręcznik użytkownika.** Pełny podręcznik obsługi oprogramowania IBM SPSS Modeler w trybie wsadowym obejmujący szczegółowe informacje na temat pracy w trybie wsadowym i korzystania z argumentów z poziomu wiersza komend. Ten podręcznik jest dostępny tylko w formacie PDF.

Dokumentacja SPSS Modeler Premium

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

- **IBM SPSS Modeler Entity Analytics — podręcznik użytkownika.** Informacje dotyczące używania techniki Entity Analytics w oprogramowaniu SPSS Modeler obejmują procedury dotyczące instalacji i konfiguracji repozytorium, węzłów technik Entity Analytics oraz zadań administracyjnych.
- **IBM SPSS Modeler Social Network Analysis — podręcznik użytkownika.** Podręcznik zawierający informacje na temat analizy sieci społecznościowych za pomocą oprogramowania SPSS Modeler, w tym informacje o analizie grup i analizie przenikania.
- **SPSS Modeler Text Analytics — podręcznik użytkownika.** Informacje na temat używania analiz tekstu za pomocą oprogramowania SPSS Modeler, obejmują procedury dotyczące węzłów eksploracji tekstu, interaktywnego pulpitu roboczego, szablonów oraz innych zasobów.

Przykłady aplikacji

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

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

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

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

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

Folder Demos

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

Monitorowanie wykorzystania licencji

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

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

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

Rozdział 2. Przegląd produktu IBM SPSS Modeler

Pierwsze kroki

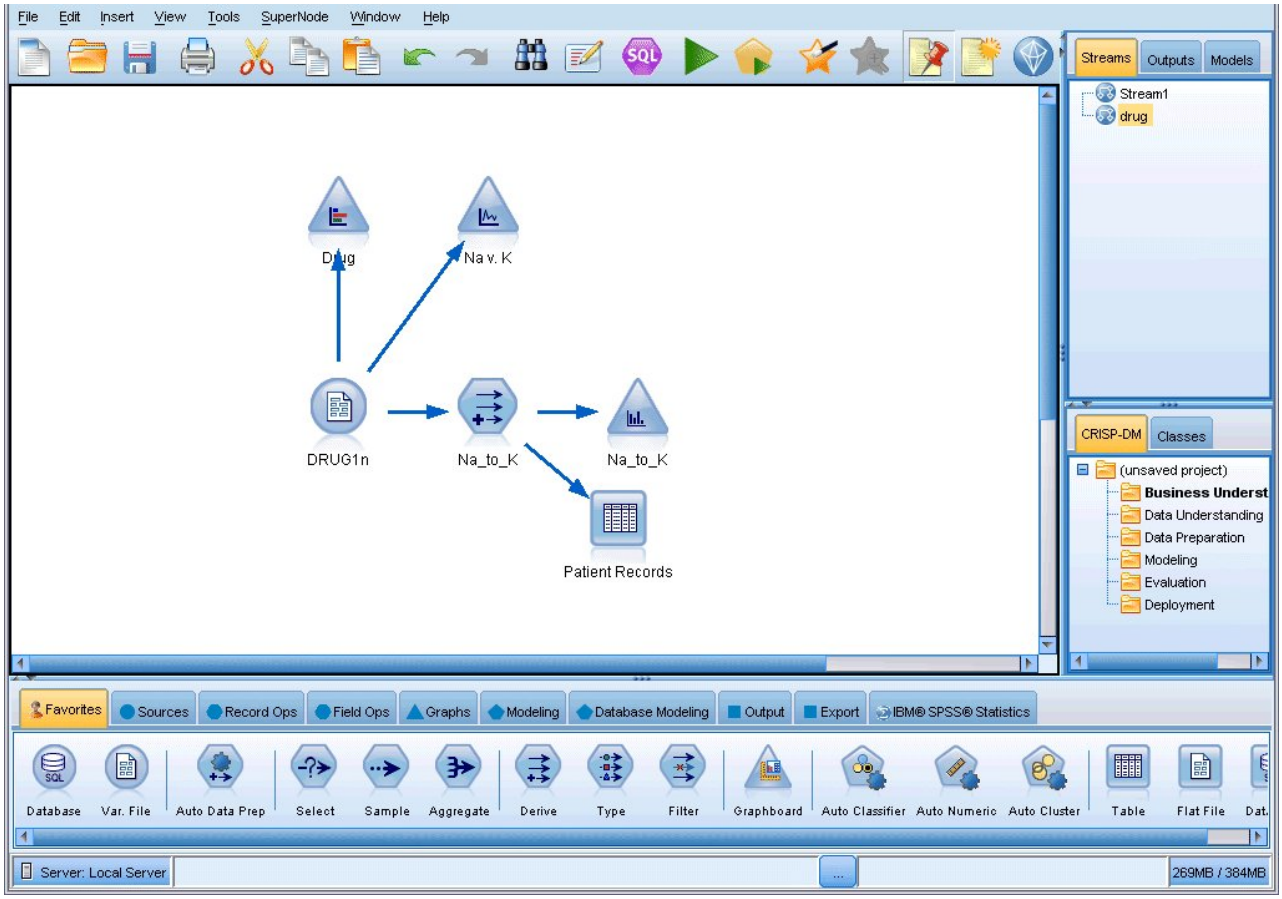
Będąc narzędziem do eksploracji danych, oprogramowanie IBM SPSS Modeler umożliwia uzyskanie użytecznych informacji o relacjach istniejących w dużych zbiorach danych. Inaczej niż w tradycyjnych metodach statystycznych użytkownik na początku nie musi wiedzieć, czego szuka. Dane można przeglądać, dopasowując różne modele, i badać różnorodne relacje do momentu znalezienia użytecznych informacji.

Uruchamianie IBM SPSS Modeler

Aby uruchomić aplikację, kliknij:

Start > [Wszystkie] Programy > IBM SPSS Modeler 18.0 > IBM SPSS Modeler 18.0

Okno główne zostanie wyświetlone po kilku sekundach.



Rysunek 1. Okno główne aplikacji IBM SPSS Modeler

Uruchamianie z wiersza komend

Można użyć wiersza komend systemu operacyjnego w celu uruchomienia programu IBM SPSS Modeler w następujący sposób:

1. Na komputerze, na którym zainstalowano program IBM SPSS Modeler, otwórz okno DOS lub okno wiersza komend.
2. Aby uruchomić interfejs IBM SPSS Modeler w trybie interaktywnym, wpisz komendę modelerclient wraz z wymaganymi argumentami, na przykład:

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

Dostępne argumenty (flagi) umożliwiają nawiązanie połączenia z serwerem, załadowanie strumieni, uruchomienie skryptów lub określenie innych parametrów, stosownie do potrzeb.

Łączenie się z serwerem IBM SPSS Modeler Server

Serwer IBM SPSS Modeler można uruchamiać jako niezależną aplikację, jako klienta połączonego bezpośrednio z serwerem IBM SPSS Modeler Server lub jako klienta połączonego z serwerem IBM SPSS Modeler Server albo z klastrem serwerów za pośrednictwem wtyczki Coordinator of Processes programu IBM SPSS Collaboration and Deployment Services. Bieżący status połączenia jest wyświetlany w lewym dolnym rogu okna programu IBM SPSS Modeler.

Przy każdej próbie połączenia się z serwerem można ręcznie wprowadzić nazwę serwera, z którym ma zostać nawiązane połączenie, lub wybrać nazwę zdefiniowaną wcześniej. Jeśli jednak zainstalowany jest program IBM SPSS Collaboration and Deployment Services, możliwe jest przeszukiwanie listy serwerów lub klastrów serwerów z okna dialogowego Logowanie do serwera. Możliwość przeglądania usług Statistics działających w sieci jest dostępna za pośrednictwem usługi Coordinator of Processes.

Aby połączyć się z serwerem

1. W menu Narzędzia kliknij opcję **Logowanie do serwera**. Zostanie otwarte okno dialogowe Logowanie do serwera. Można też dwukrotnie kliknąć obszar statusu połączenia w oknie IBM SPSS Modeler.
2. Korzystając z okna dialogowego, podaj opcje umożliwiające nawiązanie połączenia z komputerem serwera lokalnego lub wybierz połączenie z tabeli.
   - Kliknij przycisk **Dodaj** lub **Edytuj**, aby dodać lub edytować połączenie. Więcej informacji można znaleźć w temacie “Dodawanie i edytowanie połączenia z serwerem IBM SPSS Modeler Server” na stronie 9.
   - Kliknij opcję **Wyszukaj**, aby uzyskać dostęp do serwera lub klastra serwerów za pośrednictwem usługi Coordinator of Processes. Więcej informacji można znaleźć w temacie “Wyszukiwanie serwerów w programie IBM SPSS Collaboration and Deployment Services” na stronie 9.

   Tabela serwera. W tabeli tej znajduje się zestaw zdefiniowanych połączeń z serwerem. W tabeli wyświetlane są połączenie domyślne, nazwa serwera, opis i numer portu. Istnieje możliwość ręcznego dodania nowego połączenia, a także wyboru lub wyszukania istniejącego połączenia. Aby ustawić konkretny serwer jako połączenie domyślne, zaznacz pole wyboru w kolumnie Domyślne w tabeli.

   Domyślna ścieżka danych. Wskaż ścieżkę używaną przez dane na komputerze serwera. Kliknij przycisk z wielokropkiem **(...)**, aby przejść do wymaganej lokalizacji.

   Ustaw dane uwierzytelniające. To pole wyboru powinno być niezaznaczone, aby możliwe było włączenie funkcji **pojedynczego uwierzytelniania**, która próbuje zalogować użytkownika na serwerze przy użyciu nazwy użytkownika i hasła na lokalnym komputerze. Jeśli pojedyncze logowanie nie jest możliwe lub w przypadku zaznaczenia tego pola w celu wyłączenia pojedynczego logowania (na przykład w celu zalogowania się na konto administratora), uaktywniane są poniższe pola umożliwiające wprowadzenie danych uwierzytelniających.

   Identyfikator użytkownika. Wprowadź nazwę użytkownika, z użyciem której należy się logować do serwera.

   Hasło. Wprowadź hasło powiązane z określoną nazwą użytkownika.

   Domena. Określ domenę używaną do zalogowania się na serwerze. Nazwa domeny jest wymagana tylko wówczas, gdy komputer serwera znajduje się w innej domenie Windows niż komputer kliencki.
3. Kliknij przycisk **OK**, aby zakończyć nawiązywanie połączenia.

Aby odłączyć się od serwera

1. W menu Narzędzia kliknij opcję **Logowanie do serwera**. Zostanie otwarte okno dialogowe Logowanie do serwera. Można też dwukrotnie kliknąć obszar statusu połączenia w oknie IBM SPSS Modeler.

2. W oknie dialogowym wybierz pozycję Lokalny serwer, a następnie kliknij przycisk **OK**.

Dodawanie i edytowanie połączenia z serwerem IBM SPSS Modeler Server

Połączenia z serwerem można edytować lub dodawać ręcznie w oknie dialogowym Logowanie do serwera. Klikając przycisk Dodaj, można wyświetlić puste okno dialogowe Dodaj/Edytuj serwer, w którym można wprowadzać szczegółowe dane połączenia z serwerem. Po wybraniu istniejącego połączenia i kliknięciu przycisku Edytuj w oknie dialogowym Logowanie do serwera zostanie otwarte okno dialogowe zawierające szczegóły tego połączenia, co pozwala wprowadzić dowolne zmiany.

Uwaga: Nie można edytować połączenia z serwerem dodanego z programu IBM SPSS Collaboration and Deployment Services, ponieważ nazwa, port i inne szczegóły są definiowane w programie IBM SPSS Collaboration and Deployment Services. Sprawdzona procedura to użycie tych samych portów do komunikacji zarówno z programem IBM SPSS Collaboration and Deployment Services, jak i z klientem SPSS Modeler Client. Można je ustawić pod parametrami max_server_port i min_server_port w pliku options.cfg.

Aby dodać połączenia z serwerem

1. W menu Narzędzia kliknij opcję **Logowanie do serwera**. Zostanie otwarte okno dialogowe Logowanie do serwera.
2. W tym oknie dialogowym kliknij przycisk **Dodaj**. Zostanie otwarte okno dialogowe Logowanie do serwera: Dodaj/Edytuj serwer.
3. Wprowadź szczegóły połączenia z serwerem, a następnie kliknij przycisk **OK**, aby zapisać połączenie i powrócić do okna dialogowego Logowanie do serwera.

- **Serwer.** Wskaż dostępny serwer lub wybierz serwer z listy. Komputer serwera można zidentyfikować za pośrednictwem nazwy alfanumerycznej (na przykład *mój_serwer*) lub adresu IP przypisanego do komputera serwera (na przykład 202.123.456.78).
- **Port.** Podaj numer portu, na którym serwer nasłuchuje. Jeśli wartość domyślna nie działa, poproś administratora systemu o podanie poprawnego numeru portu.
- **Opis.** Wprowadź opcjonalny opis dla tego połączenia z serwerem.
- **Zapewnij bezpieczne połączenie (z pomocą SSL).** Określa, czy ma zostać użyte bezpieczne połączenie SSL (**Secure Sockets Layer**). SSL jest powszechnie używanym protokołem do zarządzania bezpieczeństwem przekazywania danych w Internecie. Aby możliwe było użycie tej funkcji, protokół SSL musi być włączony na serwerze będącym hostem programu IBM SPSS Modeler Server. W razie konieczności skontaktuj się z lokalnym administratorem systemu w celu uzyskania bardziej szczegółowych informacji.

Aby edytować połączenia z serwerem

1. W menu Narzędzia kliknij opcję **Logowanie do serwera**. Zostanie otwarte okno dialogowe Logowanie do serwera.
2. W tym oknie dialogowym wybierz połączenie, które chcesz edytować, a następnie kliknij przycisk **Edytuj**. Zostanie otwarte okno dialogowe Logowanie do serwera: Dodaj/Edytuj serwer.
3. Wprowadź szczegóły połączenia z serwerem, a następnie kliknij przycisk **OK**, aby zapisać zmiany i powrócić do okna dialogowego Logowanie do serwera.

Wyszukiwanie serwerów w programie IBM SPSS Collaboration and Deployment Services

Zamiast wprowadzania danych połączenia z serwerem ręcznie, można wybrać dostępny w sieci serwer lub klaster serwerów za pośrednictwem usługi Coordinator of Processes dostępnej w programie IBM SPSS Collaboration and Deployment Services. Klaster serwerów to grupa serwerów, z której usługa Coordinator of Processes określa serwer optymalny do odpowiedzi na żądanie przetwarzania.

Choć do okna dialogowego Logowanie do serwera można ręcznie dodawać serwery, wyszukiwanie dostępnych serwerów pozwala łączyć się z serwerami bez konieczności znajomości ich poprawnej nazwy i numeru portu. Informacje te są dostarczane automatycznie. Nadal jednak potrzebne są prawidłowe dane logowania: nazwa użytkownika, domena i hasło.

Uwaga: W przypadku braku dostępu do możliwości usługi Coordinator of Processes można nadal ręcznie wprowadzić nazwę serwera, z którym ma zostać nawiązane połączenie, lub można wybrać uprzednio zdefiniowaną nazwę. Więcej informacji można znaleźć w temacie “Dodawanie i edytowanie połączenia z serwerem IBM SPSS Modeler Server” na stronie 9.

Aby wyszukiwać serwery i klastry

1. W menu Narzędzia kliknij opcję **Logowanie do serwera**. Zostanie otwarte okno dialogowe Logowanie do serwera.
2. Kliknij przycisk **Szukaj** w tym oknie dialogowym, aby otworzyć okno dialogowe Szukaj serwerów. Jeśli podczas próby przeglądania danych usługi Coordinator of Processes okaże się, że użytkownik nie jest zalogowany do programu IBM SPSS Collaboration and Deployment Services, zostanie wyświetlony monit z prośbą o zalogowanie się.
3. Wybierz serwer lub klaster serwerów z listy.
4. Kliknij przycisk **OK**, aby zamknąć okno dialogowe i dodać to połączenie do tabeli w oknie dialogowym Logowanie do serwera.

Zmiana położenia katalogu tymczasowego

Wykonanie niektórych operacji w oprogramowaniu IBM SPSS Modeler Server może wymagać utworzenia plików tymczasowych. Domyślnie podczas obsługi plików tymczasowych program IBM SPSS Modeler korzysta z katalogu tymczasowego w systemie operacyjnym. Wykonując poniższe czynności, można zmienić położenie katalogu tymczasowego.

1. Utwórz nowy katalog o nazwie *spss* i podkatalog *servertemp*.
2. Edytuj plik *options.cfg* znajdujący się w katalogu */config* w katalogu instalacyjnym programu IBM SPSS Modeler. Edytuj parametr temp_directory zawarty w tym pliku i wprowadź w nim ciąg: temp_directory, "C:/spss/servertemp".
3. Następnie ponownie uruchom usługę IBM SPSS Modeler Server. W tym celu można kliknąć kartę **Usługi** dostępną w Panelu sterowania systemu Windows. Zatrzymaj usługę i uruchom ją ponownie, aby zastosować wprowadzone zmiany. Ponowne uruchomienie komputera również spowoduje ponowne uruchomienie usługi.

Wszystkie pliki tymczasowe będą teraz zapisywane w nowym katalogu.

Uwaga: najczęściej powtarzanym błędem podczas wykonywania tej operacji jest użycie nieprawidłowych znaków ukośnika. Należy używać ukośników zwykłych, a nie odwrotnych.

Uruchamianie wielu sesji programu IBM SPSS Modeler

Jeśli konieczne jest uruchomienie więcej niż jednej sesji programu IBM SPSS Modeler naraz, należy wprowadzić pewne zmiany w ustawieniach programu IBM SPSS Modeler i systemu Windows. Uruchomienie więcej niż jednej sesji może być konieczne na przykład w sytuacji, gdy mamy dwie odrębne licencje serwerowe i chcemy uruchomić dwa strumienie na dwóch różnych serwerach z tego samego komputera klienckiego.

Aby umożliwić jednoczesne działanie wielu sesji programu IBM SPSS Modeler:

1. Kliknij:

   Start > [Wszystkie] Programy > IBM SPSS Modeler 18.0
2. Kliknij prawym przyciskiem myszy na skrócie IBM SPSS Modeler 18 (tym z ikoną), a następnie wybierz polecenie **Właściwości**.
3. W polu tekstowym **Element docelowy** dopisz -noshare na końcu łańcucha.
4. W Eksploratorze Windows wybierz:

   Narzędzia > Opcje folderów...
5. Na karcie Typy plików wybierz opcję Strumień IBM SPSS Modeler i kliknij przycisk **Zaawansowane**.
6. W oknie dialogowym Edytowanie typu pliku wybierz opcję Otwórz za pomocą IBM SPSS Modeler i kliknij przycisk **Edytuj**.
7. W polu tekstowym **Aplikacja używana do wykonania akcji** dopisz -noshare przed argumentem **-stream**.

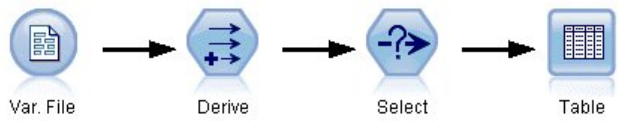
Przegląd interfejsu oprogramowania IBM SPSS Modeler

Na każdym etapie procesu eksploracji danych użytkownik może skorzystać z prostego w obsłudze interfejsu IBM SPSS Modeler. Algorytmy modelujące, takie jak predykcja, klasyfikacja, segmentacja i wykrywanie asocjacji, pozwalają na tworzenie wydajnych i dokładnych modeli. Wyniki działania modelu można w prosty sposób wdrażać i odczytywać w bazach danych, w programie IBM SPSS Statistics i wielu innych aplikacjach.

Obsługa programu IBM SPSS Modeler polega na wykonaniu trzyetapowego procesu opracowywania danych.

- Najpierw dane są wczytywane do oprogramowania IBM SPSS Modeler.
- Następnie na danych wykonywanych jest wiele operacji.
- Na koniec dane są przesyłane do położenia docelowego.

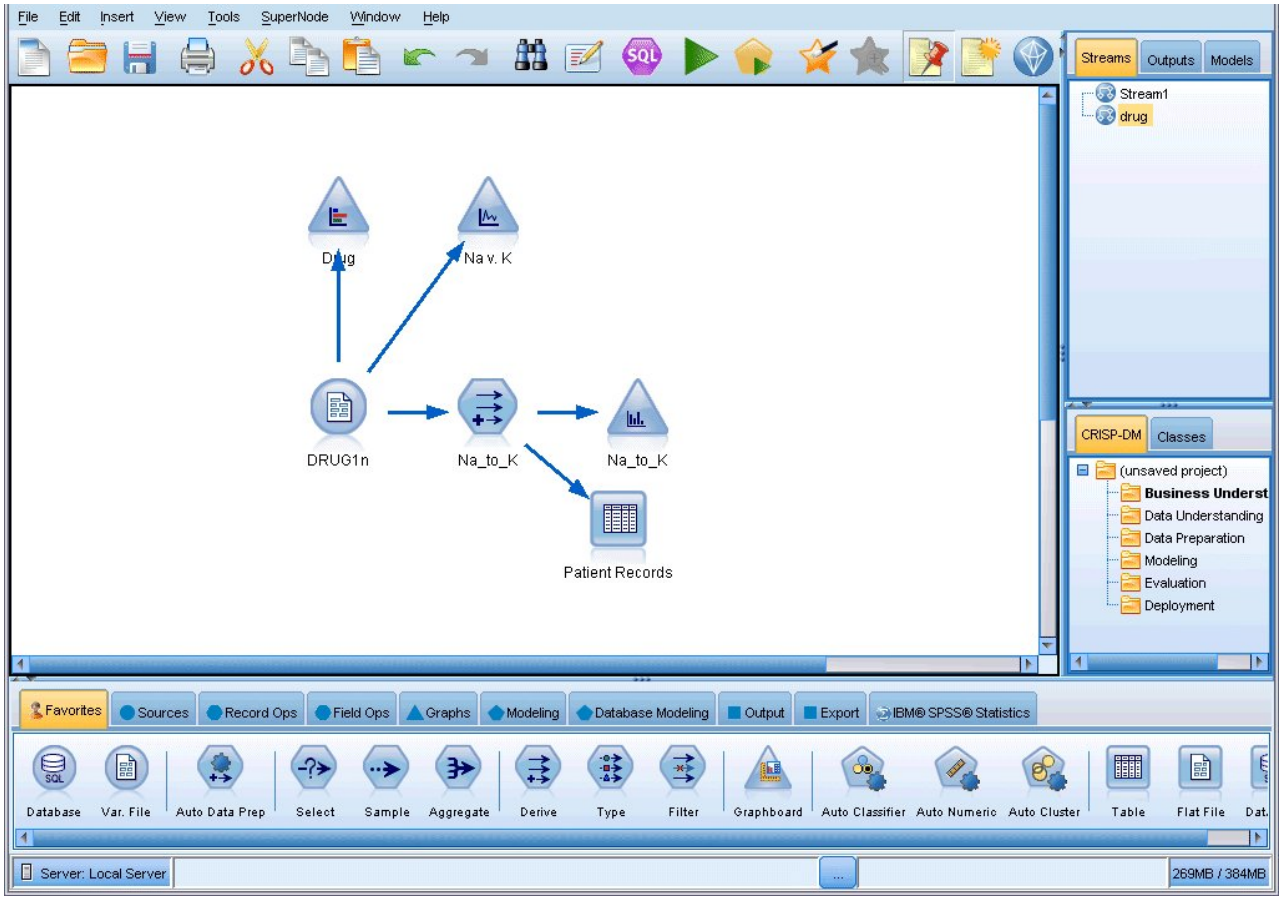
Ta sekwencja operacji jest nazywana **strumieniem danych**, ponieważ dane (rekord po rekordzie) wychodzą ze źródła, podlegają pewnym operacjom, a następnie są przekazywane do celu (modelu lub określonego typu danych wyjściowych).



Rysunek 2. Prosty strumień

Obszar roboczy strumienia IBM SPSS Modeler

Obszar roboczy strumienia to największa część okna IBM SPSS Modeler. W tym obszarze użytkownik tworzy strumienie danych i pracuje z nimi.



Rysunek 3. Obszar roboczy programu IBM SPSS Modeler (widok domyślny)

Tworzenie strumienia polega na narysowaniu (na głównym obszarze roboczym interfejsu) diagramu operacji biznesowych wykonywanych na danych. Każda operacja jest oznaczona ikoną lub **węzłem**, a węzły są łączone do postaci **strumieni** odzwierciedlających przepływ danych w ramach każdej operacji.

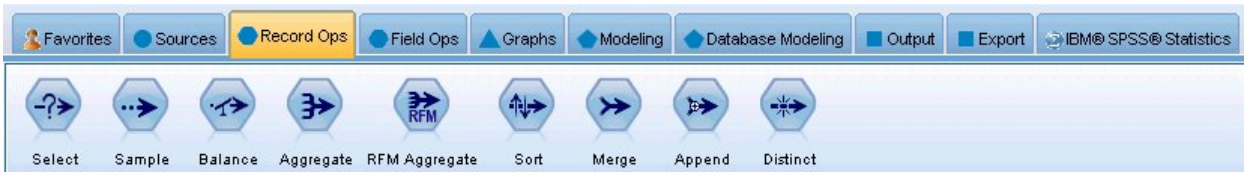
Jednocześnie w oprogramowaniu IBM SPSS Modeler można opracowywać wiele strumieni, korzystając z jednego obszaru roboczego strumienia lub wielu obszarów roboczych. W trakcie sesji strumienie są przechowywane w menedżerze strumieni dostępnym w prawym górnym narożniku okna IBM SPSS Modeler.

Paleta węzłów

Większość narzędzi do pracy z danymi i do modelowania danych w programie SPSS Modeler jest dostępnych na *palecie węzłów*, która rozciąga się u dołu okna pod obszarem roboczym strumienia.

Na przykład karta palety *rekordy* zawiera węzły, które umożliwiają wykonywanie operacji na **rekordach** danych, takich jak wybieranie, scalanie i dołączanie.

Aby dodawać węzły do obszaru roboczego, klikaj dwukrotnie ikony na palecie węzłów lub przeciągaj je do obszaru roboczego. Następnie węzły można połączyć, tworząc *strumień* odzwierciedlający przepływ danych.



Rysunek 4. Karta Rekordy na palecie węzłów

Każda karta palety zawiera zbiór pokrewnych węzłów używanych w różnych fazach działania strumienia, na przykład:

- **Źródła** Te węzły służą do wprowadzania danych do programu SPSS Modeler.
- **Rekordy** Te węzły wykonują na *rekordach* danych takie operacje, jak wybieranie, scalanie i dołączanie.
- **Zmienne** Te węzły wykonują operacje na *zmiennych*, np. filtrowanie, wywodzenie nowych zmiennych i określanie poziomu pomiaru danych zmiennych.
- **Wykresy** Te węzły prezentują dane w formie graficznej przed modelowaniem i po nim. Dostępne formy graficzne to wykresy, histogramy, węzły sieciowe i wykresy ewaluacyjne.
- **Modelowanie** Te węzły korzystają z algorytmów modelowania dostępnych w programie SPSS Modeler, takich jak sieci neuronowe, drzewa decyzyjne, algorytmy grupowania i określanie sekwencji danych.
- **Modelowanie w bazie** Te węzły korzystają z algorytmów modelowania dostępnych w bazach danych Microsoft SQL Server, IBM DB2, Oracle i Netezza.
- **Wynik** Te węzły generują różnych wyniki w postaci danych, wykresów i wyników modeli, które można przeglądać w programie SPSS Modeler.
- **Eksport** Te węzły generują różnych dane wyjściowe, które można przeglądać w aplikacjach zewnętrznych, takich jak IBM SPSS Data Collection lub Excel.
- **IBM SPSS Statistics** Te węzły importują dane z programu IBM SPSS Statistics lub eksportują do niego dane, a także wykonują procedury IBM SPSS Statistics.

W miarę nabywania doświadczenia w pracy z programem SPSS Modeler, można dostosowywać zawartość palet do indywidualnych potrzeb.

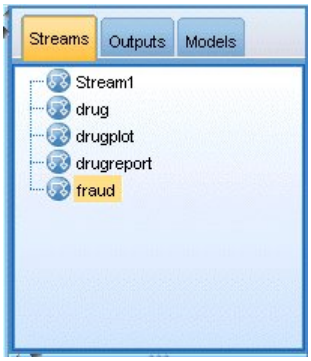
Po lewej stronie palety węzłów można filtrować wyświetlane węzły, wybierając kategorię: Analytic Server, Klasyfikacja, Związek lub Segmentacja.

Pod paletą węzłów znajduje się panel raportów, na którym wyświetlane są informacje o postępach w realizacji różnych operacji, np. wczytywania danych do strumienia. Poniżej palety węzłów znajduje się także panel statusu, na którym wyświetlane są informacje o tym, co w danej chwili robi aplikacja, a także komunikaty informujące o konieczności wprowadzenia danych przez użytkownika.

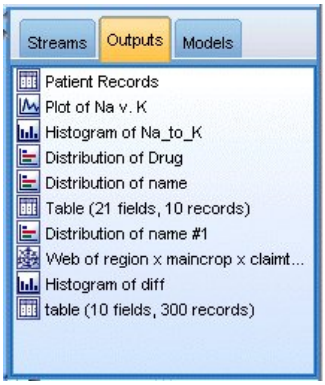
Panele zarządzania w programie IBM SPSS Modeler

W prawym górnym narożniku okna znajduje się panel zarządzania. Ten obszar składa się z trzech kart służących do zarządzania strumieniami, wynikami i modelami.

Za pomocą karty Strumienie można otwierać, zapisywać i usuwać strumienie otwarte w sesji oraz zmieniać ich nazwy.



Rysunek 5. Karta Strumienie



Rysunek 6. Karta Wyniki

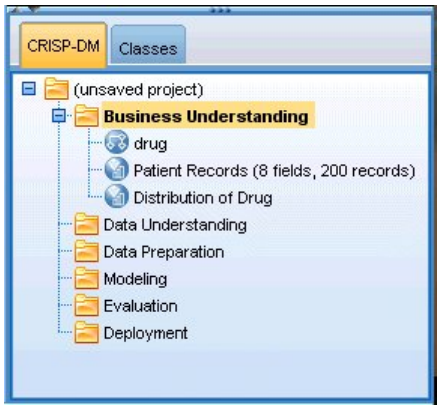
Na karcie Wyniki znajduje się wiele plików, takich jak wykresy i tabele, utworzonych w ramach obsługi strumieni w oprogramowaniu IBM SPSS Modeler. Użytkownik może wyświetlać, zapisywać i zamykać tabele, wykresy oraz raporty znajdujące się na tej karcie, jak również zmieniać ich nazwy.

Rysunek 7. Karta Modele zawierająca modele użytkowe

Karta Modele to najbardziej funkcjonalna z kart zarządzania. Na tej karcie znajdują się wszystkie **modele użytkowe**, które zawierają modele wygenerowane w bieżącej sesji oprogramowania IBM SPSS Modeler. Modele te można przeglądać bezpośrednio z poziomu karty Modele, można je również dodawać do strumienia w obszarze roboczym.

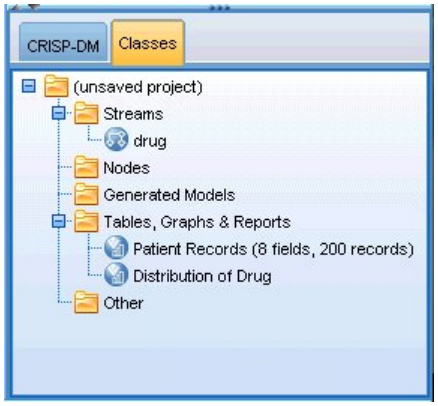
Projekty w programie IBM SPSS Modeler

W prawym dolnym narożniku okna znajduje się panel projektów służący do tworzenia **projektów** eksploracji danych (grup plików powiązanych z zadaniem eksploracji danych) oraz zarządzania nimi. Projekty utworzone w oprogramowaniu IBM SPSS Modeler można wyświetlać na dwa sposoby: w widoku Klas oraz CRISP-DM.



Rysunek 8. Widok CRISP-DM

Karta CRISP-DM pozwala na organizowanie projektów zgodnie z formatem Cross-Industry Standard Process for Data Mining, będącym sprawdzoną, branżową metodologią działań. Zarówno w przypadku doświadczonych, jak i młodych stażem eksploratorów danych, narzędzie CRISP-DM ułatwia przeprowadzanie operacji organizacyjnych i usprawnia uzyskiwanie określonych wyników pracy.



Rysunek 9. Widok Klasy

Karta Klasy umożliwia organizowanie pracy wg kategorii w oprogramowaniu IBM SPSS Modeler, czyli wg typów tworzonych obiektów. Ten widok jest przydatny podczas tworzenia zestawienia danych, strumieni i modeli.

Pasek narzędzi programu IBM SPSS Modeler

W górnej części okna IBM SPSS Modeler znajduje się pasek narzędzi umożliwiający wykonanie wielu przydatnych funkcji. Poniżej przedstawiono przyciski widoczne na pasku narzędzi wraz z opisami ich działania.

Utwórz nowy strumień

Otwórz strumień

Zapisz strumień

Wydrukuj bieżący strumień

Wytnij i przenieś do schowka

Kopiuj do schowka

Wklej zaznaczenie

Cofnij ostatnią czynność

Powtórz

Wyszukaj węzły

Edytuj właściwości strumienia

Wyświetl podgląd tworzenia kodu SQL

Wykonaj bieżący strumień

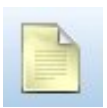
Wykonaj wybrany fragment strumienia

Zatrzymaj strumień (funkcja dostępna wyłącznie, gdy strumień jest wykonywany)

Dodaj Superwęzeł

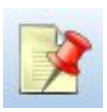
Powiększ (tylko Superwęzły)

Pomniejsz (tylko Superwęzły)



Brak znacznika w strumieniu

Wstaw komentarz



Ukryj znacznik strumienia (o ile istnieje)

Pokaż ukryte znaczniki strumienia

Otwórz strumień w oprogramowaniu IBM SPSS Modeler Advantage

Znaczniki strumienia składają się z komentarzy, łączy modelu i wskaźników gałęzi oceniania.

Łącza modelu opisano w publikacji *IBM SPSS - węzły modelowania*.

Dostosowywanie paska narzędzi

Użytkownik może zmienić wiele aspektów paska narzędzi:

- Wyświetlanie
- Dostępność podpowiedzi ikon
- Wielkość ikon

Włączanie i wyłączanie paska narzędzi:

1. W menu głównym kliknij opcje:
   Widok > Pasek narzędzi > Wyświetlanie

Zmiana ustawień rozmiarów ikon lub podpowiedzi:

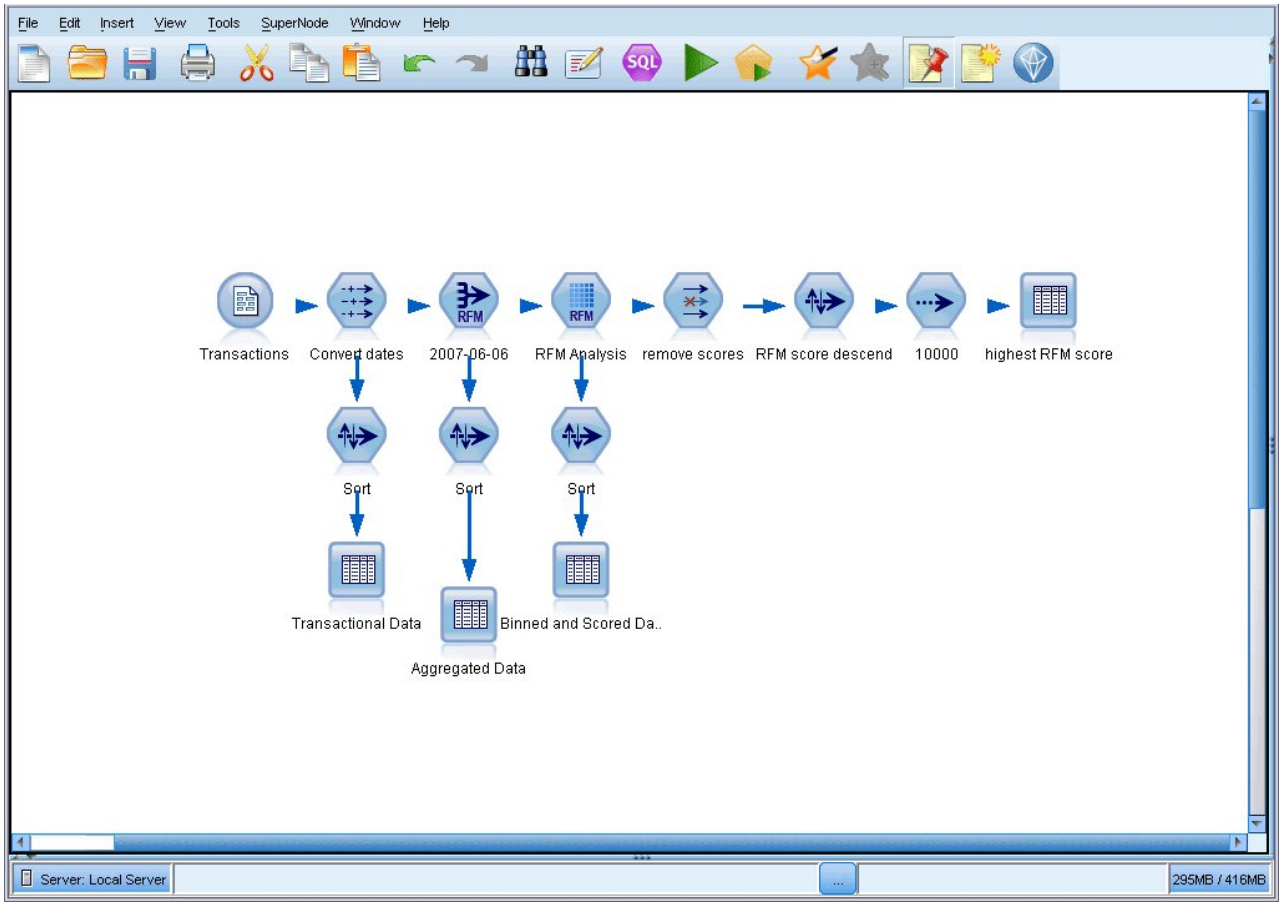
1. W menu głównym kliknij opcje:
   Widok > Paski narzędzi > Dostosuj

Kliknij pozycję **Pokaż podpowiedzi** lub **Duże przyciski**.

Dostosowywanie okna programu IBM SPSS Modeler

Korzystając z linii podziału między różnymi fragmentami interfejsu programu SPSS Modeler, można zmieniać rozmiar lub zamykać narzędzia, aby dopasować środowisko pracy do swoich preferencji. Na przykład w przypadku pracy z dużym strumieniem można użyć małych strzałek znajdujących się na każdej z linii podziału w celu zamknięcia palety węzłów, panelu zarządzania oraz panelu projektu. Powoduje to maksymalizację obszaru roboczego strumienia, udostępniając wystarczający obszar roboczy dla jednego dużego lub wielu strumieni.

W menu Widok można też kliknąć pozycję **Paleta węzłów**, **Zarządzanie** lub **Projekt**, aby włączyć lub wyłączyć wyświetlanie tych elementów.



Rysunek 10. Zmaksymalizowany obszar roboczy strumienia

Jako alternatywy wobec zamknięcia palety węzłów oraz paneli zarządzania i projektu można użyć obszaru roboczego strumienia jako strony z możliwością przewijania; w tym celu należy użyć pasków przewijania w pionie i w poziomie znajdujących się z boku i u dołu okna SPSS Modeler.

Można także sterować wyświetlaniem znaczników ekranowych, składających się z komentarzy do strumieni, łączy modelu i oznaczeń gałęzi oceniania. W celu włączenia lub wyłączenia wyświetlania kliknij opcję:

Widok > Adnotacja strumienia

Zmiana rozmiaru ikony strumienia

Użytkownik może zmienić rozmiar ikon strumienia, korzystając z poniższych metod:

- za pomocą ustawienia właściwości strumienia,
- za pomocą menu podręcznego w strumieniu,
- korzystając z klawiatury.

Cały widok strumienia można przeskalować do dowolnego rozmiaru w zakresie od 8% do 200% w odniesieniu do standardowego rozmiaru ikony.

Skalowanie całego strumienia (właściwości strumienia)

1. Z menu głównego wybierz pozycje
   Narzędzia > Właściwości strumienia > Opcje > Układ.
2. Z menu Rozmiar ikony wybierz żądany rozmiar.
3. Kliknij przycisk **Zastosuj**, aby wyświetlić wynik operacji.
4. Kliknij przycisk **OK**, aby zapisać zmiany.

Skalowanie całego strumienia (menu)

1. Prawym przyciskiem myszy kliknij tło strumienia w obszarze roboczym.
2. Wybierz pozycję **Rozmiar ikony** i wybierz żądany rozmiar.

Skalowanie całego strumienia (klawiatura)

1. Na klawiaturze głównej naciśnij klawisze Ctrl + [-], aby zredukować skalę powiększenia do kolejnej mniejszej wartości.
2. Na klawiaturze głównej naciśnij klawisze Ctrl + Shift + [+], aby zwiększyć skalę powiększenia do kolejnej większej wartości.

Ta funkcja jest szczególnie przydatna w celu wyświetlenia ogólnego widoku złożonego strumienia. Dzięki niej można również zmniejszyć liczbę stron, na których zostanie wydrukowany strumień.

Korzystanie z myszy w oprogramowaniu IBM SPSS Modeler

Najczęstsze zastosowania myszy w oprogramowaniu IBM SPSS Modeler:

- **Kliknięcie.** Za pomocą lewego lub prawego przycisku można wybierać opcje z menu, otwierać menu podręczne i obsługiwać elementy sterujące oraz opcje. Kliknięcie i przytrzymanie przycisku pozwala na przenoszenie i przeciąganie węzłów.
- **Kliknięcie dwukrotne.** Kliknięcie dwukrotne lewym przyciskiem myszy powoduje wstawianie węzłów w obszarze roboczym strumienia i pozwala na edytowanie istniejących węzłów.
- **Kliknięcie środkowym przyciskiem.** Kliknięcie środkowym przyciskiem myszy i przeciągnięcie kursora powoduje połączenie węzłów w obszarze roboczym strumienia. Węzeł można odłączyć, klikając go dwukrotnie środkowym przyciskiem myszy. Jeśli użytkownik nie dysponuje myszą trzyprzyciskową, to trzeci przycisk można emulować, przytrzymując klawisz Alt podczas klikania lub przeciągania.

Używanie skrótów klawiaturowych

Wiele operacji programistycznych w środowisku wizualnym IBM SPSS Modeler można wykonać za pomocą przypisanych do nich skrótów klawiaturowych. Przykładowo: węzeł można usunąć, klikając go i naciskając klawisz Delete. Podobnie strumień można szybko zapisać, przytrzymując klawisz Ctrl i naciskając klawisz S. Komendy sterujące są uruchamiane za pomocą kombinacji klawisza Ctrl i innych klawiszy, np. Ctrl+S.

Używanych jest również wiele standardowych skrótów klawiaturowych systemu Windows, takich jak Ctrl+X (wycinanie). Są one obsługiwane w oprogramowaniu IBM SPSS Modeler wraz ze skrótami dostępnymi tylko w tym programie.

Uwaga: niekiedy stare skróty klawiaturowe IBM SPSS Modeler powodują konflikt ze standardowymi skrótami klawiaturowymi systemu Windows. Stare skróty są uruchamiane w oprogramowaniu z użyciem klawisza Alt. Przykładowo: skrót Ctrl+Alt+C służy do włączania i wyłączania pamięci podręcznej.

Tabela 1. Obsługiwane skróty

Klawisz skrótu	**Funkcja**
Ctrl+A	Zaznacz wszystko
Ctrl+X	Wytnij
Ctrl+N	Nowy strumień
Ctrl+O	Otwórz strumień
Ctrl+P	Drukuj
Ctrl+C	Kopiuj
Ctrl+V	Wklej
Ctrl+Z	Cofnij
Ctrl+Q	Zaznacz wszystkie węzły znajdujące się poniżej danego węzła
Ctrl+W	Usuń zaznaczenie wszystkich poniższych węzłów (przełączanie za pomocą klawiszy Ctrl+Q)
Ctrl+E	Wykonaj od wybranego węzła
Ctrl+S	Zapisz bieżący strumień
Alt+klawisze strzałek	Przenieś węzeł wybrany w obszarze roboczym strumienia w kierunku wskazanym za pomocą strzałki
Shift+F10	Otwórz menu podręczne wybranego węzła

Tabela 2. Obsługiwane skróty - stare klawisze skrótów

Klawisz skrótu	**Funkcja**
Ctrl+Alt+D	Duplikuj węzeł
Ctrl+Alt+L	Załaduj węzeł
Ctrl+Alt+R	Zmień nazwę węzła
Ctrl+Alt+U	Utwórz węzeł danych użytkownika
Ctrl+Alt+C	Wł./wył. pamięć podręczną
Ctrl+Alt+F	Opróżnij pamięć podręczną
Ctrl+Alt+X	Rozwiń Superwęzeł
Ctrl+Alt+Z	Powiększ/pomniejsz
Usuń	Usuń węzeł lub połączenie

Drukowanie

W programie IBM SPSS Modeler możliwe jest wydrukowanie następujących obiektów:

- Diagramy strumienia
- Wykresy
- Tabele
- Raporty (z węzła Raport oraz z raportów projektu)
- Skrypty (z Właściwości strumienia, z okien dialogowych Skrypt samodzielny lub Skrypt superwęzła)

- Modele (przeglądarki modeli, karty okien dialogowych z aktualnym fokusem, widoku drzewa)
- Adnotacje (za pośrednictwem karty Adnotacje dla danych wynikowych)

Aby wydrukować projekt:

- Aby wydrukować bez podglądu, kliknij przycisk Drukuj na pasku narzędzi.
- Aby skonfigurować stronę przed przystąpieniem do drukowania, wybierz opcję **Ustawienia strony** z menu Plik.
- Aby przed drukowaniem wyświetlić podgląd, wybierz pozycję **Podgląd wydruku** z menu Plik.
- Aby wyświetlić standardowe okno dialogowe z opcjami wyboru drukarek oraz aby określić opcje wyglądu, wybierz pozycję **Drukuj** z menu Plik.

Automatyzacja procesów programu IBM SPSS Modeler

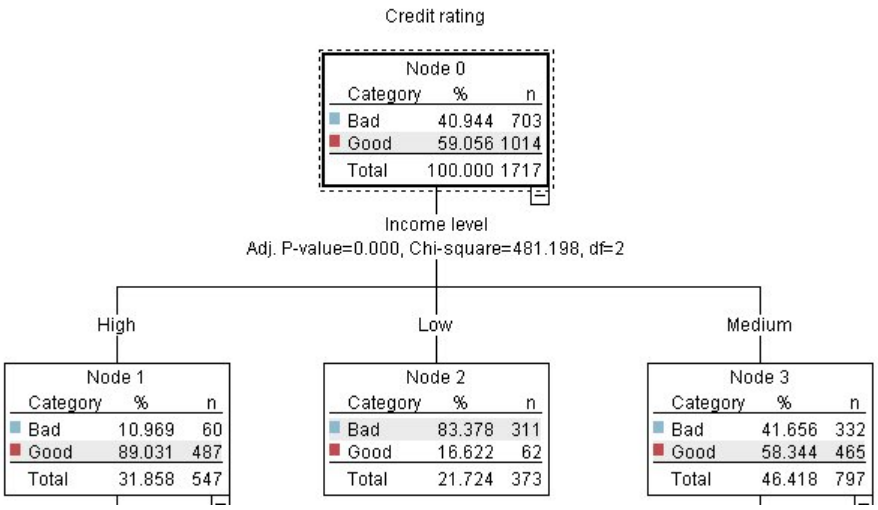
Ponieważ zaawansowana eksploracja danych może być procesem złożonym i niekiedy dość długotrwałym, program IBM SPSS Modeler oferuje różne możliwości programowania i automatyzacji.

- Język CLEM (ang. **Control Language for Expression Manipulation**) to język do analizowania i manipulowania danymi przechodzącymi przez strumienie IBM SPSS Modeler. Narzędzia do eksploracji danych wykorzystują język CLEM w szerokim zakresie w operacjach strumieniowych, do wykonywania zadań zarówno tak prostych, jak wyliczanie zysku na podstawie kosztów i przychodów, jak i tak złożonych, jak transformowanie treści blogów w zestaw zmiennych i rekordów zawierający użyteczne informacje.
- **Skrypty** są potężnym narzędziem automatyzacji procesów interfejsu użytkownika. Skrypty umożliwiają wykonywanie tego samego rodzaju czynności, jakie wykonywane są przez użytkownika za pomocą myszy lub klawiatury. Można także określić dane wynikowe i manipulować wygenerowanymi modelami.

Rozdział 3. Wstęp do modelowania

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

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



Rysunek 11. Prosty model drzewa decyzyjnego

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

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

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

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

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

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

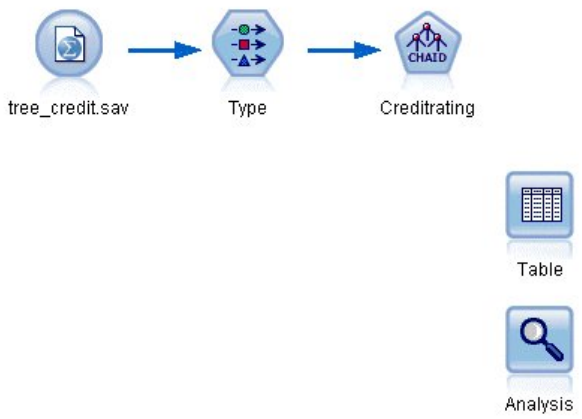
Używając modelu drzewa decyzyjnego, można przeprowadzić analizę cech dwóch grup klientów i przewidzieć prawdopodobieństwo niespłacania kredytu.

W tym przykładzie zastosowano strumień o nazwie *modelingintro.str*, który jest dostępny w folderze *Demos*, podfolder *streams*. Plik danych to *tree_credit.sav*. Więcej informacji można znaleźć w temacie “Folder Demos” na stronie 5.

Przyjrzyjmy się strumieniowi.

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

Tworzenie strumienia



Rysunek 12. Strumień modelowania

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

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

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

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

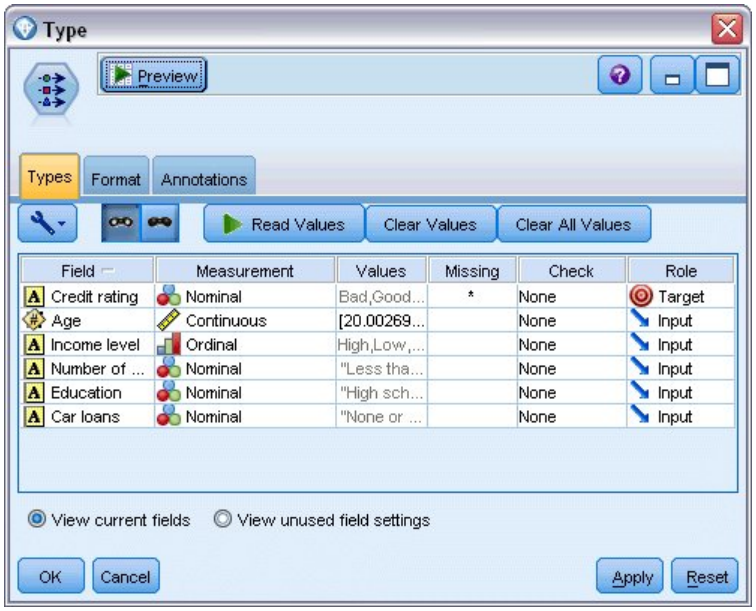
Ten strumień zawiera również węzły Tabela i Analiza, które będą używane do wyświetlania wyników oceniania po utworzeniu modelu użytkowego i dodaniu go do strumienia.

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

Rysunek 13. Odczyt danych z użyciem węzła źródłowego Plik Statistics

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

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



Rysunek 14. Ustawienie zmiennych przewidywanych i wejściowych w węźle Typ

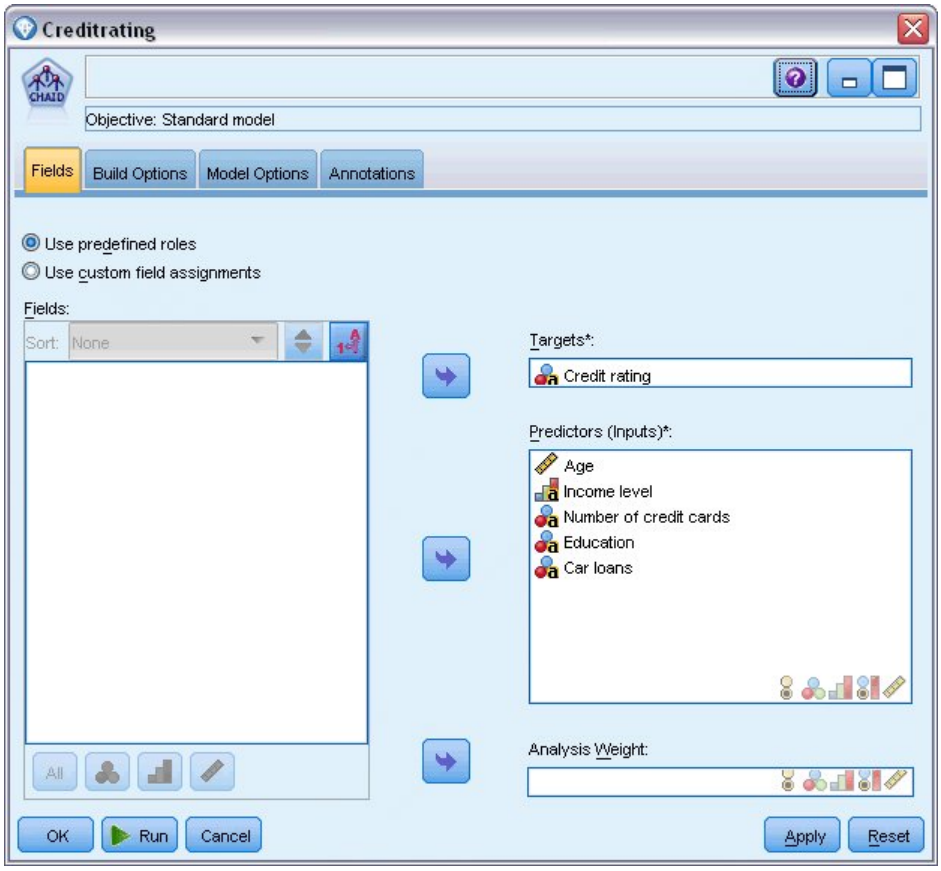
Dla każdej zmiennej węzeł Typ również określa **rolę** wskazującą udział poszczególnych zmiennych w modelowaniu. Rola jest ustawiana na wartość *Przewidywana* dla zmiennej *Credit rating*, która wskazuje, czy dany klient nie spłaca kredytu. Jest to **zmienna przewidywana** lub zmienna, dla której zamierzamy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

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

Węzeł modelowania CHAID generuje model.

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

1. Kliknij zakładkę Opcje budowania.



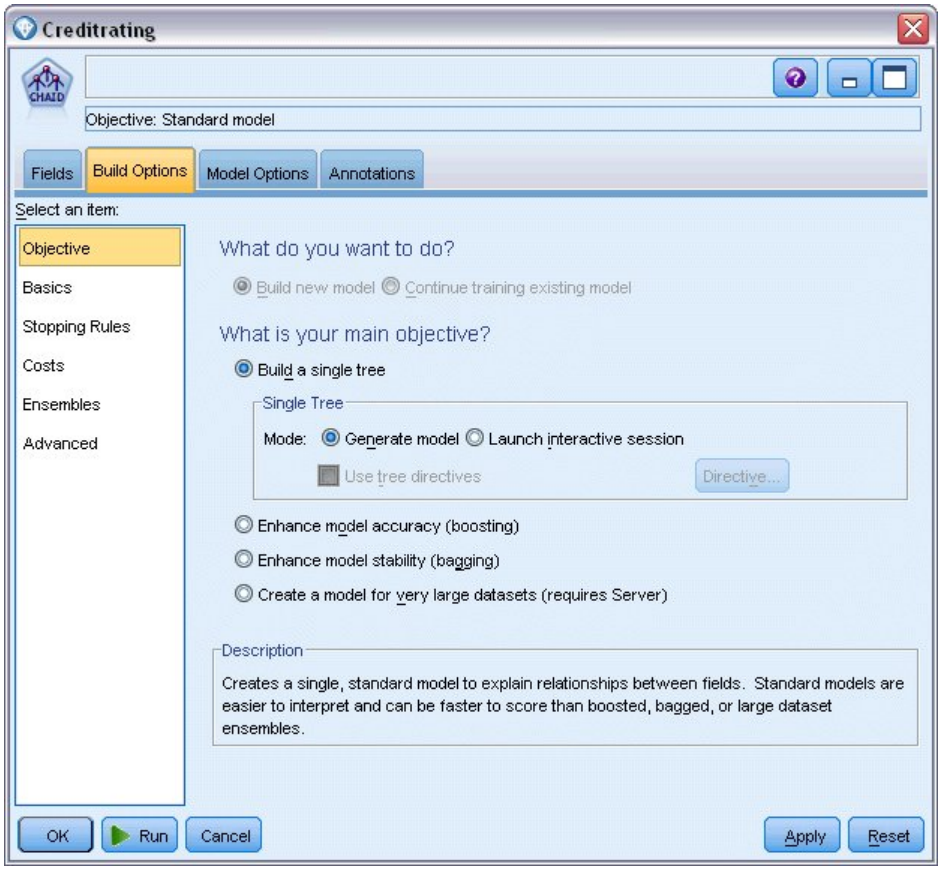
Rysunek 15. Węzeł modelowania CHAID, zakładka Zmienne

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

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

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

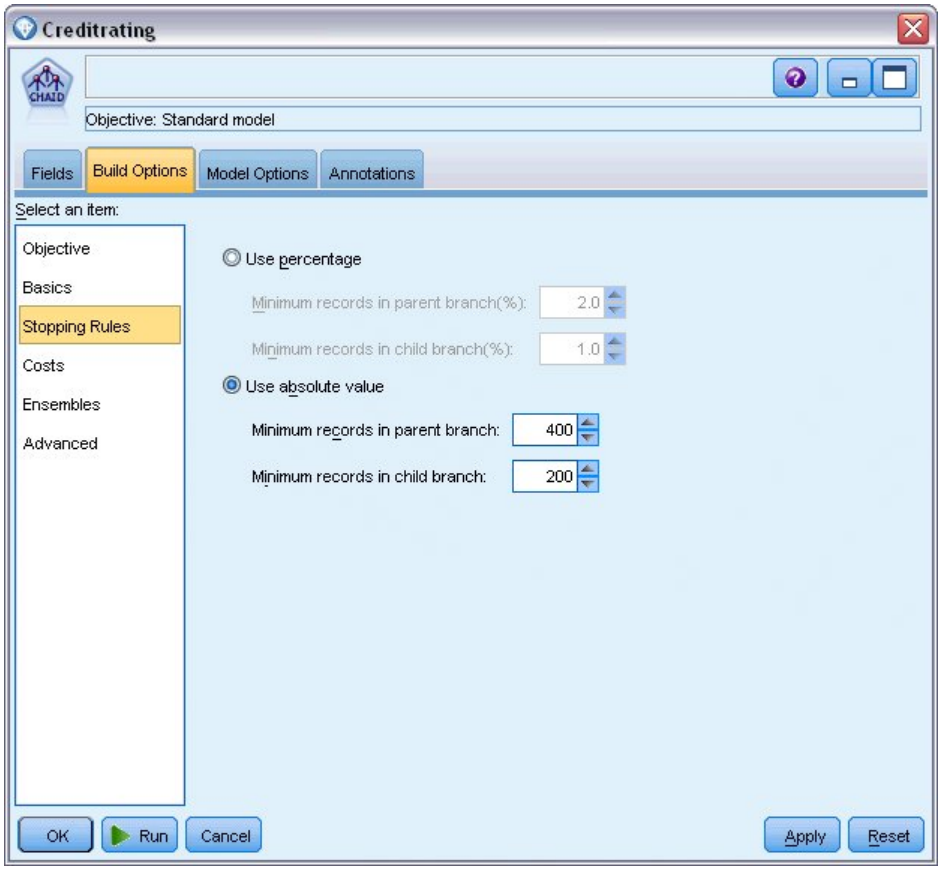
Teraz można opcjonalnie uruchomić interaktywną sesję modelowania, która pozwoli na dostosowanie modelu, jednak w tym przykładzie zostanie po prostu wygenerowany model z zastosowaniem domyślnego ustawienia trybu: **Generuj model**.



Rysunek 16. Węzeł modelowania CHAID, karta Opcje budowania

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

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

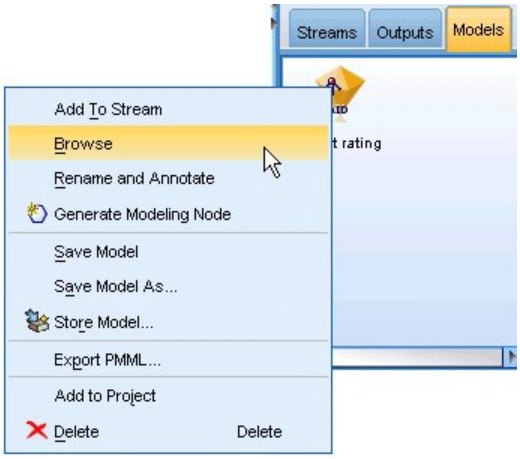


Rysunek 17. Ustawianie kryteriów zatrzymywania dla budowania drzewa decyzyjnego

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

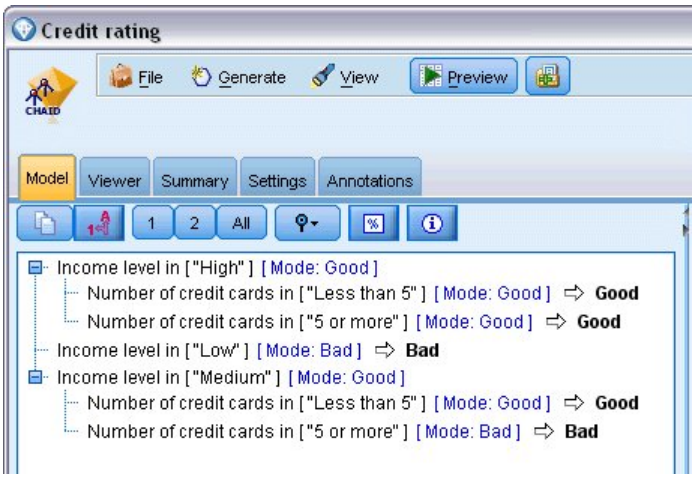
Przeglądanie modelu

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



Rysunek 18. Paleta modeli

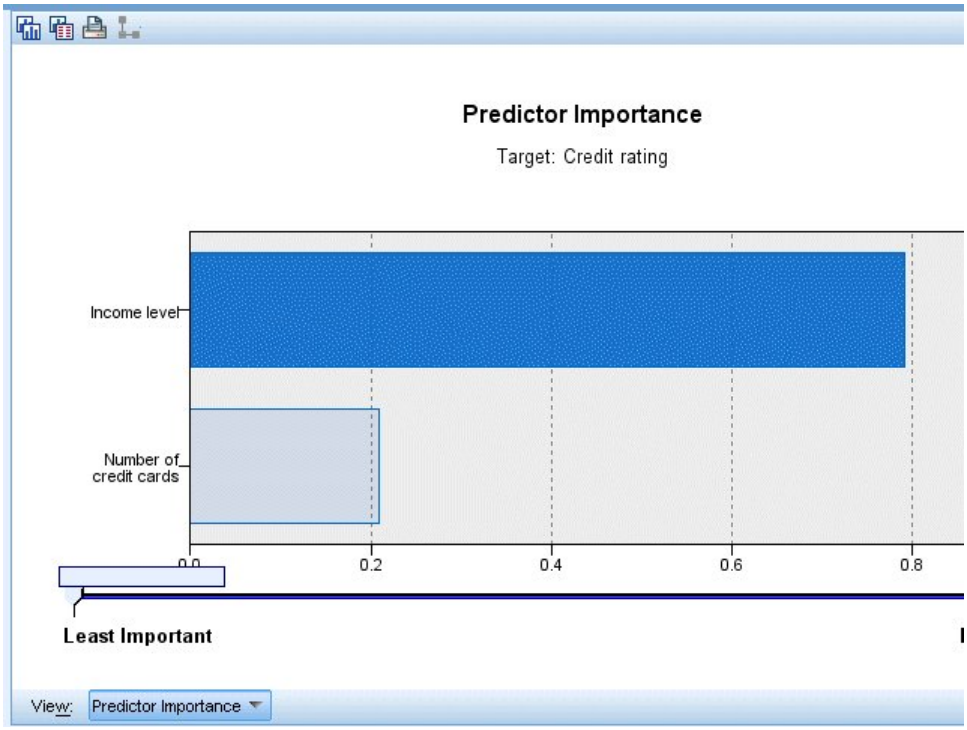
W przypadku modelu użytkowego CHAID na karcie Model szczegóły są wyświetlane w postaci zestawu reguł — zwykle jest to szereg reguł, jakie można zastosować w celu przypisania poszczególnych rekordów do węzłów podrzędnych w oparciu o wartości różnych zmiennych wejściowych.



Rysunek 19. Model użytkowy CHAID, zestaw reguł

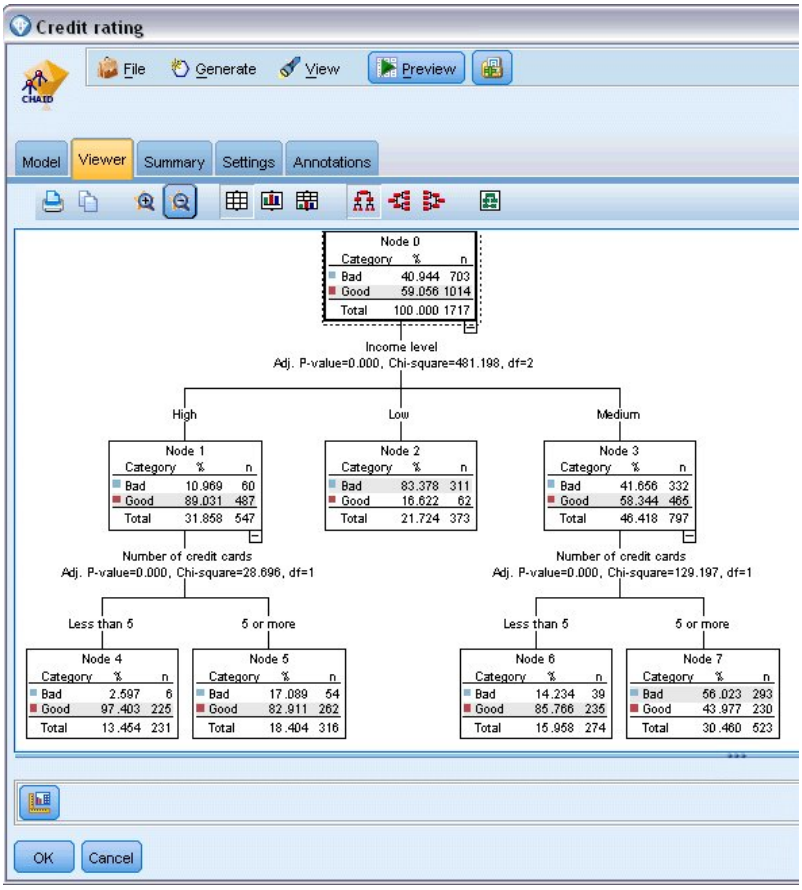
Dla każdego węzła końcowego drzewa decyzyjnego — to znaczy dla tych węzłów, które nie są dalej podzielone — zwracana jest predykcja *Good* lub *Bad*. W każdym przypadku predykcja jest określana według **dominanty** lub najczęściej udzielanej odpowiedzi dla rekordów, które należą do tego węzła.

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



Rysunek 20. Wykres ważności predyktorów

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



Rysunek 21. Karta Przegląd w modelu użytkowym, z wybraną opcją pomniejszania

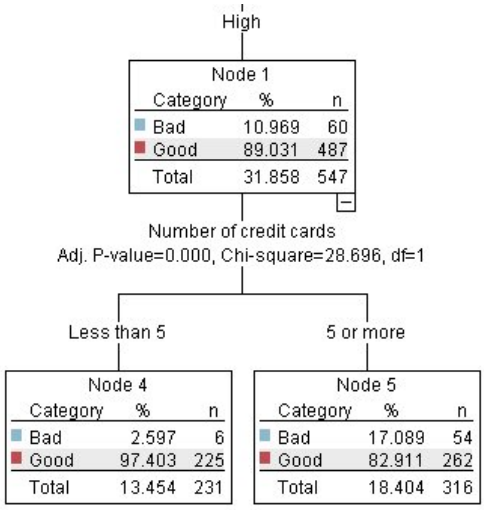
Patrząc na górną część drzewa w pierwszym węźle (węzeł 0), widzimy podsumowanie dla wszystkich rekordów ze zbioru danych. Tylko ponad 40% obserwacji ze zbioru danych jest klasyfikowanych jako wysokie ryzyko. Jest to dość duża proporcja, dlatego sprawdźmy, czy drzewo zawiera jakieś wskazówki, jakie czynniki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

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

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

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

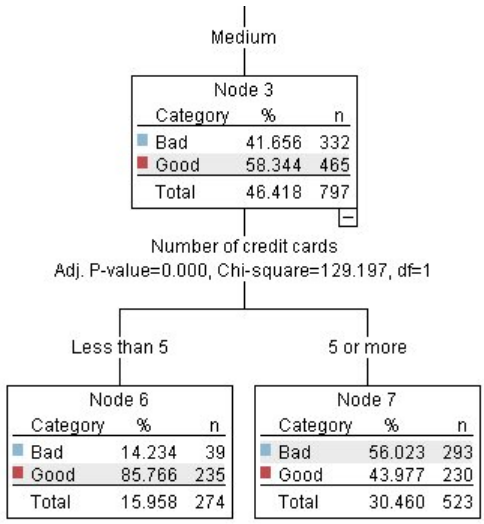
Należy zwrócić uwagę, jak model podzielił klientów na dwie podkategorie (węzły 4 i 5) w oparciu o liczbę posiadanych kart kredytowych. W przypadku klientów z wysokim dochodem, jeśli kredyt zostanie udzielony tylko tym osobom, które mają mniej niż 5 kart kredytowych, można zwiększyć wskaźnik sukcesu z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



Rysunek 22. Widok drzewa klientów z wysokim dochodem

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

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



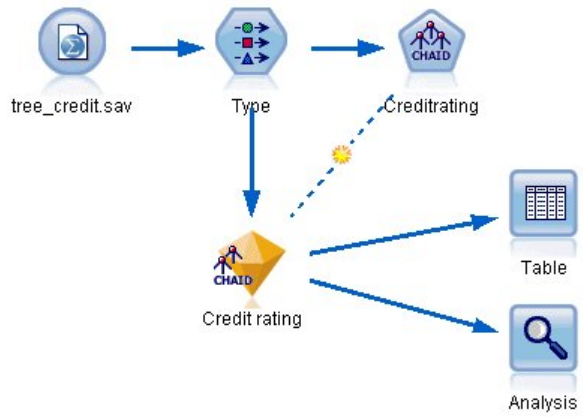
Rysunek 23. Widok drzewa klientów ze średnim dochodem

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

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

Ewaluacja modelu

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

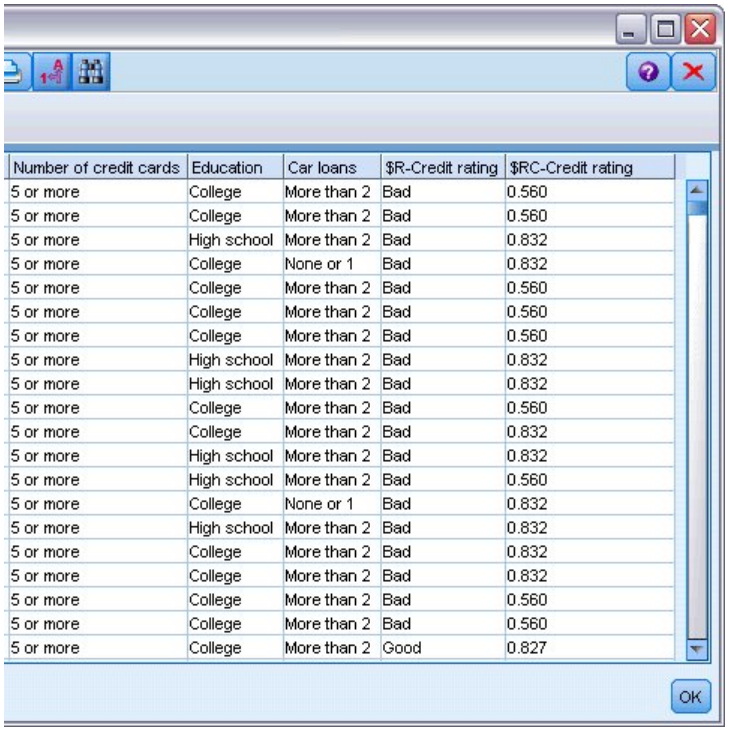


Rysunek 24. Dołączanie modelu użytkowego do węzłów wynikowych w celu przeprowadzenia ewaluacji modelu

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

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

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



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

OK

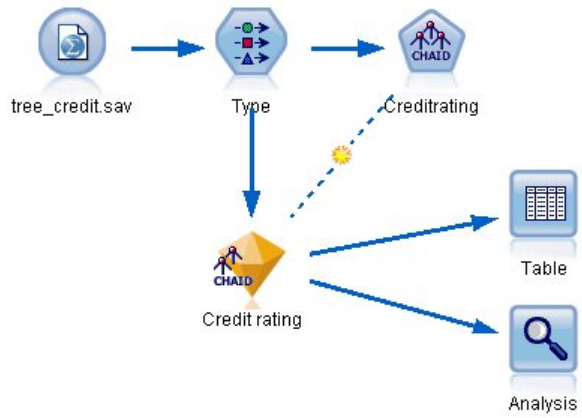
Rysunek 25. Tabela przedstawiająca wygenerowane oceny i współczynniki ufności

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

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

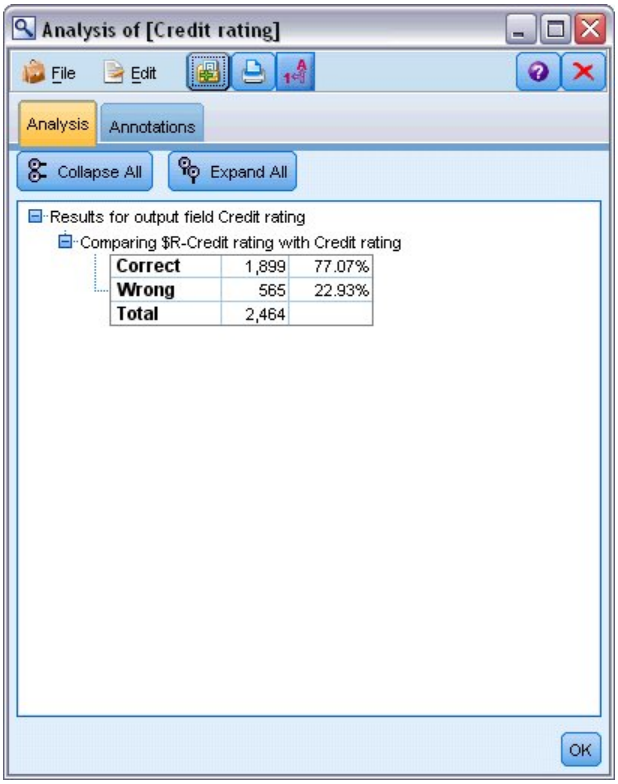
Aby dokładnie dowiedzieć się, ile predykcji jest poprawnych, można przejrzeć tabelę i zliczyć liczbę rekordów, w których wartość przewidzianej zmiennej *$R-Credit rating* jest zgodna z wartością zmiennej *Credit rating*. Na szczęście istnieje dużo łatwiejszy sposób — można użyć węzła Analiza, który zrobi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

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



Rysunek 26. Dołączanie węzła Analiza

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



Rysunek 27. Porównywanie wyników analizy z odpowiedziami obserwowanymi i rzeczywistymi

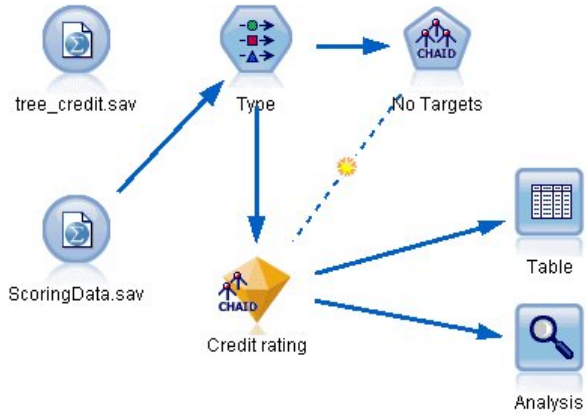
Wynik jest ograniczony przez fakt, że oceniane są te same rekordy, jakie zostały użyte do oszacowania modelu. W rzeczywistości można użyć węzła Podział, aby podzielić dane na osobne próby do uczenia i ewaluacji.

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

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

Ocenianie rekordów

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



Rysunek 28. Dołączanie nowych danych do oceny

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

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

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

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

Podsumowanie

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

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

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

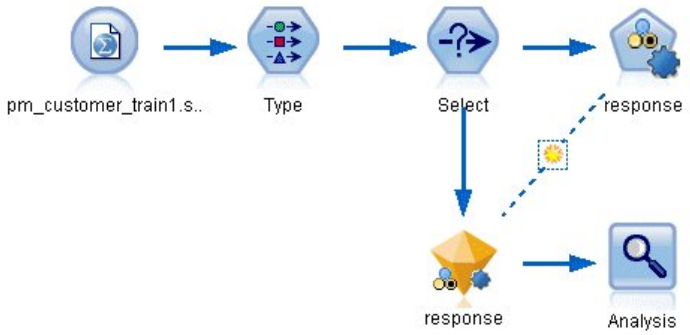
Rozdział 4. Zautomatyzowane modelowanie dla przewidywanej zmiennej typu flaga

Modelowanie odpowiedzi klienta (Auto Klasyfikacja)

Węzeł Auto Klasyfikacja pozwala użytkownikowi na automatyczne tworzenie i porównywanie wielu różnych modeli dla flagi (takiej jak, czy dany klient nie spłaci kredytu lub odpowie na określoną ofertę) lub nominalnych zmiennych przewidywanych. W tym przykładzie wyszukamy flagę (tak lub nie) wyniku. W ramach relatywnie prostego strumienia węzeł generuje i określa rangę zestawu modeli kandydackich, wybiera jeden, który działa najlepiej, i łączy je w pojedynczy zagregowany (zespolony) model. Takie podejście łączy łatwość automatyzacji z korzyściami łączenia wielu modeli, które często zwracają dokładniejsze predykcje, niż można uzyskać z jednego modelu.

W tym przykładzie fikcyjna firma chce uzyskać wyższy zysk, dobierając właściwą ofertę dla każdego klienta.

To podejście podkreśla korzyści automatyzacji. Podobny przykład używający przewidywanej zmiennej ilościowej (przedział numeryczny) przedstawiono w sekcji Wartości właściwości (Auto Predykcja).



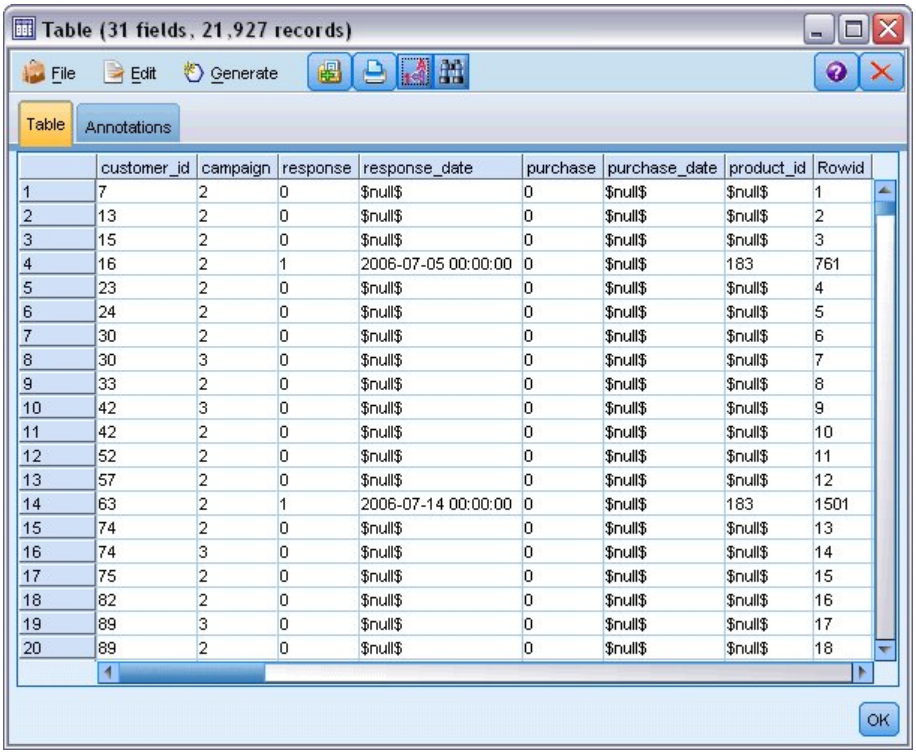
Rysunek 29. Przykładowy strumień węzła Auto Klasyfikacja

W tym przykładzie zastosowano strumień o nazwie *pm_binaryclassifier.str* zainstalowany w folderze Demo w podfolderze *streams*. Używany plik danych to *pm_customer_train1.sav*. Więcej informacji można znaleźć w temacie “Dane historyczne”.

Dane historyczne

Plik *pm_customer_train1.sav* zawiera dane historyczne śledzące oferty złożone określonym klientom w minionych kampaniach, zgodnie ze wskazaniem wartości w polu *campaign*. Największa liczba rekordów przypada w kampanii *Premium account*.

Wartości zmiennej *campaign* są w rzeczywistości zakodowane jako liczby całkowite w danych (na przykład *2 = Premium account*). W dalszej części użytkownik zdefiniuje etykiety dla tych wartości, których można używać, aby zapewnić bardziej wartościowe wyniki.



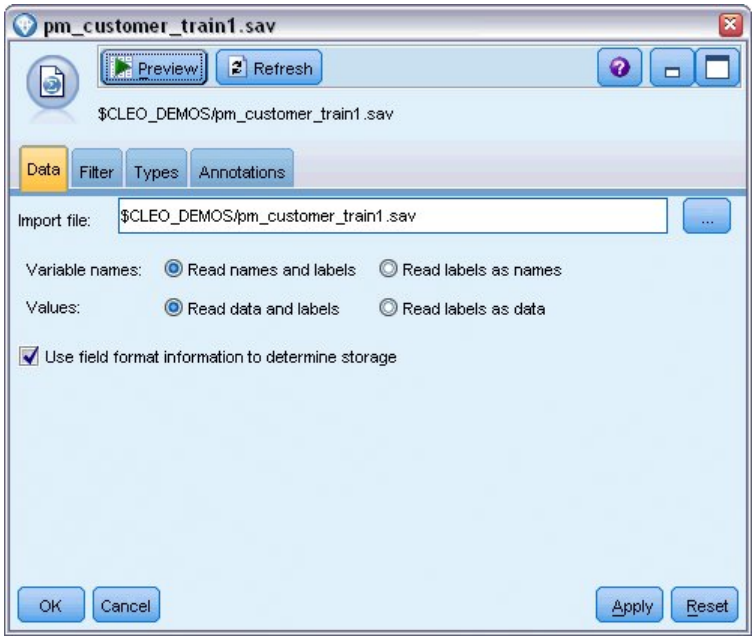
	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7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

Rysunek 30. Dane dotyczące wcześniejszych promocji

Plik zawiera również zmienną *response*, która wskazuje, czy oferta została zaakceptowana (0 = *nie* i 1 = *tak*). Będzie to **zmienna przewidywana** lub wartość, którą chcesz przewidzieć. Uwzględnionych jest też kilka zmiennych zawierających informacje demograficzne i finansowe o każdym z klientów. Zmiennych tych można użyć do zbudowania lub „uczenia" modelu, który przewiduje wskaźniki odpowiedzi dla osób lub grup na podstawie takich cech jak dochód, wiek i liczba transakcji na miesiąc.

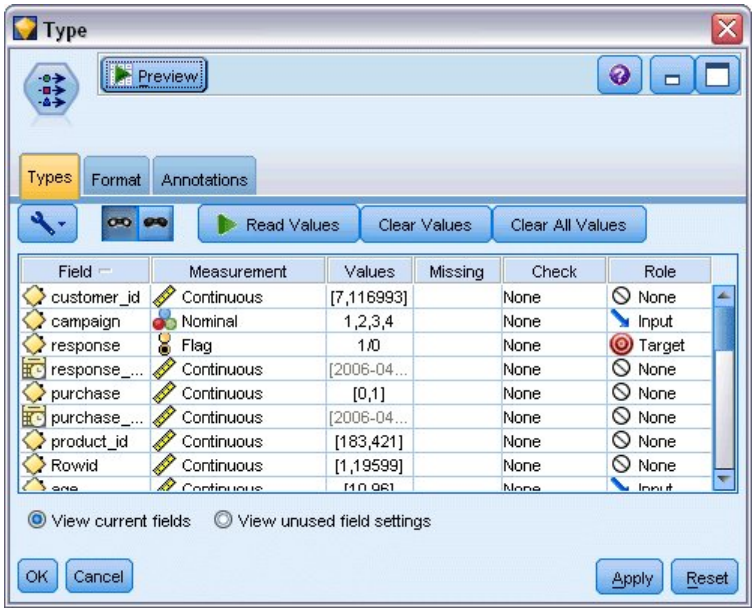
Tworzenie strumienia

1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *pm_customer_train1.sav* znajdujący się w folderze *Demos* w folderze instalacji IBM SPSS Modeler. (W ścieżce do pliku można określić parametr $CLEO_DEMOS/ jako skrót do tego folderu. Uwaga: W ścieżce należy używać ukośników, a nie ukośników odwrotnych, jak pokazano w przykładzie).



Rysunek 31. Odczytywanie danych

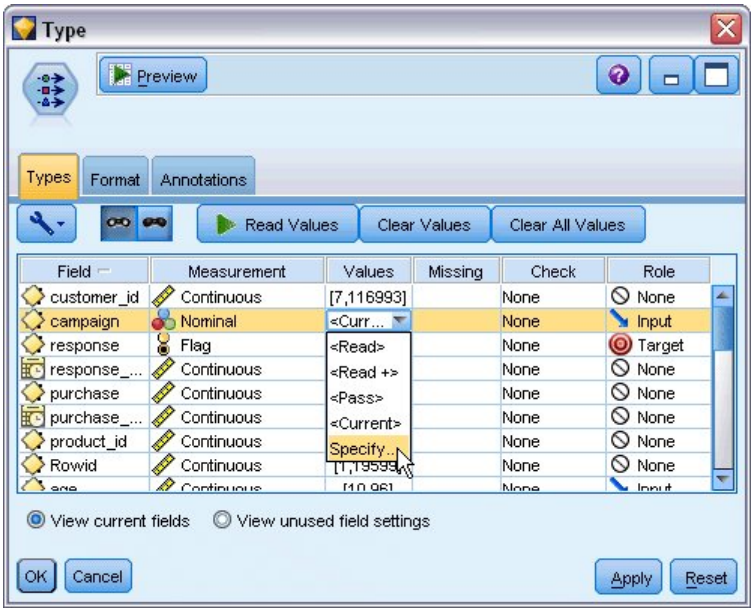
2. Dodaj węzeł typu i wybierz zmienną *response* jako zmienną przewidywaną (Rola = **Przewidywana**). Pozycję Pomiar dla tej zmiennej ustaw na **Flaga**.



Rysunek 32. Ustawianie poziomu pomiaru i roli

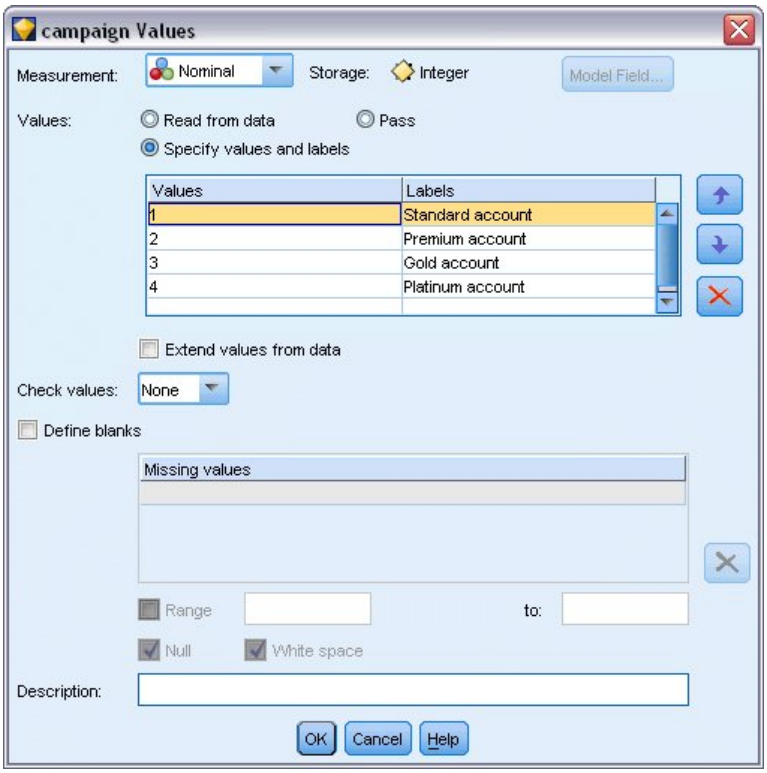
3. Ustaw rolę na **Brak** dla następujących zmiennych: *customer_id*, *campaign*, *response_date*, *purchase*, *purchase_date*, *product_id*, *Rowid* i *X_random*. Te zmienne będą ignorowane podczas budowania modelu.
4. Kliknij przycisk **Odczytaj wartości** w węźle Typ, aby zapewnić, że wartości są określone.

   Jak widzieliśmy wcześniej, dane źródłowe zawierają informacje o czterech różnych kampaniach, z których każda skierowana jest do innego typu konta klienta. Te kampanie są kodowane w danych jako liczby całkowite, więc aby ułatwić zidentyfikowanie, który typ konta reprezentują liczby całkowite, zdefiniujemy dla nich etykiety.



Rysunek 33. Określanie wartości dla zmiennych

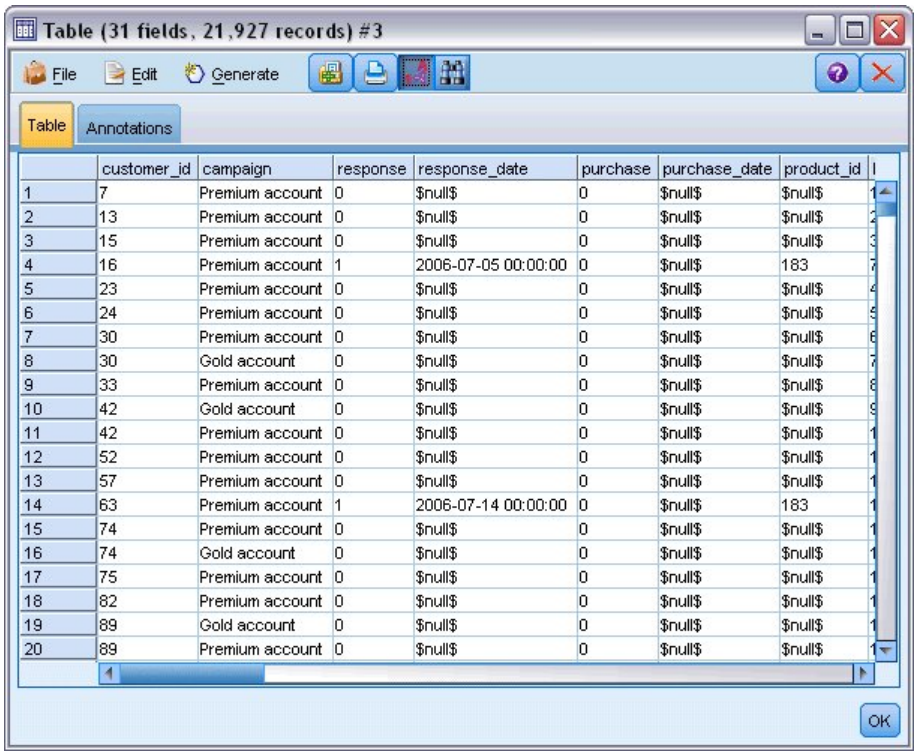
5. W wierszu dla zmiennej **campaign** kliknij wpis w kolumnie **Wartości**.
6. Z listy rozwijanej wybierz pozycję **Określ**.



Rysunek 34. Definiowanie etykiet dla wartości zmiennej

7. W kolumnie **Etykiety** wpisz etykiety dla każdej z czterech wartości zmiennej **campaign**, jak pokazano na rysunku.
8. Kliknij przycisk **OK**.

Teraz w oknach wyników można wyświetlać etykiety zamiast liczb całkowitych.

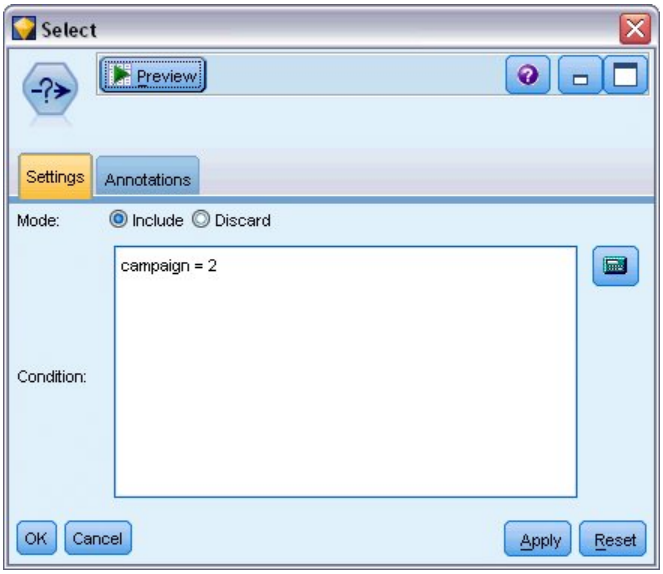


	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	l
1	7	Premium account	0	$null$	0	$null$	$null$	1
2	13	Premium account	0	$null$	0	$null$	$null$	2
3	15	Premium account	0	$null$	0	$null$	$null$	3
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183	7
5	23	Premium account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

Rysunek 35. Wyświetlanie etykiet wartości zmiennej

9. Dołącz węzeł tabeli do węzła typu.
10. Otwórz węzeł Tabela i kliknij przycisk **Uruchom**.
11. W oknie wyników kliknij przycisk paska narzędzi **Wyświetl etykiety zmiennej i wartości**, aby wyświetlić etykiety.
12. Kliknij przycisk **OK**, aby zamknąć okno wyników.

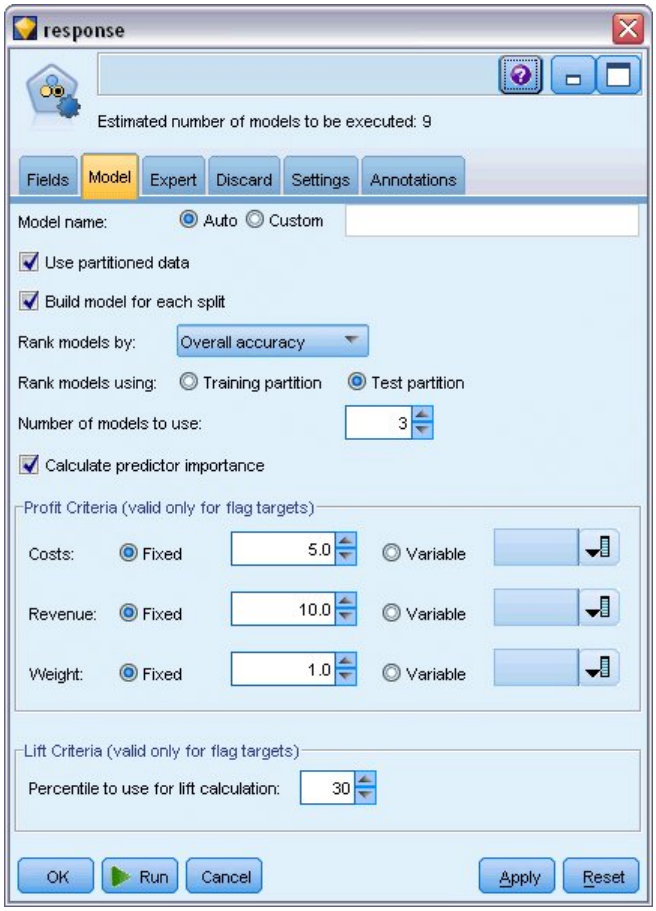
Mimo że dane zawierają informacje o czterech różnych kampaniach, każdorazowo skoncentrujemy analizę na jednej kampanii. Ponieważ największa liczba rekordów przypada dla kampanii Premium account (z kodem w danych *campaign=2*), można użyć węzła selekcji, aby uwzględnić tylko te rekordy w strumieniu.



Rysunek 36. Wybieranie rekordów dla pojedynczej kampanii

Generowanie i porównywanie modeli

1. Dołącz węzeł Auto Klasyfikacja i wybierz **Ogólna dokładność** jako metrykę używaną do uszeregowania modeli.
2. Ustaw wartość **Liczba modeli do wykorzystania** na 3. Oznacza to, że po wykonaniu węzła utworzone zostaną trzy najlepsze modele.

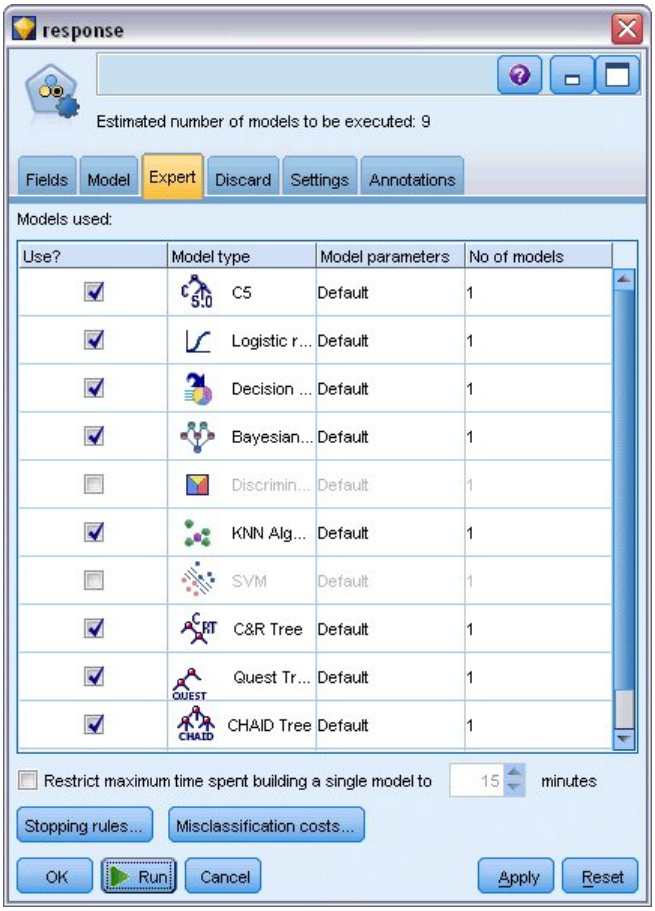


Rysunek 37. Karta Model węzła Auto Klasyfikacja

Na karcie Zaawansowany można wybrać do 11 różnych algorytmów modeli.

3. Usuń zaznaczenie typów modeli **Analiza dyskryminacyjna** i **SVM**. (Nauka tych modeli dla określonych danych zajmuje więcej czasu, więc usunięcie zaznaczenia spowoduje przyspieszenie przykładu. Jeśli czas nie jest taki ważny, możesz pozostawić te opcje zaznaczone).

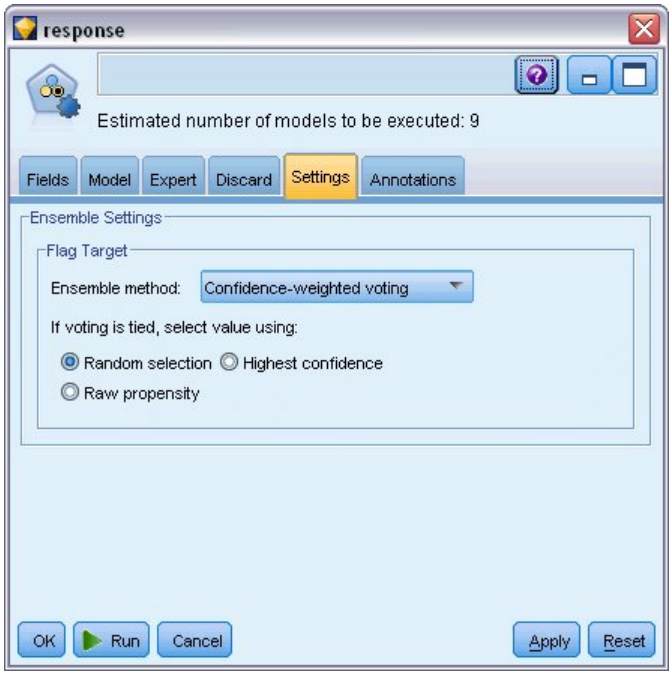
   Ponieważ na karcie Model ustawiono wartość 3 dla parametru **Liczba modeli do wykorzystania**, węzeł obliczy dokładność pozostałych dziewięciu algorytmów i utworzy pojedynczy model użytkowy łączący trzy najdokładniejsze algorytmy.



Rysunek 38. Karta Zaawansowany węzła Auto Klasyfikacja

4. Na karcie Ustawienia dla metody zespolenia wybierz opcję **Głosowanie ważone ufnością**. Określa to, w jaki sposób tworzona jest zagregowana ocena dla każdego rekordu.

   Przy prostym głosowaniu, jeśli dwa z trzech modeli przewidują wartość *tak*, to *tak* wygrywa głosowaniem 2 do 1. Przy głosowaniu ważonym ufnością głosy są ważone na podstawie wartości ufności dla każdej predykcji. Dlatego też, jeśli jeden model przewiduje *nie* z wyższą ufnością niż pozostałe predykcje *tak* łącznie, to wartość *nie* wygrywa.



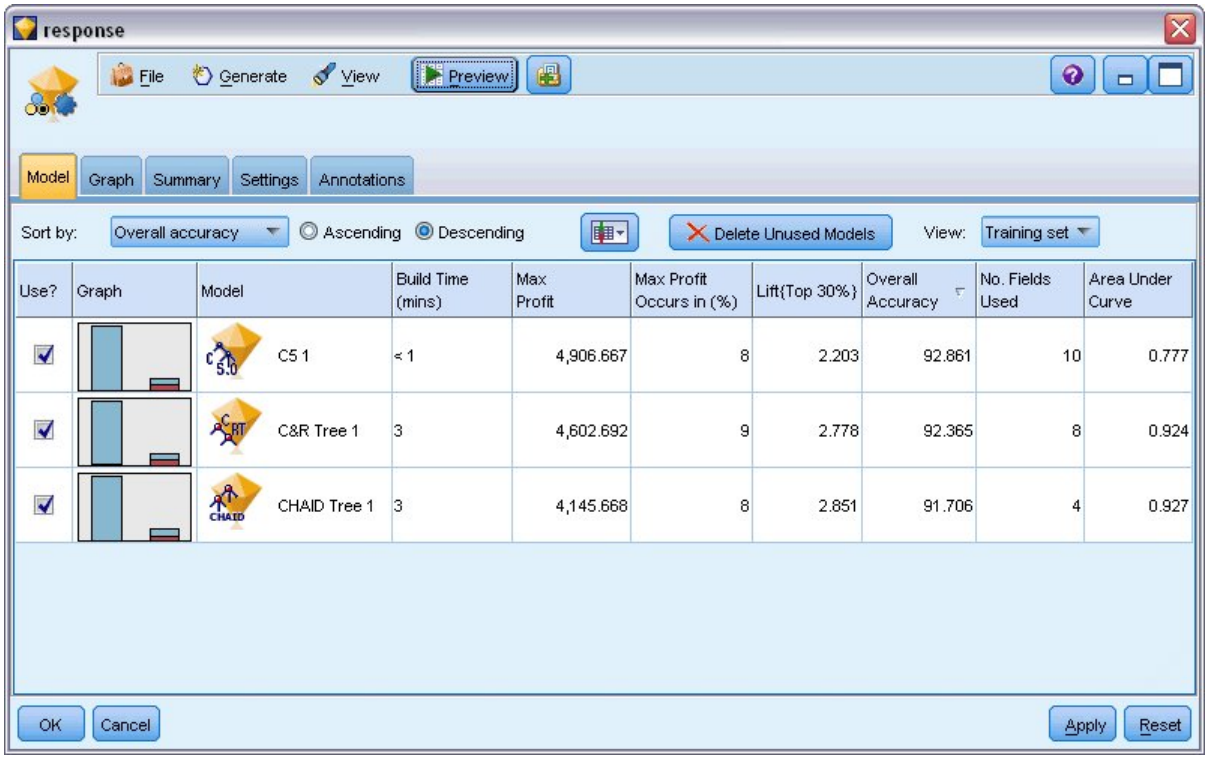
Rysunek 39. Karta Ustawienia węzła Auto Klasyfikacja

5. Kliknij przycisk **Uruchom**.

Po kilku minutach wygenerowany model użytkowy zostaje utworzony i umieszczony w obszarze roboczym oraz na palecie modeli w prawym górnym rogu okna. Model użytkowy można przeglądać lub zapisać, lub wdrożyć na wiele innych sposobów.

Otwórz model użytkowy. Wymienia on szczegóły każdego z modeli utworzonych podczas uruchomienia. (W rzeczywistej sytuacji, w której można utworzyć setki modeli dla dużych zbiorów danych, może to zająć wiele godzin). Patrz Rys. 29 na stronie 37.

Jeśli chcesz dalej eksplorować indywidualny model, możesz dwukrotnie kliknąć ikonę modelu użytkowego w kolumnie **Model**, aby zejść niżej i przeglądać wyniki indywidualnego modelu. Z tego poziomu można wygenerować węzły modelowania, modele użytkowe lub wykresy ewaluacyjne. W kolumnie **Wykres** w celu wygenerowania wykresu w pełnym wymiarze można kliknąć dwukrotnie ikonę.



Rysunek 40. Wyniki węzła Auto Klasyfikacja

Domyślnie modele są sortowane na podstawie ogólnej dokładności, ponieważ taką miarę wybrano na karcie Model węzła Auto Klasyfikacja. Model C51 ma najlepsze wyniki według tej miary, ale modele C&R Tree i CHAID są prawie tak samo dokładne.

Możliwe jest sortowanie różnych kolumn, klikając nagłówek dla tej kolumny lub można wybrać jedną z miar z listy rozwijanej **Sortuj według** na pasku narzędzi.

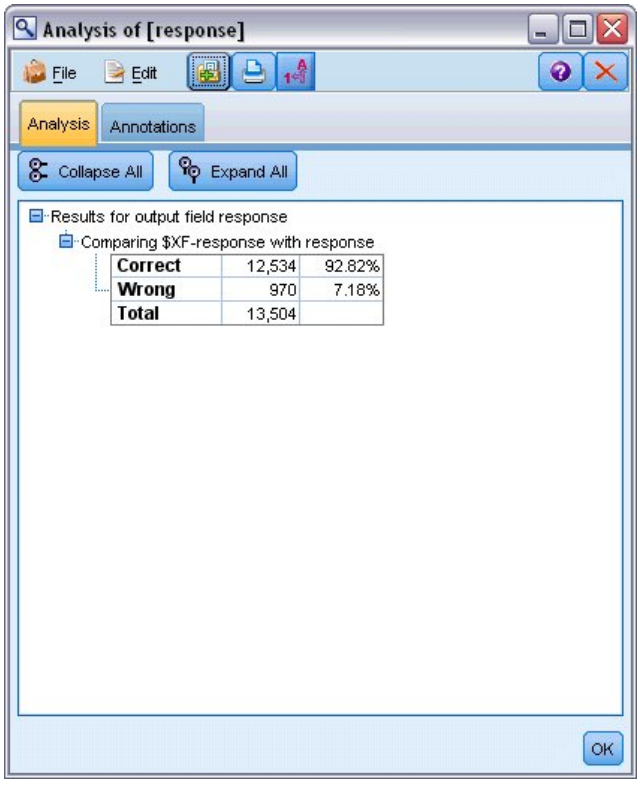
Na podstawie tych wyników użytkownik decyduje o użyciu trzech najdokładniejszych modeli. Połączenie predykcji z wielu modeli umożliwia obejście ograniczeń w poszczególnych modelach, co często powoduje wyższą ogólną dokładność.

W kolumnie **Wykorzystanie** wybierz modele C51, C&R Tree i CHAID.

Dołącz węzeł analizy (paleta Wynik) za modelem użytkowym. Prawym przyciskiem myszy kliknij węzeł analizy i wybierz opcję **Uruchom**, aby uruchomić strumień.

Zagregowana ocena wygenerowana przez model zespolony przedstawiona jest w zmiennej o nazwie *$XF-response*. Przy pomiarze dla danych uczących przewidywana wartość odpowiada rzeczywistej odpowiedzi (zarejestrowanej w oryginalnej zmiennej *response*) z ogólną dokładnością wynoszącą 92,82%.

Mimo że wynik nie jest tak dokładny jak najlepszy z trzech modeli w tym przypadku (92,86% dla C51), różnica jest zbyt mała, aby była istotna. Zasadniczo model zespolony zazwyczaj będzie działał lepiej przy zastosowaniu dla zbiorów danych innych niż dane uczące.



Rysunek 41. Analiza trzech zespolonych modeli

Podsumowanie

Podsumowując, użytkownik użył węzła Auto Klasyfikacja do porównania kilku różnych modeli, użył trzech najdokładniejszych modeli i dodał je do strumienia w zespolonym modelu użytkowym Auto Klasyfikacja.

- Biorąc pod uwagę ogólną dokładność, najlepsze wyniki dla danych uczących osiągnęły modele C51, Drzewo C&R i CHAID.
- Model zespolony pokazał wydajność prawie tak dobrą jak w przypadku najlepszych indywidualnych modeli i może działać lepiej przy zastosowaniu na innych zbiorach danych. Jeśli celem jest jak największa automatyzacja procesu, to podejście pozwala na osiągnięcie wydajnego modelu w większości wypadków bez potrzeby zbytniego zagłębiania się w specyfikę modelu.

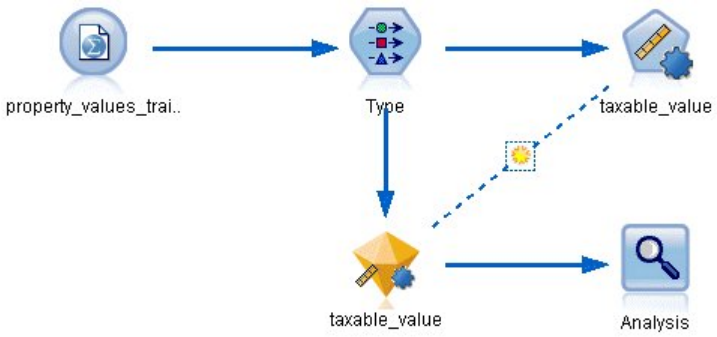
Rozdział 5. Zautomatyzowane modelowanie dla docelowej wartości ilościowej

Wartości właściwości (Auto Predykcja)

Węzeł Auto Predykcja pozwala na automatyczne tworzenie i porównywanie różnych modeli dla wyników ilościowych (zakresów numerycznych), takich jak przewidywanie wartości nieruchomości podlegającej opodatkowaniu. Za pomocą pojedynczego węzła można oszacować i porównać zestaw modeli kandydackich i wygenerować podzbiór modeli do dalszej analizy. Węzeł działa w taki sam sposób, jak węzeł Auto Klasyfikacja, ale raczej dla ilościowych zmiennych przewidywanych, a nie flag lub nominalnych zmiennych przewidywanych.

Węzeł łączy najlepsze modele kandydackie w pojedynczy zagregowany (zespolony) model użytkowy. Takie podejście łączy łatwość automatyzacji z korzyściami łączenia wielu modeli, które często zwracają dokładniejsze predykcje, niż można uzyskać z jednego modelu.

Ten przykład koncentruje się na fikcyjnej gminie, która jest odpowiedzialna za dostosowanie i ocenę podatku od nieruchomości. Aby wykonać tę czynność bardzo dokładnie, stworzony zostanie model, który przewiduje wartość nieruchomości w oparciu o rodzaj budynku, sąsiedztwo, wielkość i inne znane czynniki.



Rysunek 42. Przykładowy strumień węzła Auto Predykcja

W tym przykładzie zastosowano strumień o nazwie *property_values_numericpredictor.str* zainstalowany w folderze Demos w podfolderze *streams*. Używany plik danych to *property_values_train.sav.* Więcej informacji można znaleźć w temacie “Folder Demos” na stronie 5.

Dane uczące

Plik danych zawiera zmienną o nazwie *taxable_value*, która jest **zmienną przewidywaną** lub wartością, którą chcesz przewidzieć. Inne zmienne zawierają informacje, takie jak sąsiedztwo, rodzaj budynku oraz powierzchnia wewnętrzna, które mogą być używane jako predyktory.

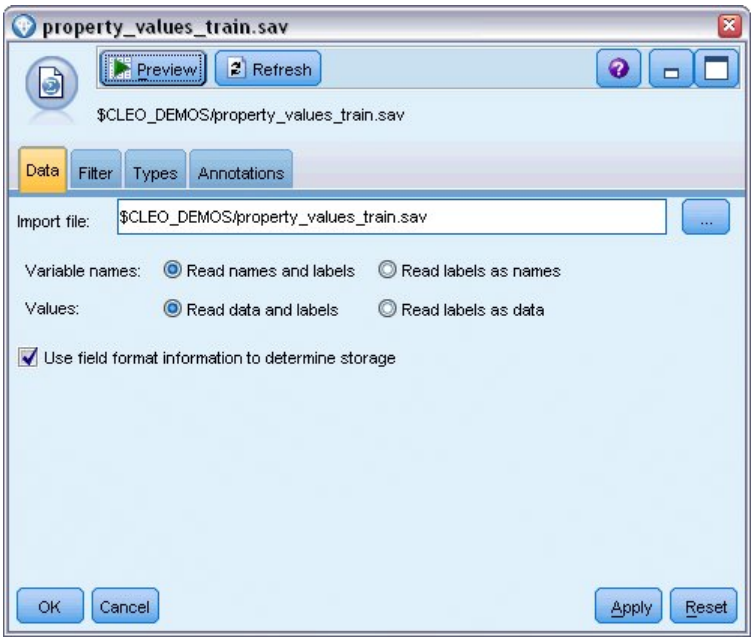
Nazwa zmiennej	Etykieta
property_id	Identyfikator nieruchomości
sąsiedztwo	Obszar w mieście
building_type	Rodzaj budynku
year_built	Rok budowy
volume_interior	Powierzchnia wewnętrzna
volume_other	Powierzchnia garażu i budynków dodatkowych

Nazwa zmiennej	Etykieta
lot_size	Wielkość działki
taxable_value	Wartość podlegająca opodatkowaniu

W folderze Demos znajduje się również plik oceniania danych *property_values_score.sav*. Zawiera te same zmienne, ale bez zmiennej *taxable_value*. Po nauczeniu modeli używających zbioru danych, w którym wartość podlegająca opodatkowaniu jest znana, można ocenić rekordy, w których ta wartość nie jest jeszcze znana.

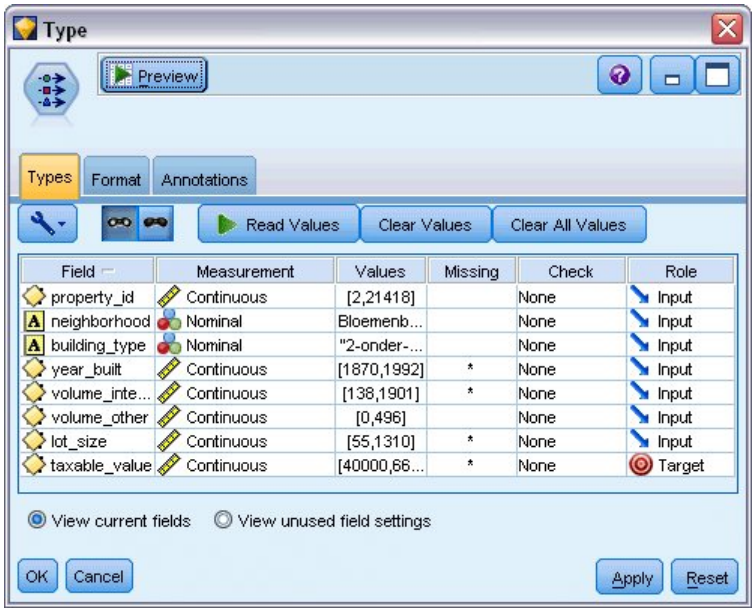
Tworzenie strumienia

1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *property_values_train.sav* znajdujący się w folderze *Demos* w folderze instalacji IBM SPSS Modeler. (W ścieżce do pliku można określić parametr $CLEO_DEMOS/ jako skrót do tego folderu. Uwaga: w ścieżce należy używać ukośników, a nie ukośników odwrotnych, jak pokazano w przykładzie.)



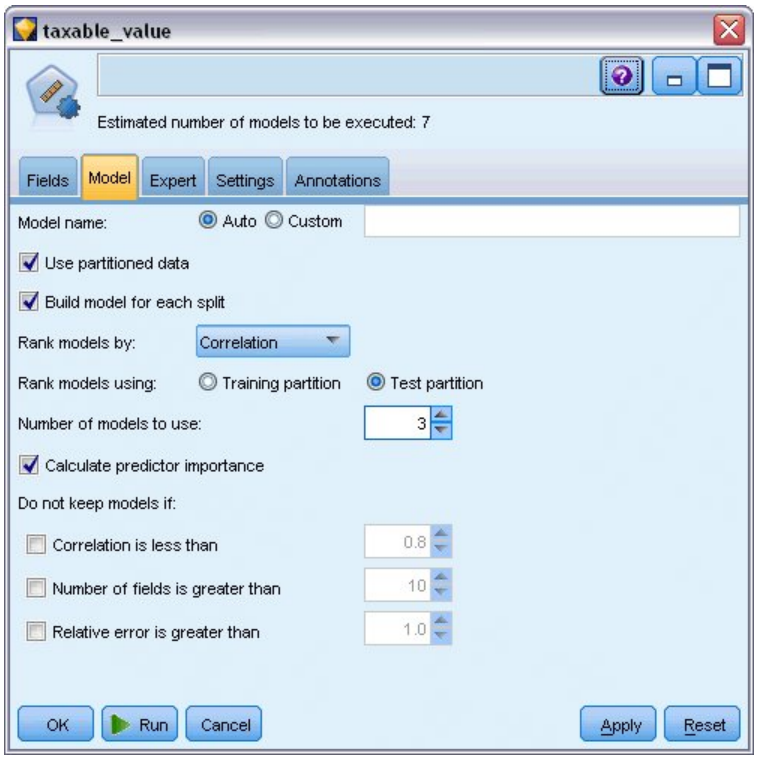
Rysunek 43. Odczytywanie danych

2. Dodaj węzeł typu i wybierz zmienną *taxable_value* jako zmienną przewidywaną (Rola = **Przewidywana**). Dla wszystkich innych zmiennych Rola powinna być ustawiona jako **Dane wejściowe**, wskazując, że zmienne te będą używane jako predyktory.



Rysunek 44. Ustawianie zmiennej przewidywanej

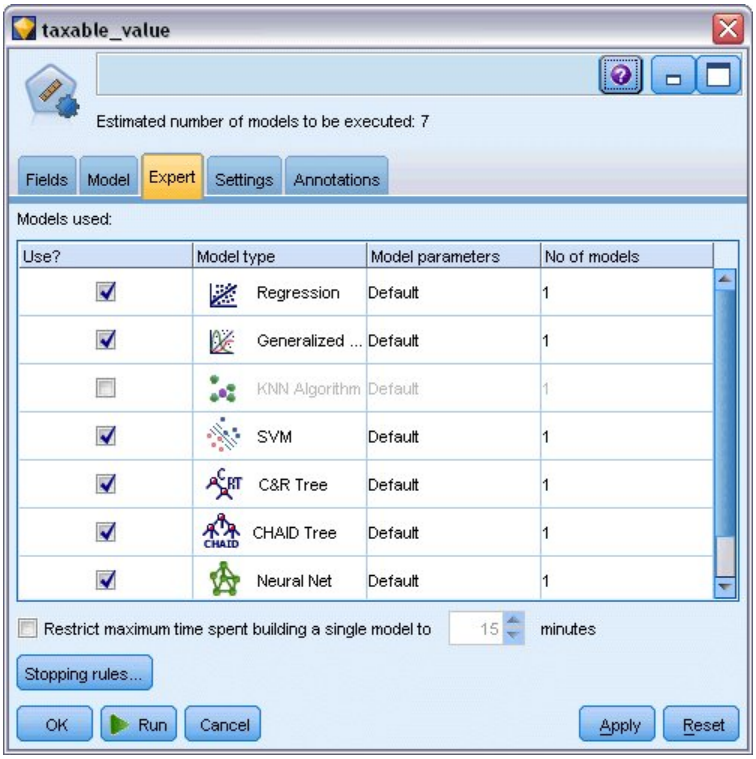
3. Dołącz węzeł Auto Predykcja i wybierz **Korelacja** jako metrykę używaną do uszeregowania modeli.
4. Ustaw wartość **Liczba modeli do wykorzystania** na 3. Oznacza to, że po wykonaniu węzła utworzone zostaną trzy najlepsze modele.



Rysunek 45. Karta Model węzła Auto Predykcja

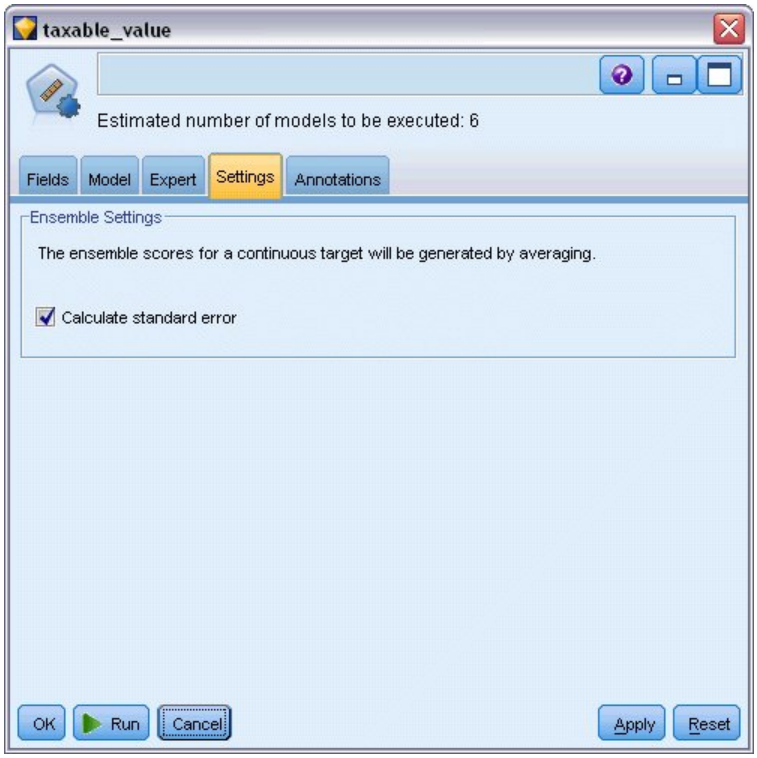
5. Na karcie Zaawansowany pozostaw domyślne ustawienia. Węzeł oszacuje pojedynczy model dla każdego algorytmu, w sumie dla siedmiu modeli. (Można również zmodyfikować te ustawienia, aby porównać wiele wariantów dla każdego typu modelu).

Ponieważ na karcie Model ustawiono wartość 3 dla parametru **Liczba modeli do wykorzystania**, węzeł obliczy dokładność siedmiu algorytmów i utworzy pojedynczy model użytkowy łączący trzy najdokładniejsze algorytmy.



Rysunek 46. Karta Zaawansowany węzła Auto Predykcja

6. Na karcie Ustawienia pozostaw ustawienia domyślne. Ponieważ jest to docelowa wartość ilościowa, ocena zespolona jest generowana przez uśrednienie ocen indywidualnych modeli.



Rysunek 47. Karta Ustawienia węzła Auto Predykcja

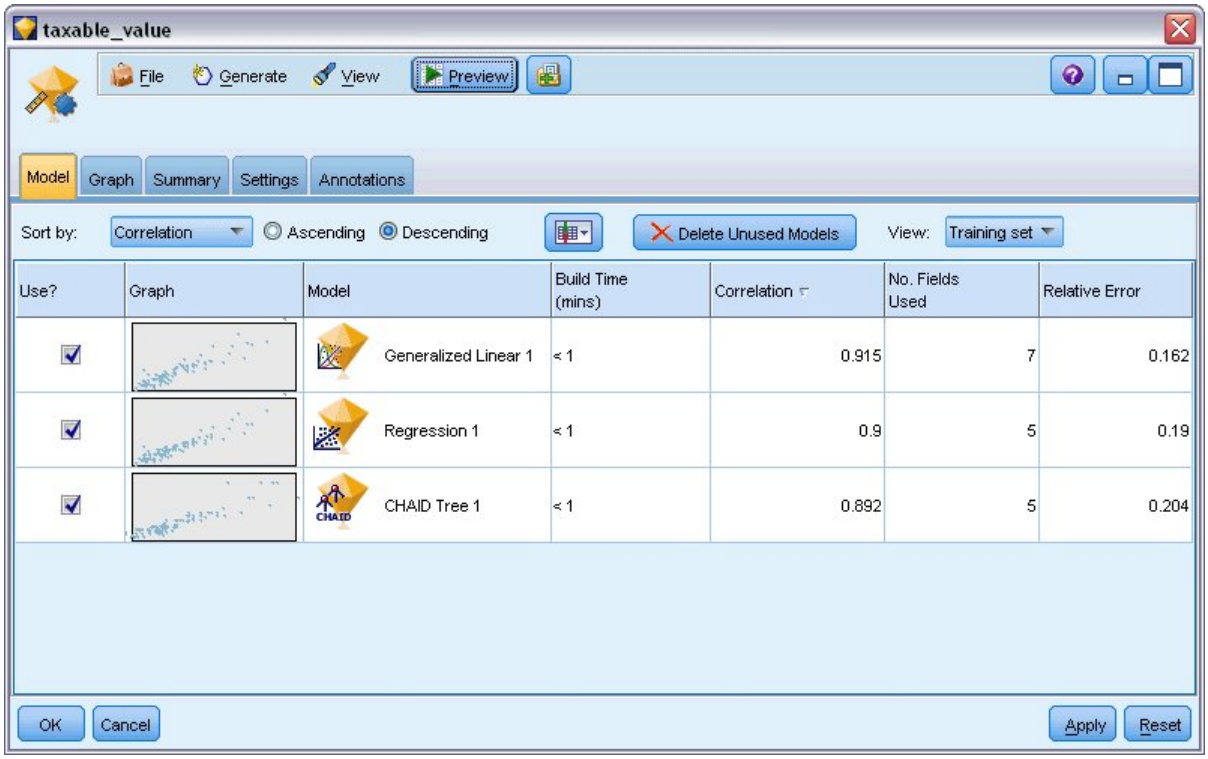
Porównywanie modeli

1. Kliknij przycisk Uruchom.

Model użytkowy zostaje utworzony i umieszczony w obszarze roboczym oraz na palecie modeli w prawym górnym rogu okna. Model użytkowy można przeglądać lub zapisać, lub wdrożyć na wiele innych sposobów.

Otwórz model użytkowy. Wymienia on szczegóły każdego z modeli utworzonych podczas uruchomienia. (W rzeczywistej sytuacji, w której setki modeli wykonują oszacowania dla dużych zbiorów danych, może to zająć wiele godzin). Patrz Rys. 42 na stronie 49.

Jeśli chcesz dalej eksplorować indywidualny model, możesz dwukrotnie kliknąć ikonę modelu użytkowego w kolumnie **Model**, aby zejść niżej i przeglądać wyniki indywidualnego modelu. Z tego poziomu można wygenerować węzły modelowania, modele użytkowe lub wykresy ewaluacyjne.



Rysunek 48. Wyniki węzła Auto Predykcja

Domyślnie modele są sortowane na podstawie korelacji, ponieważ taką miarę wybrano w węźle Auto Predykcja. Do celu rangowania używana jest wartość bezwzględna korelacji, przy wartościach bliższych 1 wskazujących na silne relacje. Uogólniony model liniowy jest najlepszy według tej miary, ale kilka innych ma zbliżoną dokładność. Uogólniony model liniowy ma również najniższy błąd względny.

Możliwe jest sortowanie różnych kolumn, klikając nagłówek dla tej kolumny lub można wybrać jedną z miar z listy **Sortuj według** na pasku narzędzi.

Każdy rysunek przedstawia wykres obserwowanych wartości w porównaniu do przewidywanych wartości dla modelu, zapewniając szybki wizualny wskaźnik korelacji pomiędzy nimi. W dobrym modelu punkty powinny być skupione wzdłuż przekątnej, co jest spełnione dla wszystkich modeli w tym przykładzie.

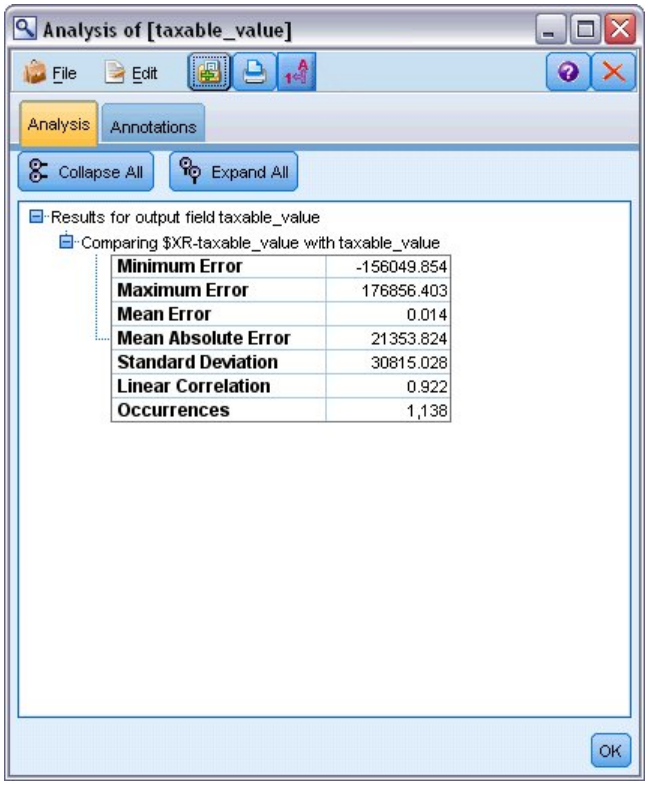
W kolumnie **Wykres** w celu wygenerowania wykresu w pełnym wymiarze można kliknąć dwukrotnie ikonę.

Na podstawie tych wyników użytkownik decyduje o użyciu trzech najdokładniejszych modeli. Połączenie predykcji z wielu modeli umożliwia obejście ograniczeń w poszczególnych modelach, co często powoduje wyższą ogólną dokładność.

Należy upewnić się, czy w kolumnie **Wykorzystanie** wybrane są wszystkie trzy modele.

Dołącz węzeł analizy (paleta Wynik) za modelem użytkowym. Prawym przyciskiem myszy kliknij węzeł analizy i wybierz opcję **Uruchom**, aby uruchomić strumień.

Uśredniona ocena wygenerowana przez model zespolony jest dodawana w zmiennej o nazwie *$XR-taxable_value* i ma korelację wynoszącą 0,922, która jest wyższa niż wartości korelacji trzech indywidualnych modeli. Oceny zespolone wykazują również niski średni błąd bezwzględny i mogą działać lepiej niż indywidualne modele przy zastosowaniu na innych zbiorach danych.



Rysunek 49. Przykładowy strumień węzła Auto Predykcja

Podsumowanie

Podsumowując, użytkownik użył węzła Auto Predykcja do porównania kilku różnych modeli, wybrał trzy najdokładniejsze modele i dodał je do strumienia w zespolonym modelu użytkowym Auto Predykcja.

- Biorąc pod uwagę ogólną dokładność, najlepsze wyniki dla danych uczących osiągnęły modele Uogólniony model liniowy, Regresja i CHAID.
- Model zespolony pokazał wydajność, która była lepsza niż w przypadku dwóch indywidualnych modeli i może działać lepiej przy zastosowaniu na innych zbiorach danych. Jeśli celem jest jak największa automatyzacja procesu, to podejście pozwala na osiągnięcie wydajnego modelu w większości wypadków bez potrzeby zbytniego zagłębiania się w specyfikę modelu.

Rozdział 6. Automatyczne przygotowywanie danych (Automated Data Preparation — ADP)

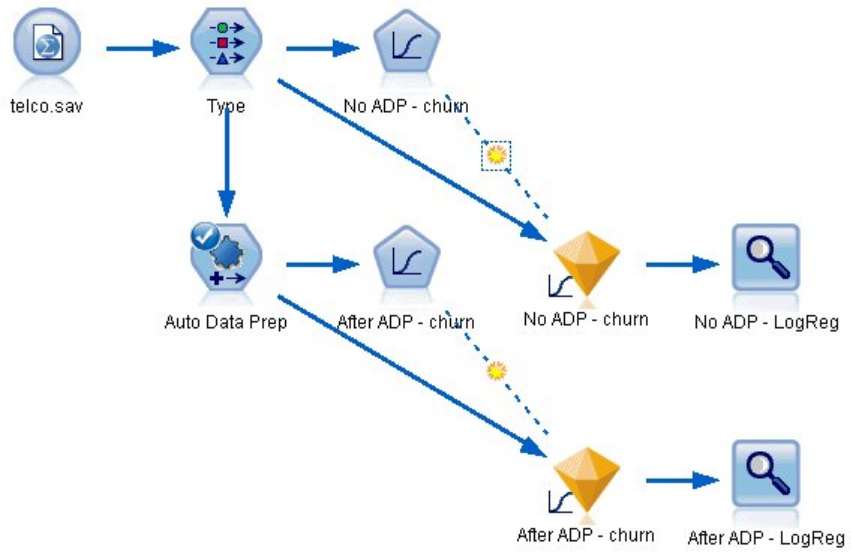
Przygotowanie danych do analizy jest jednym z najważniejszych kroków każdego projektu eksploracji danych i tradycyjnie zajmuje najwięcej czasu. Węzeł Automatyczne przygotowanie danych (ADP) obsługuje to zadania za użytkownika przez analizę danych i identyfikację poprawek, klasyfikację zmiennych, które są problematyczne lub mają małe prawdopodobieństwo bycia użytecznymi, w razie potrzeby obliczanie nowych atrybutów i zwiększanie wydajności poprzez wykorzystywanie inteligentnych technik klasyfikowania. Tego węzła można używać w sposób w pełni automatyczny, pozwalając mu na wybór i zastosowanie poprawek, lub przeglądać zmiany przed ich dokonaniem i akceptować je lub odrzucać.

Używanie węzła ADP pozwala na szybkie przygotowanie do eksploracji danych bez potrzeby wcześniejszej znajomości stosowanych koncepcji statystycznych. Jeśli uruchamiasz węzeł z ustawieniami domyślnymi, modele będą budowane i oceniane szybciej.

Ten przykład używa strumienia o nazwie *ADP_basic_demo.str*, który odwołuje się do pliku danych o nazwie *telco.sav*, aby zademonstrować zwiększoną dokładność, którą można uzyskać, używając domyślnych ustawień węzła ADP podczas budowania modeli. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *ADP_basic_demo.str* znajduje się w katalogu *streams*.

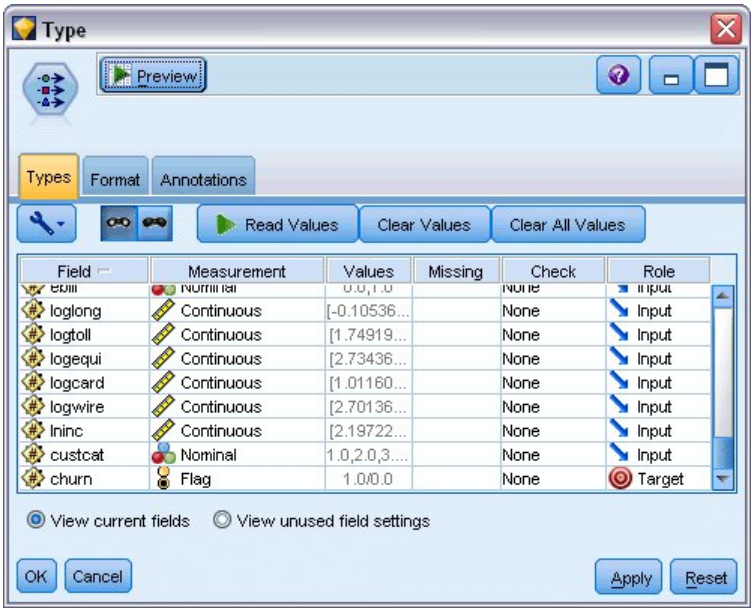
Tworzenie strumienia

1. Aby zbudować strumień, dodaj węzeł źródłowy Plik Statistics wskazujący na plik *telco.sav* znajdujący się w katalogu *Demos* instalacji programu IBM SPSS Modeler.



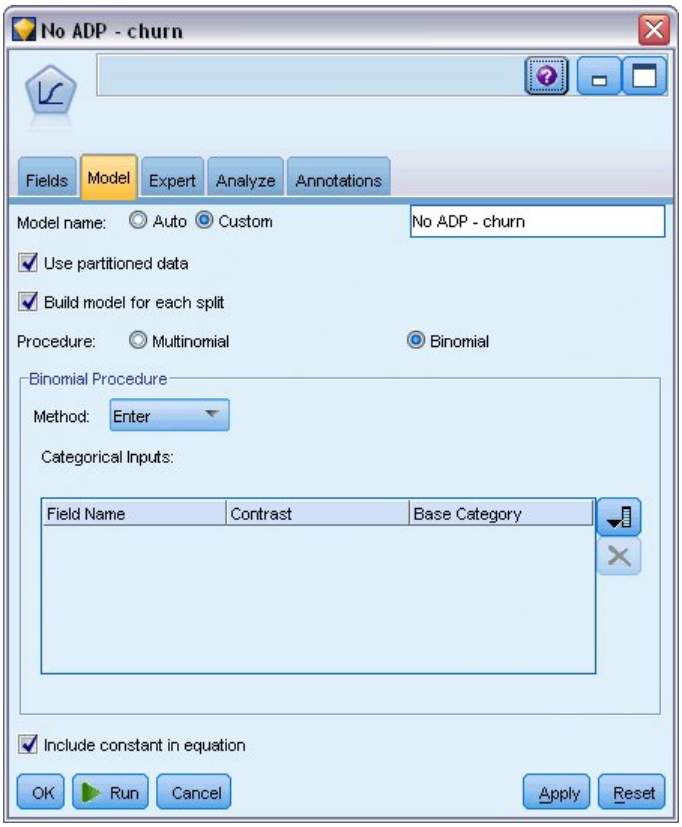
Rysunek 50. Tworzenie strumienia

2. Załącz węzeł typu do węzła źródłowego, ustaw poziom pomiaru dla zmiennej *odchodzenie* na wartość **Flaga** i ustaw rolę na **Przewidywana**. Wszystkie pozostałe zmienne powinny mieć ustawioną rolę na **Dane wejściowe**.



Rysunek 51. Wybieranie zmiennej przewidywanej

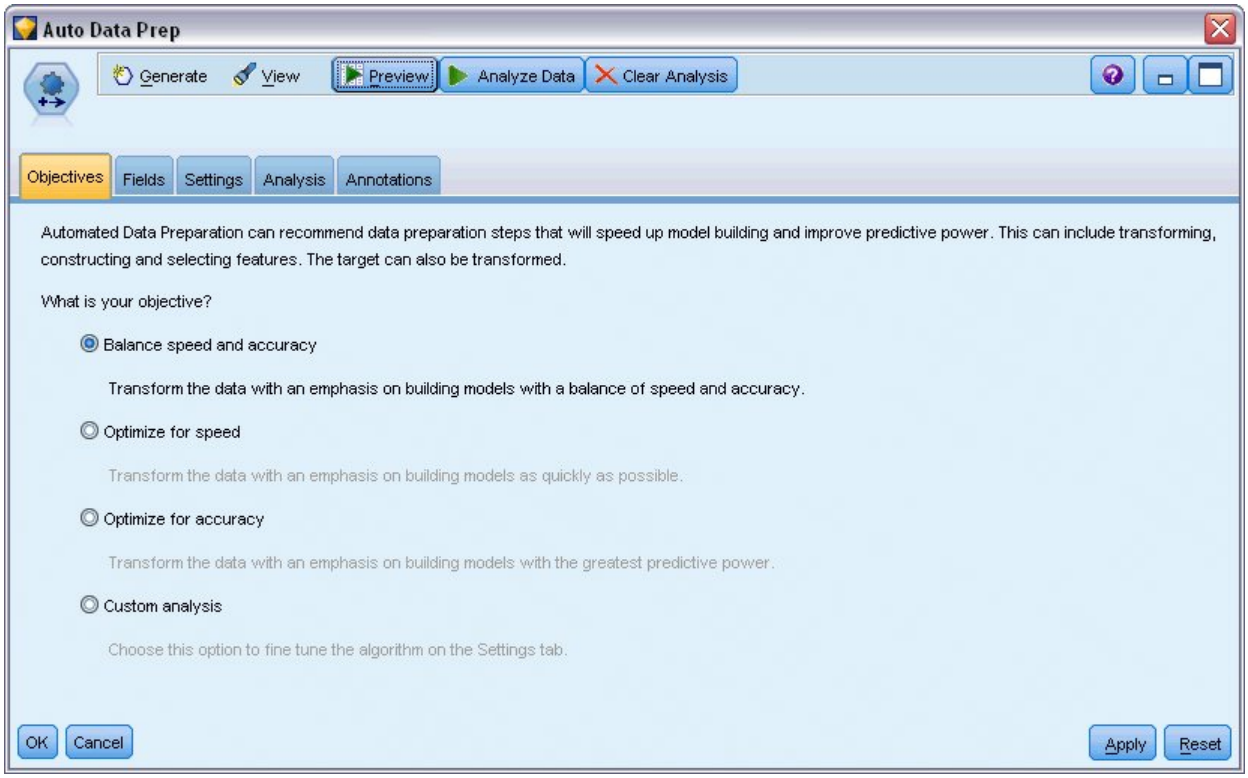
3. Dołącz węzeł logistyczny do węzła typu.
4. W węźle logistycznym kliknij kartę Model i wybierz procedurę **Dwumianowa**. W polu *Nazwa modelu* wybierz pozycję **Użytkownika** i wprowadź Bez ADP - odchodzenie.



Rysunek 52. Wybieranie opcji modelu

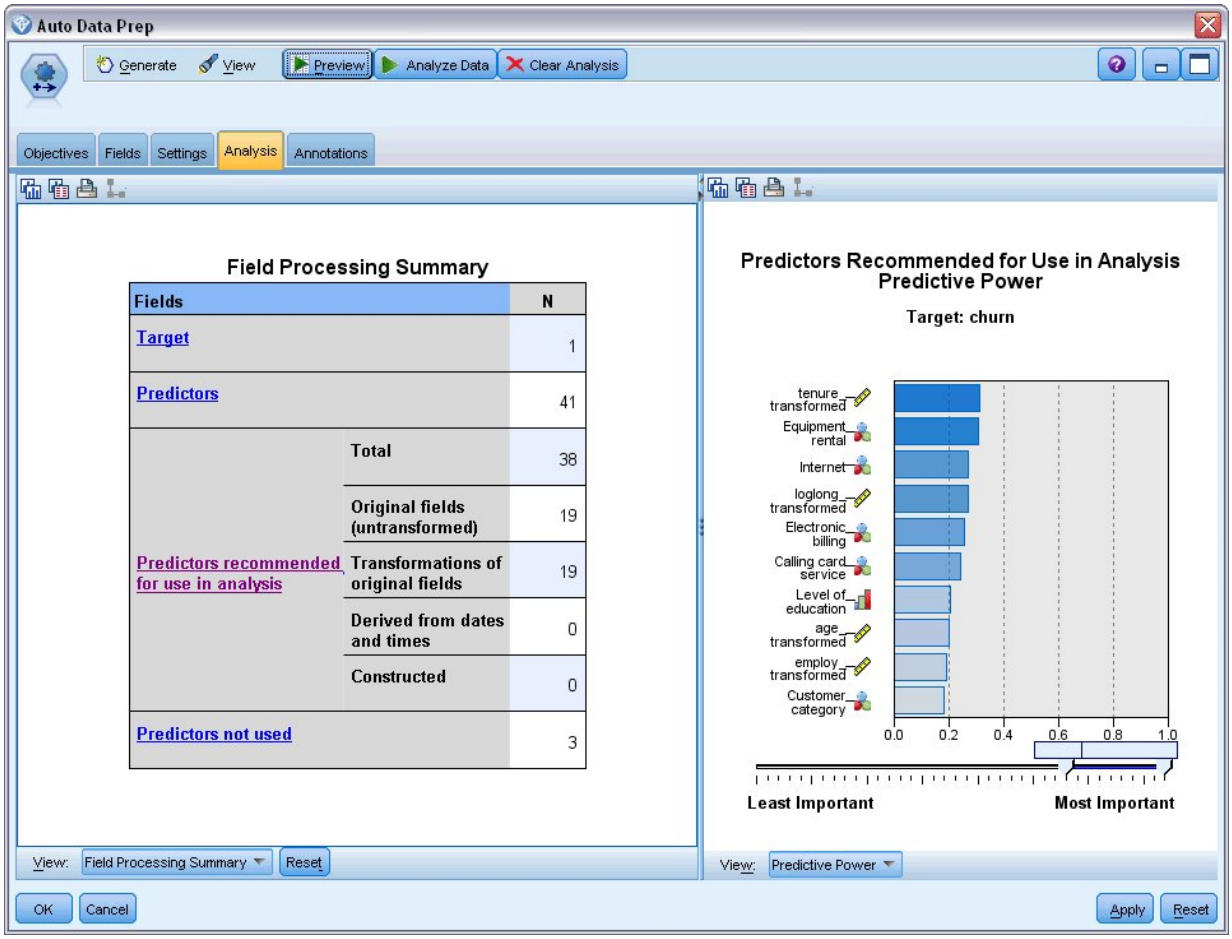
5. Dołącz węzeł ADP do węzła typu. Na karcie Cele pozostaw ustawienia domyślne, aby przeanalizować i przygotować dane, równoważąc szybkość i dokładność.
6. W górnej części karty Cele kliknij przycisk **Analizuj dane**, aby przeanalizować i przetworzyć dane.

   Inne opcje w węźle ADP pozwalają na określenie, aby proces koncentrował się na dokładności lub szybkości. Możliwe jest też dostosowanie wielu kroków przetwarzania podczas przygotowania danych.



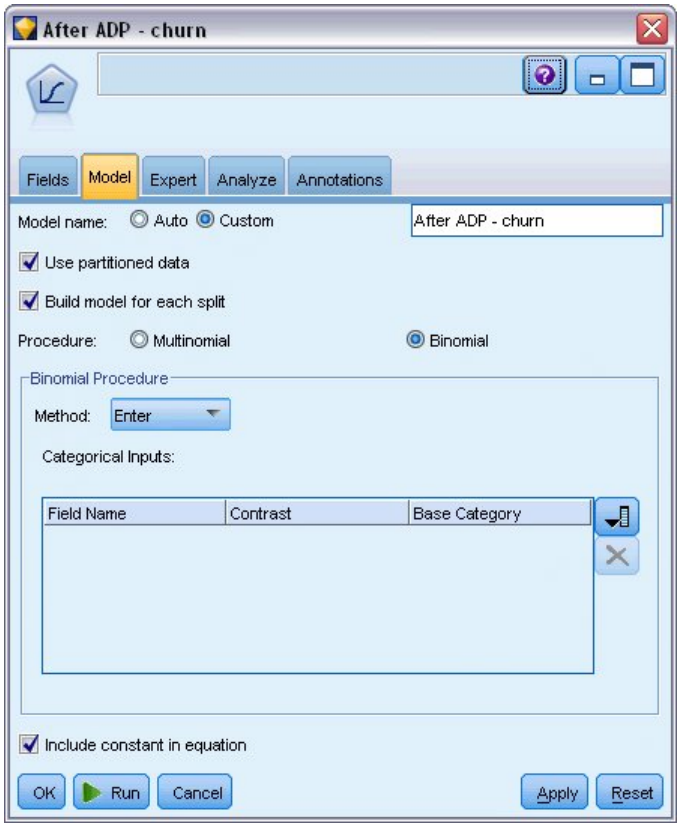
Rysunek 53. Domyślne cele ADP

Wyniki przetwarzania danych są wyświetlane na karcie Analiza. **Podsumowanie przetwarzania zmiennych** pokazuje, że z 41 właściwości danych przekazanych do węzła ADP 19 zostało przekształconych, aby wspierać przetwarzanie, a 3 zostały odrzucone i nie są używane.



Rysunek 54. Podsumowanie przetwarzania danych

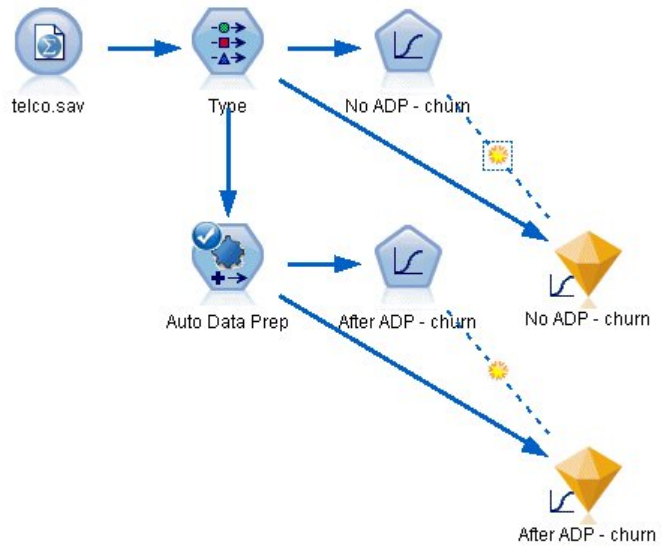
7. Dołącz węzeł logistyczny do węzła ADP.
8. W węźle logistycznym kliknij kartę Model i wybierz procedurę **Dwumianowa**. W polu *Nazwa modelu* wybierz pozycję **Użytkownika** i wprowadź Za ADP - odchodzenie.



Rysunek 55. Wybieranie opcji modelu

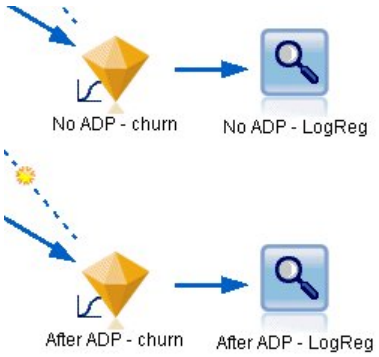
Porównywanie dokładności modeli

1. Uruchom oba węzły logistyczne, aby utworzyć modele użytkowe, które są dodawane do strumienia i do palety modeli w prawym górnym rogu.



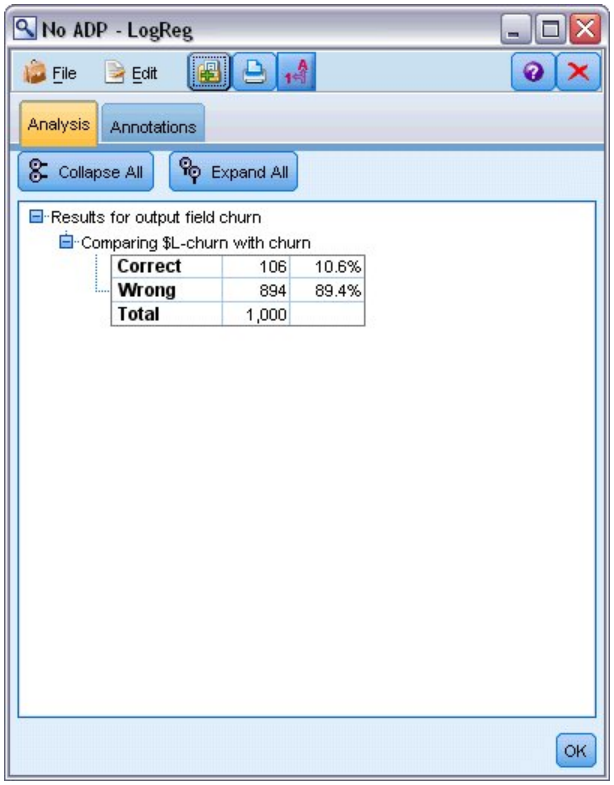
Rysunek 56. Załączanie modeli użytkowych

2. Załącz węzły analizy do modeli użytkowych i uruchom węzły analizy, używając ich domyślnych ustawień.



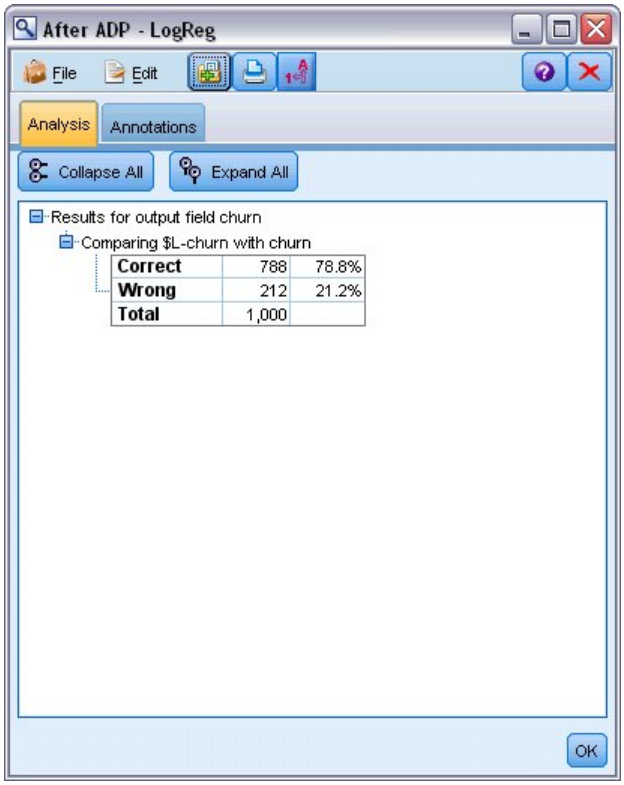
Rysunek 57. Załączanie węzłów analizy

Analiza modelu wyliczonego bez ADP pokazuje, że przepuszczenie danych przez węzeł regresji logistycznej z ustawieniami domyślnymi daje modelowi małą dokładność — tylko 10,6%.



Rysunek 58. Wyniki modelu wyliczonego bez ADP

Analiza modelu wyliczonego na podstawie ADP pokazuje, że przepuszczenie danych z domyślnymi ustawieniami ADP powoduje utworzenie dużo bardziej dokładnego modelu, który jest prawidłowy w 78,8%.



Rysunek 59. Wyniki modelu wyliczonego na podstawie ADP

Podsumowując, uruchamiając tylko węzeł ADP, aby dostosować przetwarzanie danych, możliwe było zbudowanie dużo bardziej dokładnego modelu przy małej bezpośredniej manipulacji danymi.

Oczywiście, jeśli użytkownik chce udowodnić lub obalić określoną teorię lub chce utworzyć konkretny model, korzystna może być praca bezpośrednio na ustawieniach modelu. Dla osób, które dysponują ograniczonym czasem lub muszą przygotować duże ilości danych, użycie węzła ADP może być jednak korzystne.

Wyjaśnienia podstaw matematycznych metod modelowania używanych w programie IBM SPSS Modeler przedstawiono w publikacji *IBM SPSS Modeler Algorithms Guide* dostępnej w katalogu *\Documentation* na dysku instalacyjnym.

Należy zauważyć, że wyniki w tym przykładzie opierają się tylko na danych uczących. Aby ocenić, jak dobrze modele uogólniają inne dane w rzeczywistych przypadkach, należałoby użyć węzła podziału na podzbiory, aby zatrzymać podzbiór rekordów do celów testowania i walidacji.

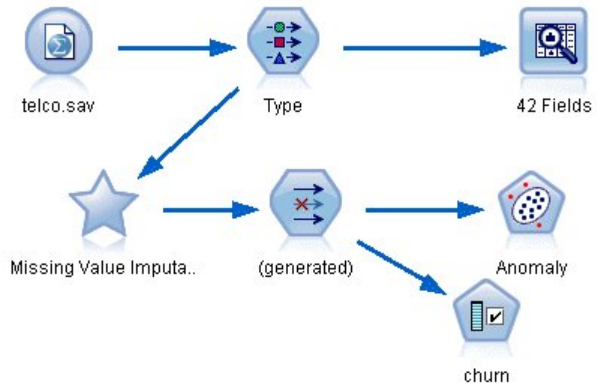
Rozdział 7. Przygotowanie danych do analizy (audyt danych)

Węzeł audytu danych zapewnia obszerny wgląd w dane dostarczone do aplikacji IBM SPSS Modeler. Używany często podczas wstępnej eksploracji danych, raport z audytu danych przedstawia statystyki podsumowujące, jak również histogramy i wykresy rozkładu dla każdej zmiennej danych i pozwala na określenie sposobu traktowania brakujących wartości, wartości odstających i wartości skrajnych.

W tym przykładzie zastosowano strumień o nazwie *telco_dataaudit.str*, który odwołuje się do pliku danych o nazwie *telco.sav*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *telco_dataaudit.str* znajduje się w katalogu *streams*.

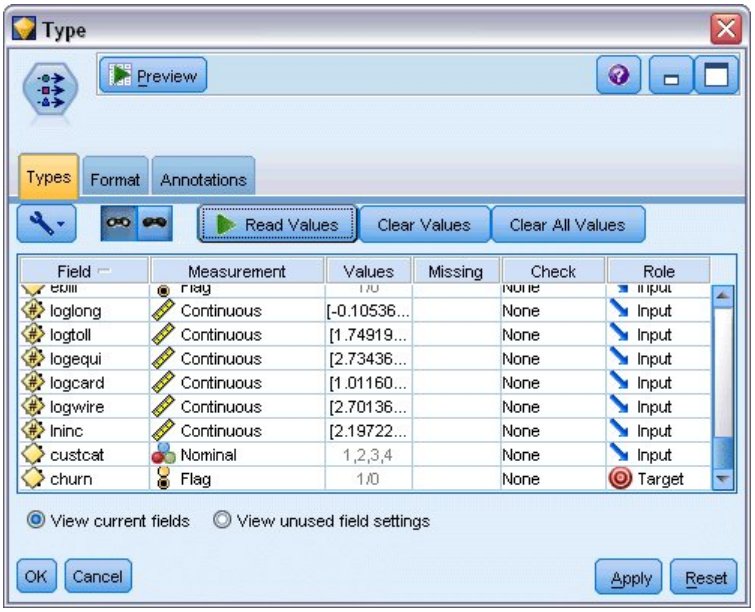
Tworzenie strumienia

1. Aby zbudować strumień, dodaj węzeł źródłowy Plik Statistics wskazujący na plik *telco.sav* znajdujący się w katalogu *Demos* instalacji programu IBM SPSS Modeler.



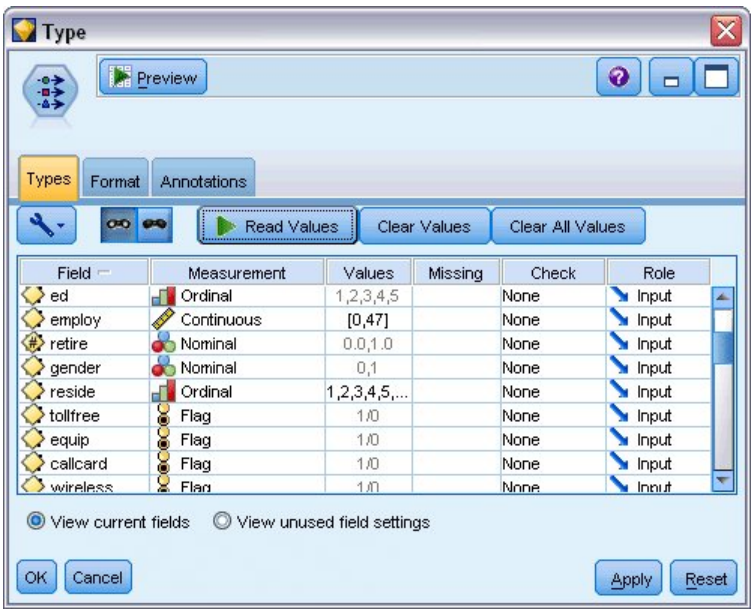
Rysunek 60. Tworzenie strumienia

2. Dodaj węzeł typu, aby zdefiniować zmienne, i wybierz zmienną *odchodzenie* jako zmienną przewidywaną (Rola = **Przewidywana**). Wartość roli należy ustawić na **Dane wejściowe** dla wszystkich innych zmiennych, aby była to jedyna zmienna przewidywana.



Rysunek 61. Ustawianie zmiennej przewidywanej

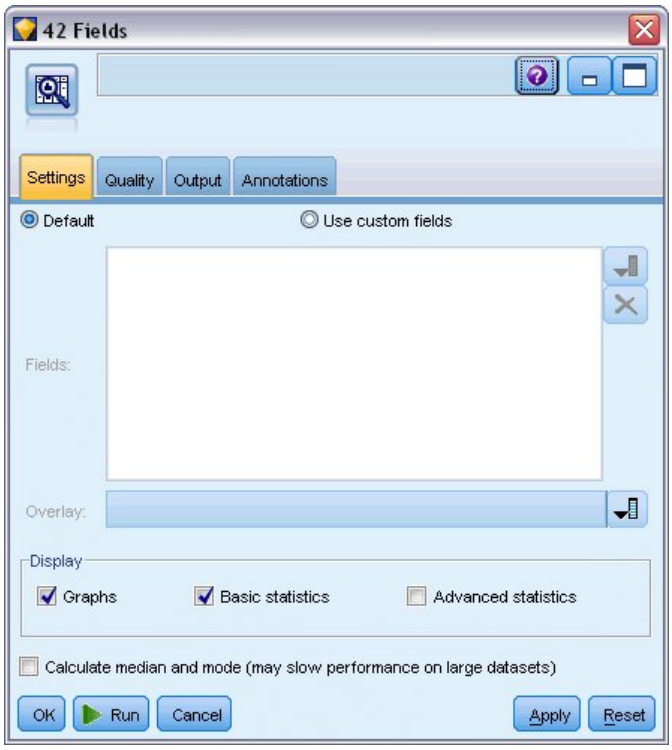
3. Potwierdź, że poziomy pomiaru zmiennych są zdefiniowane poprawnie. Na przykład większość zmiennych z wartościami 0 i 1 można traktować jako flagi, ale niektóre zmienne, takie jak np. płeć, lepiej są przedstawiane jako zmienna nominalna z dwiema wartościami.



Rysunek 62. Ustawianie poziomów pomiaru

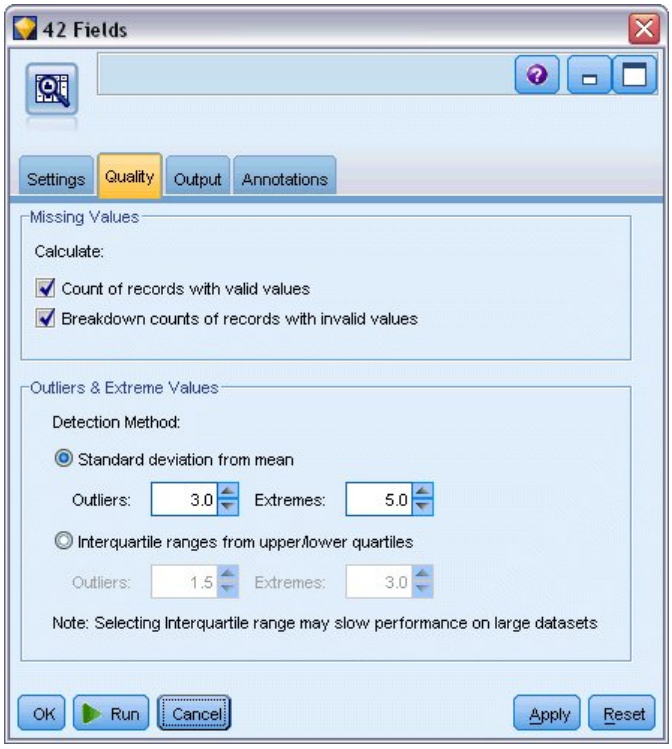
Wskazówka: Aby zmienić właściwości dla wielu zmiennych z podobnymi wartościami (takimi jak 0/1), kliknij nagłówek kolumny *Wartości*, aby posortować zmienne według tej kolumny, i użyj klawisza Shift, aby wybrać wszystkie zmienne, które chcesz zmienić. Możesz następnie kliknąć wybrany zakres prawym klawiszem myszy, aby zmienić poziom pomiaru lub inne atrybuty dla wszystkich wybranych pól.

4. Załącz węzeł audytu danych do strumienia. Na karcie Ustawienia pozostaw domyślne ustawienia, aby uwzględnić wszystkie zmienne w raporcie. Ponieważ *odchodzenie* jest jedyną zmienną przewidywaną w węźle Typ, będzie ona automatycznie użyta jako nałożenie.



Rysunek 63. Węzeł Audyt danych, karta Ustawienia

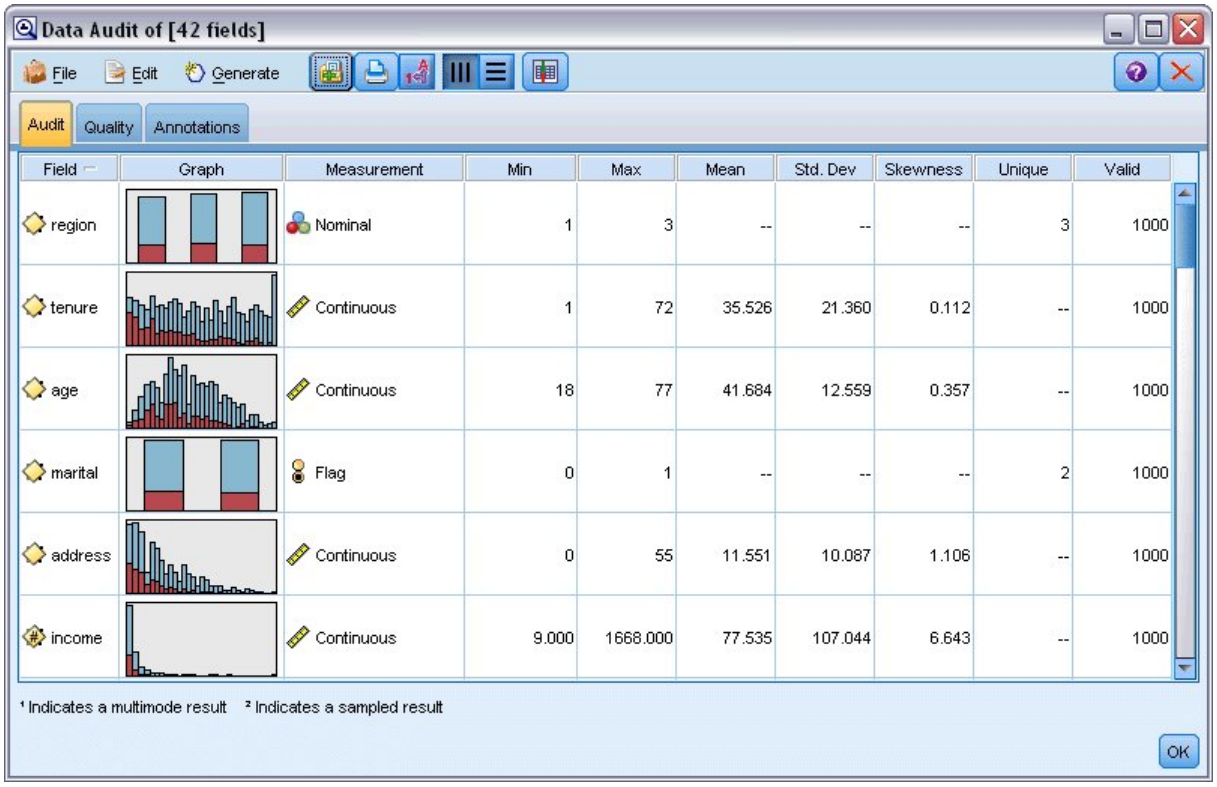
Na karcie Jakość pozostaw ustawienia domyślne, aby wykryć brakujące wartości, wartości odstające i wartości skrajne i kliknij przycisk **Uruchom**.



Rysunek 64. Węzeł Audyt danych, karta Jakość

Przeglądanie statystyk i wykresów

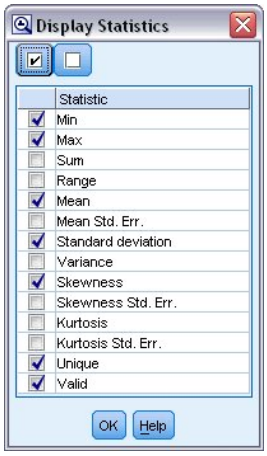
Przeglądarka audytu danych jest wyświetlana z wykresami miniaturowymi oraz statystykami opisowymi dla każdej zmiennej



Rysunek 65. Przeglądarka audytu danych

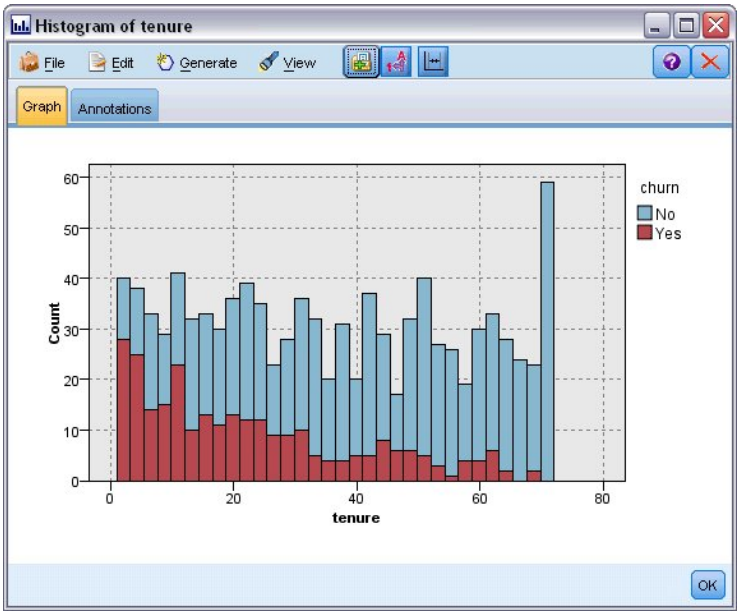
Użyj paska narzędzi, aby wyświetlić etykiety zmiennych i wartości oraz aby przełączyć wyrównanie wykresów z poziomego na pionowy (tylko dla zmiennych jakościowych).

1. Możesz również użyć paska narzędzi lub menu Edycja do wybrania statystyki, która będzie wyświetlana.



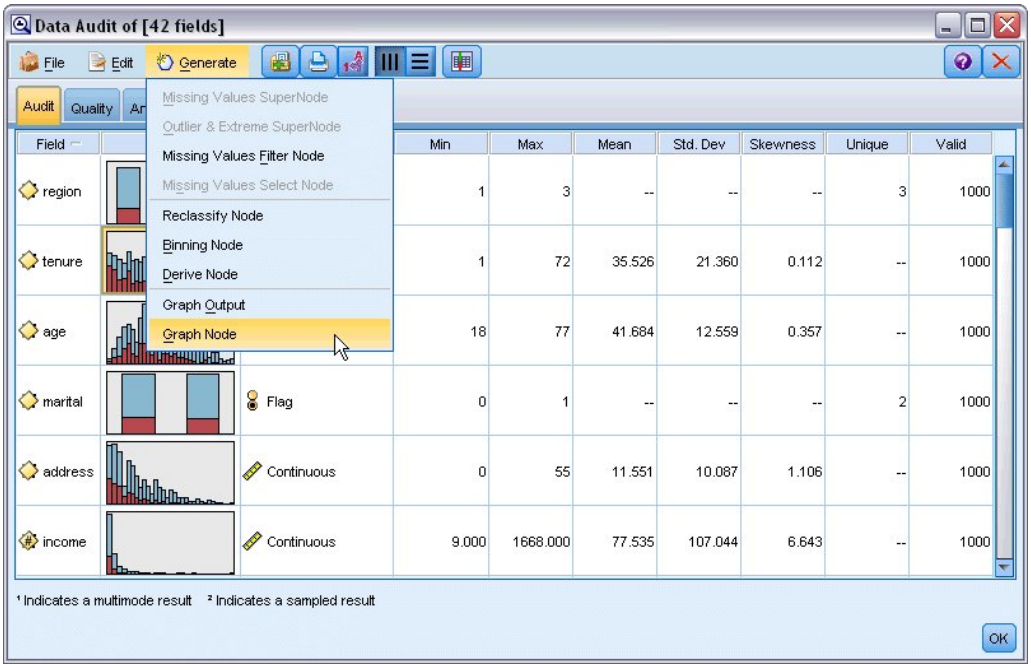
Rysunek 66. Wyświetl statystyki

Kliknij dwukrotnie dowolny wykres miniaturowy w raporcie audytu, aby wyświetlić pełnowymiarową wersję wykresu. Ponieważ *odchodzenie* jest jedyną zmienną przewidywaną w strumieniu, jest ona automatycznie używana jako nałożenie. Można wyłączyć wyświetlanie etykiet zmiennych i wartości, używając paska narzędzi okna wykresów, lub kliknąć przycisk trybu Edycja, aby dalej dostosować wykres.



Rysunek 67. Histogram czasu pracy

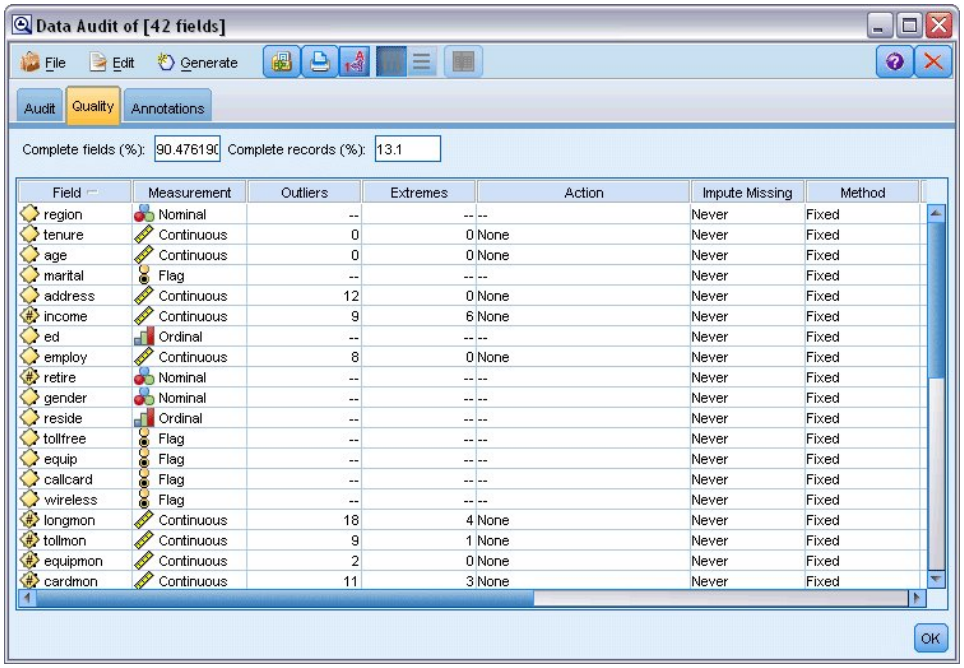
Można też wybrać jedną lub więcej miniaturek i wygenerować węzeł wykresu dla każdej z nich. Wygenerowane węzły są umieszczane w obszarze roboczym strumienia i można je dodać do strumienia, aby odtworzyć konkretny wykres.



Rysunek 68. Generowanie węzła wykresu

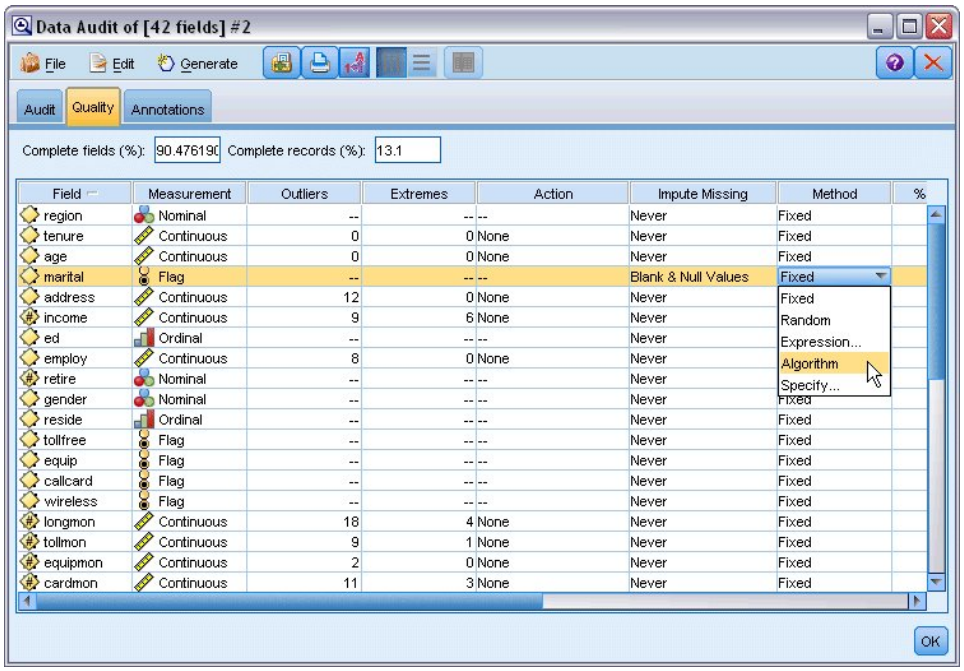
Obsługa wartości odstających i braków danych

Karta Jakość w raporcie z kontroli wyświetla informacje o wartościach odstających, wartościach skrajnych i brakach danych.



Rysunek 69. Przeglądarka audytu danych — karta Jakość

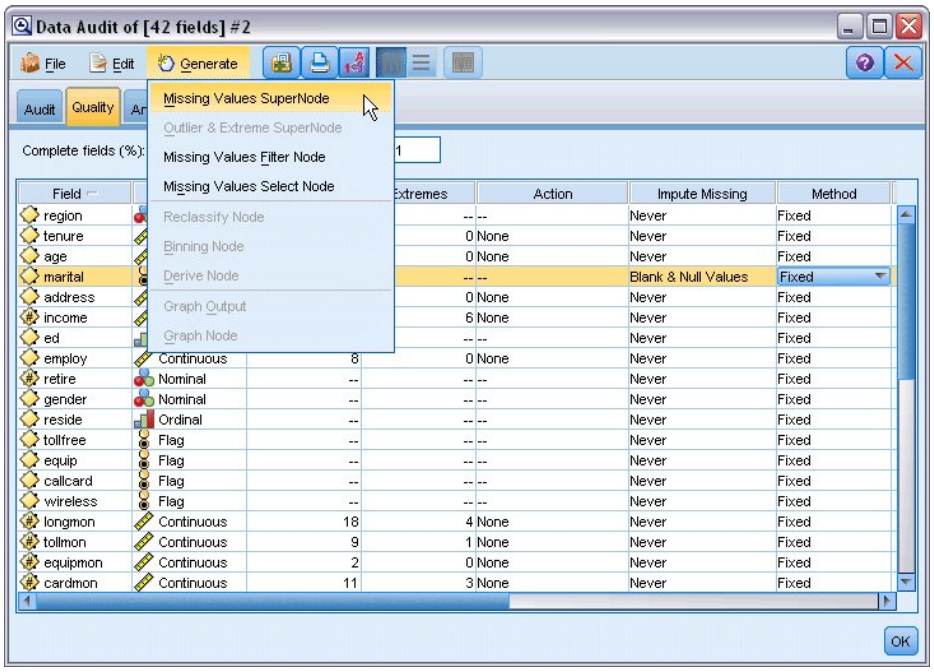
Można również określić metody obsługi tych wartości i generowania superwęzłów, aby automatycznie stosować przekształcenia. Można na przykład wybrać jedną lub wiele zmiennych i zadecydować o podstawieniu albo zastąpieniu brakujących wartości dla tych zmiennych, używając różnych metod, łącznie z algorytmem C&RT.



Rysunek 70. Wybieranie metody podstawiania

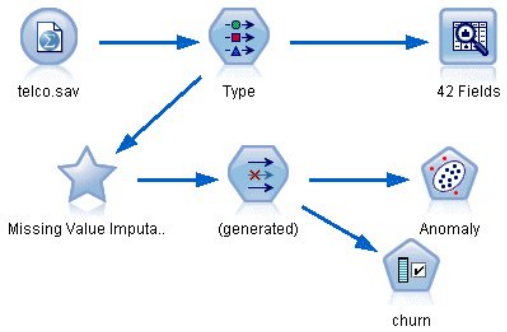
Po określeniu metody podstawiania dla jednego lub wielu zmiennych, aby wygenerować superwęzeł braków danych, wybierz z menu:

Utwórz > Superwęzeł braków danych



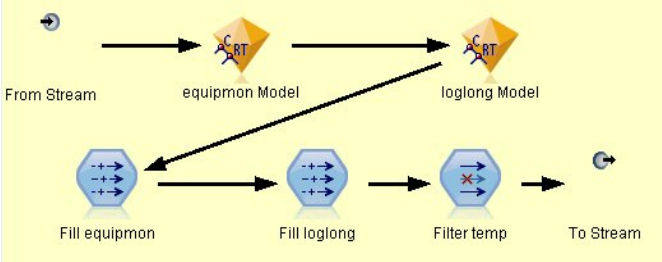
Rysunek 71. Generowanie superwęzła

Wygenerowany superwęzeł jest dodawany do obszaru roboczego strumienia, gdzie można załączyć go do strumienia, aby zastosować przekształcenia.



Rysunek 72. Strumień z superwęzłem braków danych

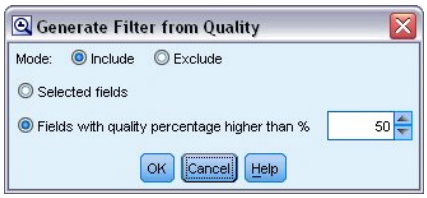
Superwęzeł w rzeczywistości zawiera serię węzłów, które wykonują żądane przekształcenia. Aby zrozumieć sposób działania superwęzła, edytuj go i kliknij opcję **Powiększ**.



Rysunek 73. Powiększanie superwęzła

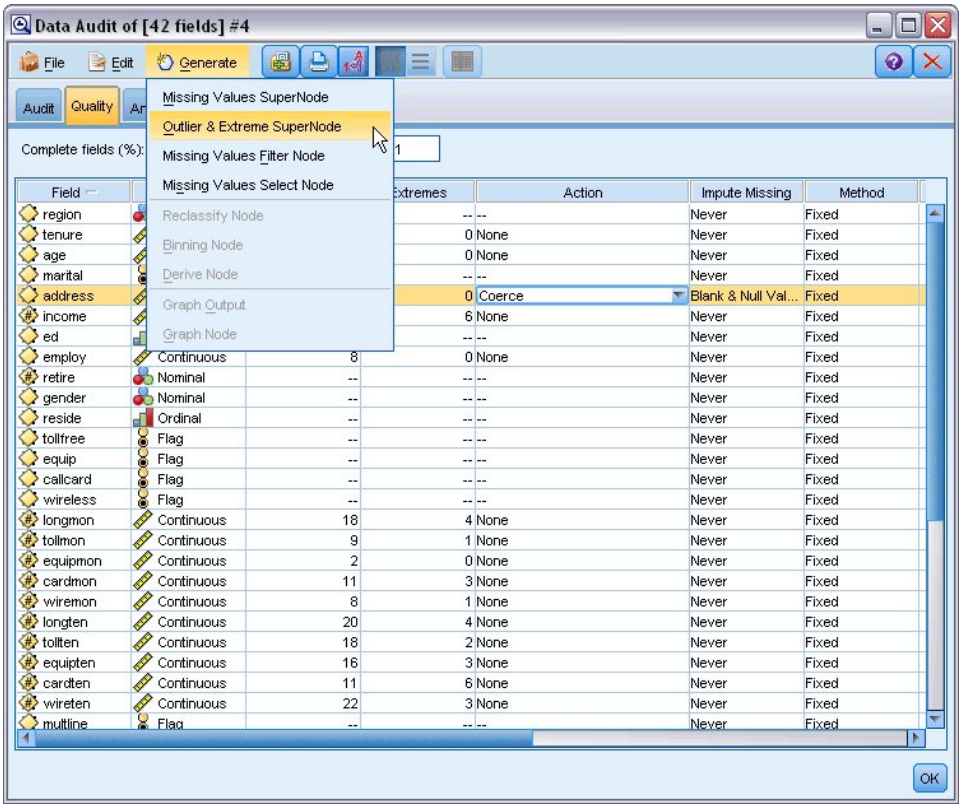
Dla każdej zmiennej wprowadzonej metodą algorytmu dostępny będzie na przykład osobny model C&RT oraz węzeł wypełniania zastępujący wartości puste i wartości null wartością przewidywaną przez model. Można dodawać, edytować lub usuwać określone węzły w ramach superwęzła, aby dalej dostosować działanie.

Można też wygenerować węzeł wyboru lub filtrowania, aby usunąć zmienne lub rekordy z brakami danych. Można na przykład odfiltrować wszystkie zmienne z wartością procentową jakości poniżej określonego progu.



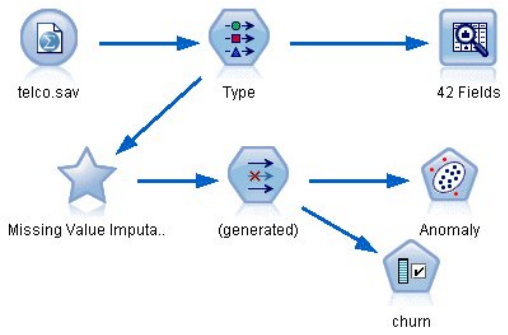
Rysunek 74. Generowanie węzła filtrowania

Wartości skrajne i odstające można potraktować w podobny sposób. Określ czynność, którą chcesz wykonać dla każdej zmiennej (wymuś, odrzuć lub wyzeruj) i wygeneruj superwęzeł, aby zastosować przekształcenia.



Rysunek 75. Generowanie węzła filtrowania

Po zakończeniu audytu i dodaniu wygenerowanych węzłów do strumienia można kontynuować analizę. Można też dalej monitorować dane, używając funkcji Wykrywanie anomalii, Dobór predyktorów lub innych metod.



Rysunek 76. Strumień z superwęzłem braków danych

Rozdział 8. Badanie skuteczności leków (wykresy badawcze/C5.0)

W tej sekcji omówimy przypadek badacza medycznego, który kompiluje dane do badania. Użytkownik zebrał dane o zbiorze pacjentów cierpiących na tę samą chorobę. W trakcie leczenia każdy pacjent zareagował na jeden z pięciu leków. Częścią zadania jest użycie eksploracji danych, aby dowiedzieć się, który lek może być odpowiedni dla przyszłego pacjenta z tą samą chorobą.

W tym przykładzie zastosowano strumień o nazwie *druglearn.str*, który odwołuje się do pliku danych o nazwie *DRUG1n*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *druglearn.str* znajduje się w katalogu *streams*.

Zmienne danych używane w przykładzie demonstracyjnym to:

Zmienna danych	Opis
Age	(Liczba)
Sex	*M* lub *F*
BP	Ciśnienie krwi: *HIGH*, *NORMAL* lub *LOW*
Cholesterol	Cholesterol we krwi: *NORMAL* lub *HIGH*
Na	Poziom sodu we krwi
K	Poziom potasu we krwi
Drug	Lek, na który zareagował pacjent

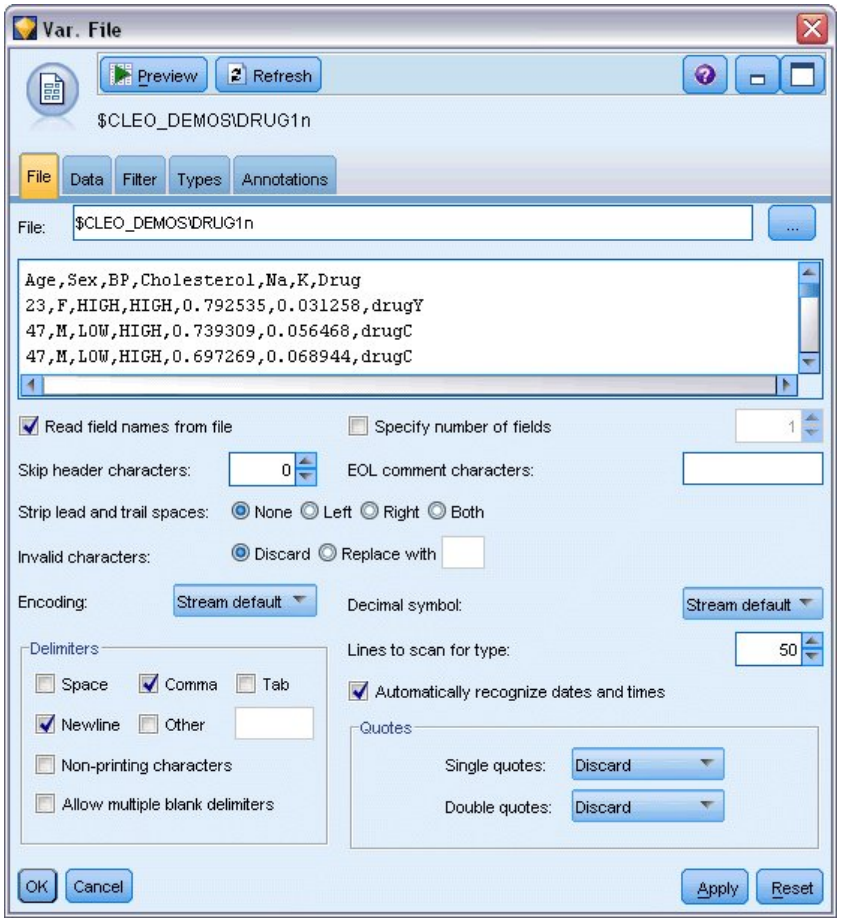
Odczytywanie danych tekstowych

Rysunek 77. Dodawanie węzła pliku zmiennych

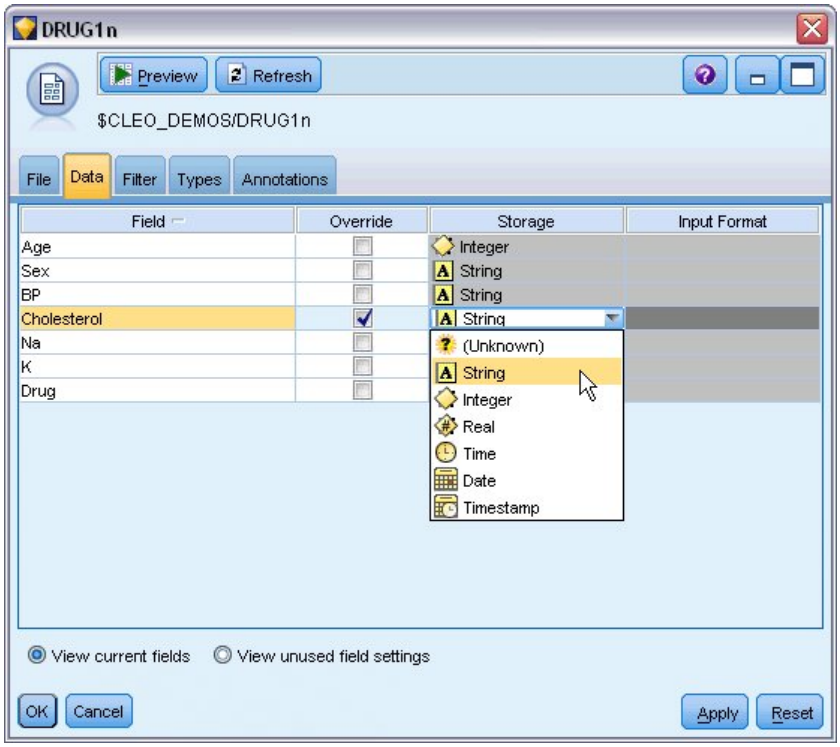
Separowane dane tekstowe można odczytać za pomocą **węzła pliku zmiennych**. Węzeł pliku zmiennych można dodać z palet — kliknij kartę **Źródła**, aby znaleźć węzeł, lub użyj karty **Ulubione**, która domyślnie zawiera ten węzeł. Następnie należy dwukrotnie kliknąć nowo umieszczony węzeł, aby otworzyć jego okno dialogowe.

Kliknij przycisk po prawej stronie pola Plik oznaczony wielokropkiem (...), aby przejść do katalogu, w którym zainstalowany jest program IBM SPSS Modeler. Otwórz katalog *Demos* i wybierz plik o nazwie *DRUG1n*.

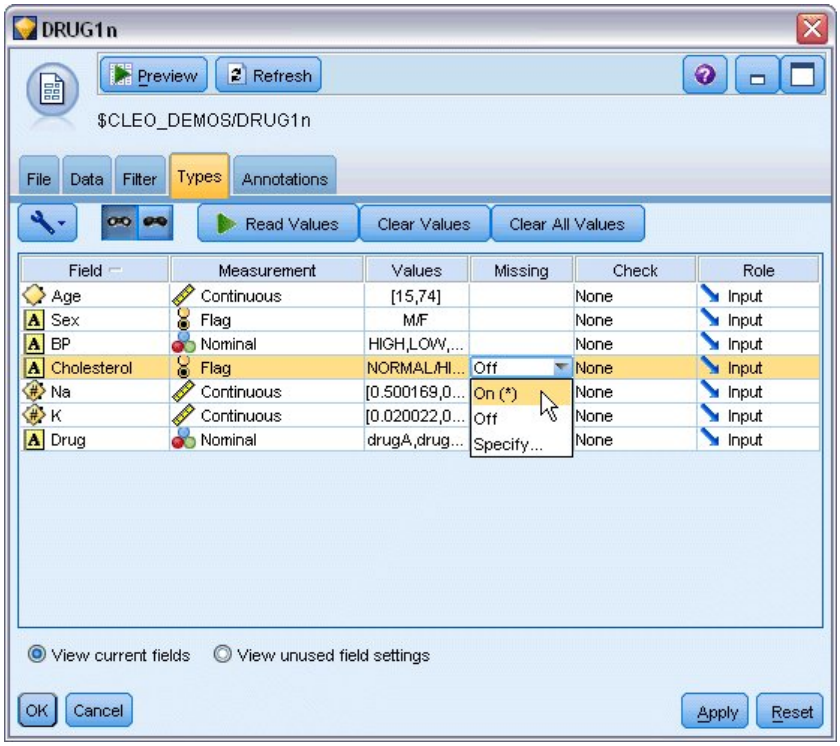
Upewnij się, że zaznaczona jest opcja **Odczytaj nazwy zmiennych z pliku** — zwróć uwagę na zmienne i wartości, które zostały wczytane do okna dialogowego.



Rysunek 78. Okno dialogowe pliku zmiennych.



Rysunek 79. Zmiana typu składowania dla zmiennej



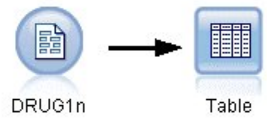
Rysunek 80. Wybieranie opcji wartości na karcie Typy

Kliknij kartę **Dane**, aby zastąpić i zmienić wartość **Składowanie** dla zmiennej. Należy zauważyć, że składowanie różni się od pozycji **Poziom pomiaru**, to znaczy poziomu pomiaru (lub typu użycia) zmiennej danych. Karta **Typy** pozwala

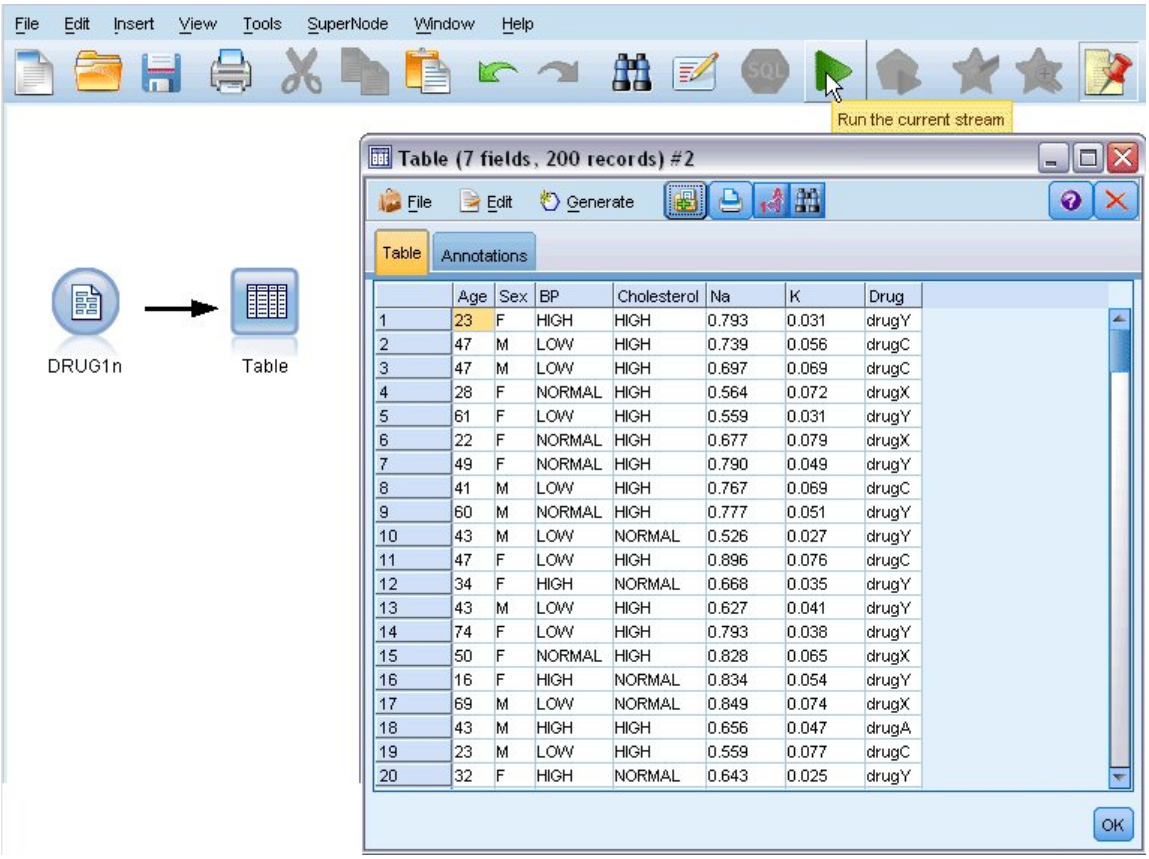
dowiedzieć się więcej o typie zmiennych w danych. Można również wybrać opcję **Odczytaj wartości**, aby wyświetlić rzeczywiste wartości dla każdej zmiennej na podstawie wyboru dokonanego w kolumnie *Wartość*. Ten proces jest znany również jako **tworzenie instancji**.

Dodawanie tabeli

Po załadowaniu pliku z danymi można przejrzeć wartości niektórych rekordów. W tym celu można na przykład zbudować strumień, który zawiera węzeł tabeli. Aby umieścić węzeł tabeli w strumieniu, należy dwukrotnie kliknąć ikonę w palecie lub przeciągnąć ją i upuścić w obszarze roboczym.



Rysunek 81. Węzeł tabeli podłączony do źródła danych



	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

Rysunek 82. Uruchamianie strumienia z paska narzędzi

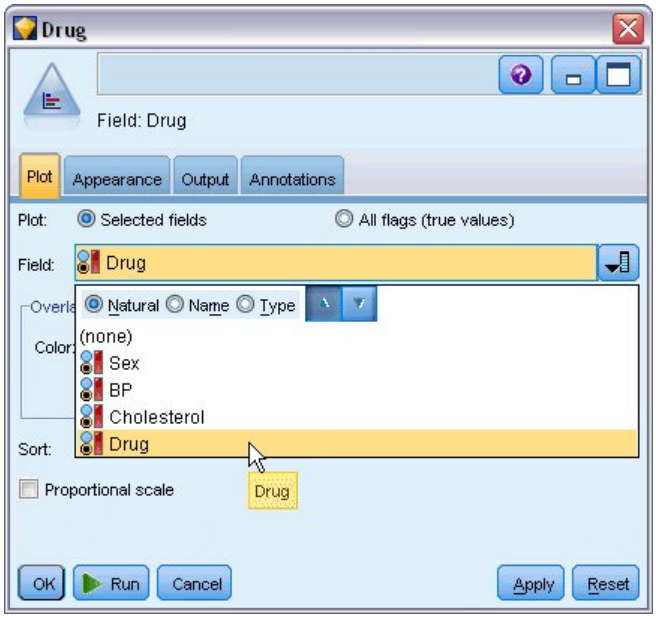
Dwukrotne kliknięcie węzła na palecie spowoduje automatyczne podłączenie go do wybranego węzła w obszarze roboczym strumienia. Jeśli węzły nie są jeszcze połączone, można też użyć środkowego przycisku myszy, aby połączyć węzeł źródłowy z węzłem tabeli. Aby zasymulować środkowy przycisk myszy, przytrzymaj klawisz Alt podczas używania myszy. Aby wyświetlić tabelę, kliknij przycisk zielonej strzałki na pasku narzędzi, aby uruchomić strumień lub prawym przyciskiem myszy kliknij węzeł tabeli i wybierz opcję **Uruchom**.

Tworzenie wykresu rozkładu

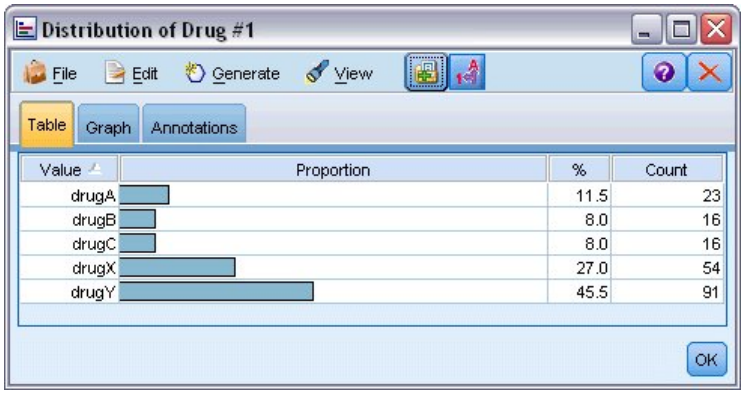
Podczas eksploracji danych często przydatne jest eksplorowanie danych, tworząc podsumowania wizualne. Program IBM SPSS Modeler oferuje kilka różnych rodzajów wykresów do wyboru w zależności od rodzajów danych, które mają być podsumowane. Na przykład, aby dowiedzieć się, jaka część pacjentów zareagowała na każdy lek, należy użyć węzła rozkładu.

Dodaj węzeł rozkładu do strumienia i połącz go z węzłem źródłowym, a następnie kliknij dwukrotnie węzeł, aby edytować opcje wyświetlania.

Wybierz pozycję *Drug* jako zmienną przewidywaną, której rozkład chcesz wyświetlić. Następnie kliknij przycisk **Uruchom** w oknie dialogowym.

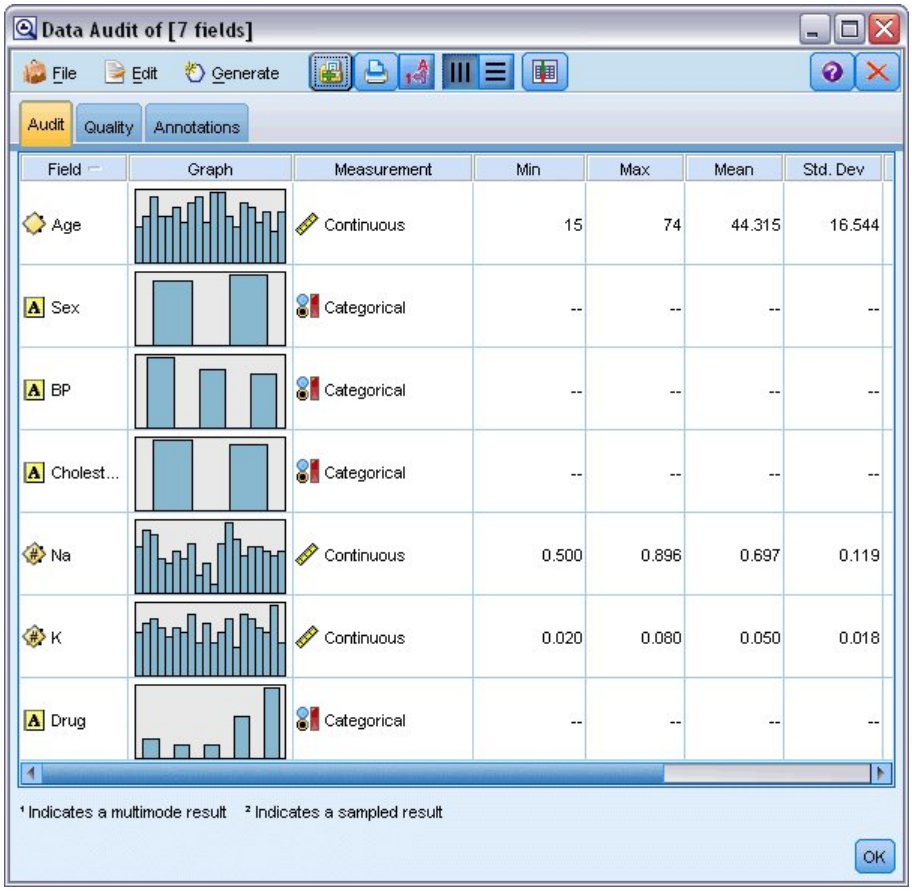


Rysunek 83. Wybieranie pozycji Drug jako zmiennej przewidywanej



Rysunek 84. Rozkład odpowiedzi na typ leku

Wynikowy wykres pozwala zobaczyć „kształt” danych. Wykres pokazuje, że pacjenci reagują najczęściej na lek *Y*, najrzadziej na leki *B* i *C*.



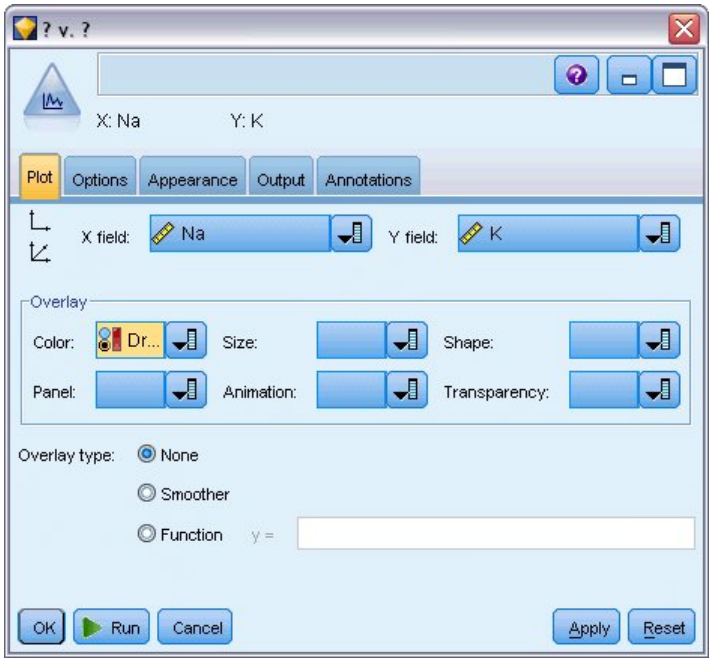
Rysunek 85. Wyniki audytu danych

Można też załączyć i wykonać węzeł audytu danych, aby uzyskać szybki wgląd w rozkłady i histogramy dla wszystkich zmiennych jednocześnie. Węzeł audytu danych jest dostępny na karcie Wynik.

Tworzenie wykresu rozrzutu

Następnie zbadamy, jakie czynniki mogą wpływać na zmienną przewidywaną *Drug*. Badacz wie, że istotnym czynnikiem jest poziom sodu i potasu we krwi. Ponieważ są to wartości liczbowe, można utworzyć wykres rozrzutu sodu i potasu, używając kategorii leków jako kolorowych nałożeń.

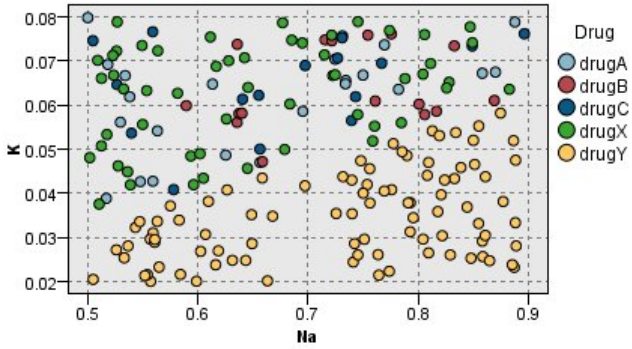
Umieść węzeł wykresu w obszarze roboczym i połącz go z węzłem źródłowym, a następnie dwukrotnie kliknij, aby edytować węzeł.



Rysunek 86. Tworzenie wykresu rozrzutu

Na karcie Wykres wybierz pozycję *Na* jako zmienną X, *K* jako zmienną Y i *Drug* jako zmienną nakładania. Następnie kliknij przycisk **Uruchom**.

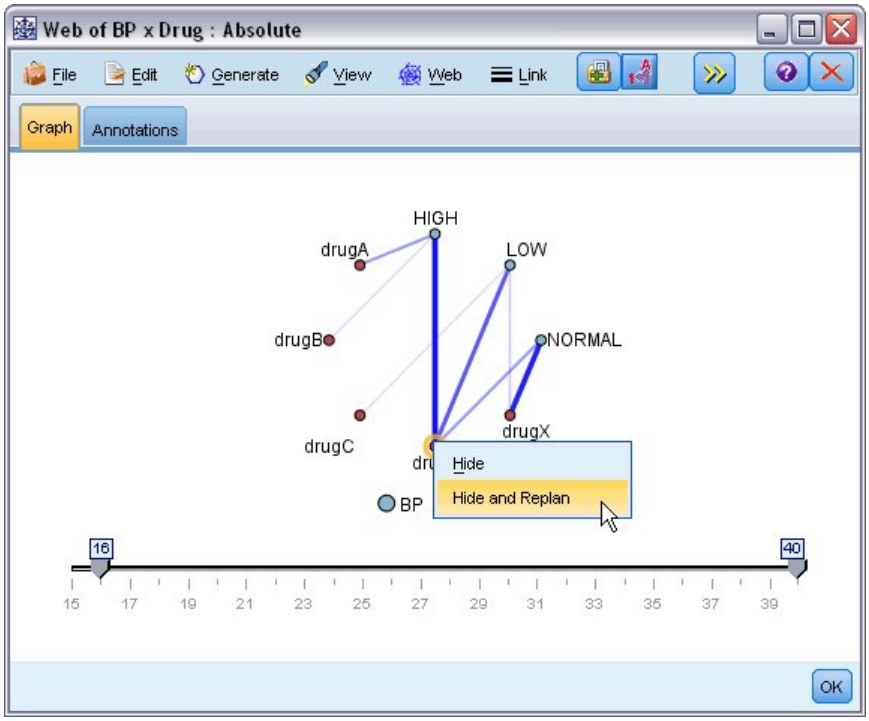
Wykres wyraźnie przedstawia próg, powyżej którego poprawny lek to zawsze lek *Y* i poniżej którego poprawny lek to nigdy nie jest lek *Y*. Ten próg jest współczynnikiem sodu (*Na*) do potasu (*K*).



Rysunek 87. Wykres rozrzutu leków

Tworzenie wykresu sieciowego

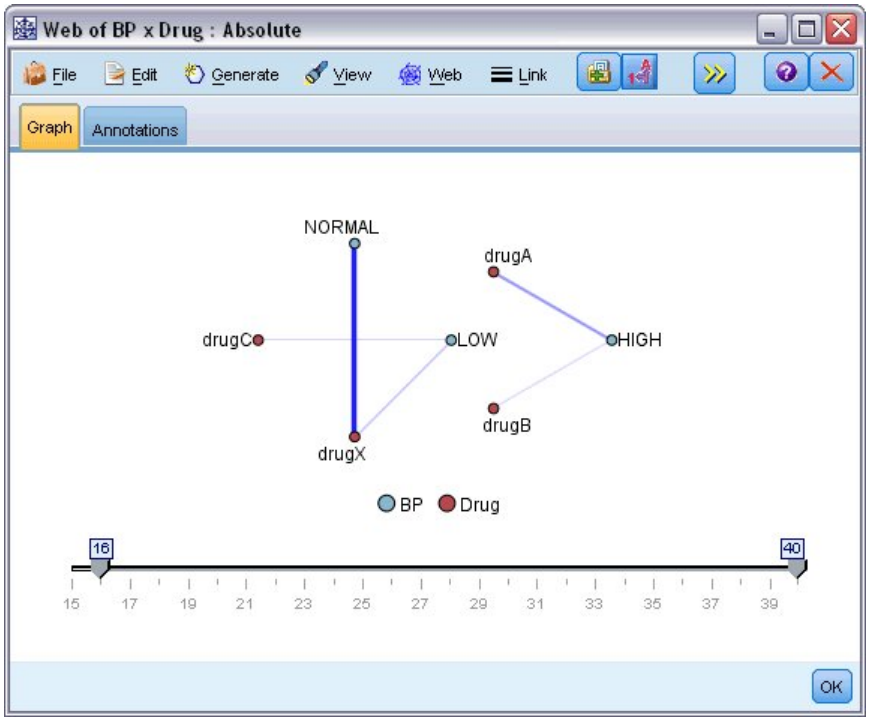
Ponieważ wiele z tych zmiennych danych jest jakościowych, można również spróbować utworzyć wykres sieciowy, który odwzorowuje powiązania między różnymi kategoriami. Rozpocznij, podłączając węzeł sieciowy do węzła źródłowego w obszarze roboczym. W oknie dialogowym węzła sieciowego wybierz zmienne *BP* (ciśnienie krwi) i *Drug*. Następnie kliknij przycisk **Uruchom**.



Rysunek 88. Wykres sieciowy leków i ciśnienia krwi

Na wykresie widać, że lek *Y* jest powiązany ze wszystkimi trzema poziomami ciśnienia krwi. To nie jest zaskoczenie, ponieważ określono już, że lek *Y* jest najlepszy. Aby skoncentrować się na innych lekach, można ukryć lek *Y*. W menu **Widok** wybierz opcję **Tryb edycji**, a następnie kliknij prawym przyciskiem myszy nad punktem *Y* i wybierz opcję **Ukryj i zaplanuj ponownie**.

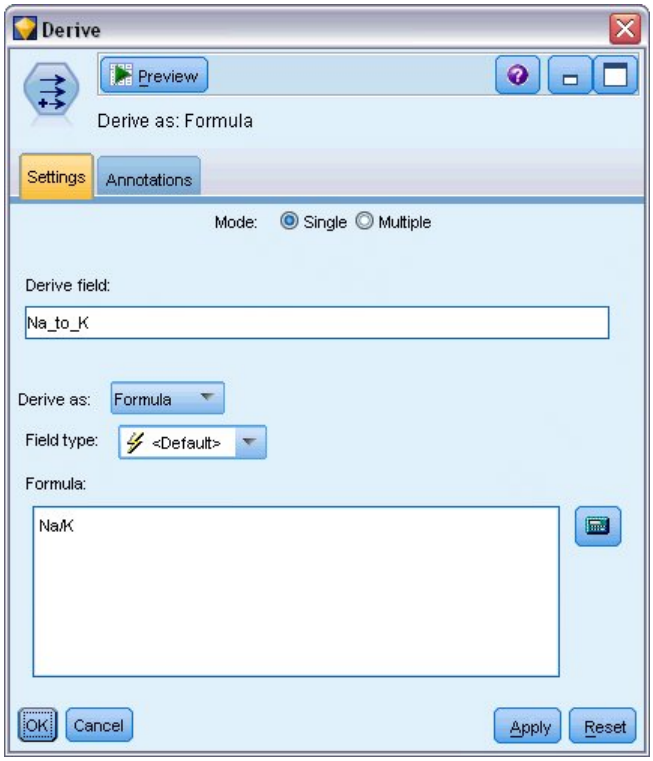
Na uproszczonym wykresie lek *Y* i wszystkie jego łącza są ukryte. Teraz wyraźnie widoczne jest, że tylko leki *A* i *B* są powiązane z wysokim ciśnieniem krwi. Tylko leki *C* i *X* są powiązane z niskim ciśnieniem krwi. Dodatkowo normalne ciśnienie krwi jest powiązane tylko z lekiem *X*. W tym miejscu jednak wciąż jeszcze nie wiadomo, jak wybrać pomiędzy lekami *A* i *B* lub pomiędzy lekami *C* i *X* dla określonego pacjenta. W takiej sytuacji może pomóc modelowanie.



Rysunek 89. Wykres sieciowy z ukrytym lekiem Y i jego łączami.

Wyliczanie nowej zmiennej

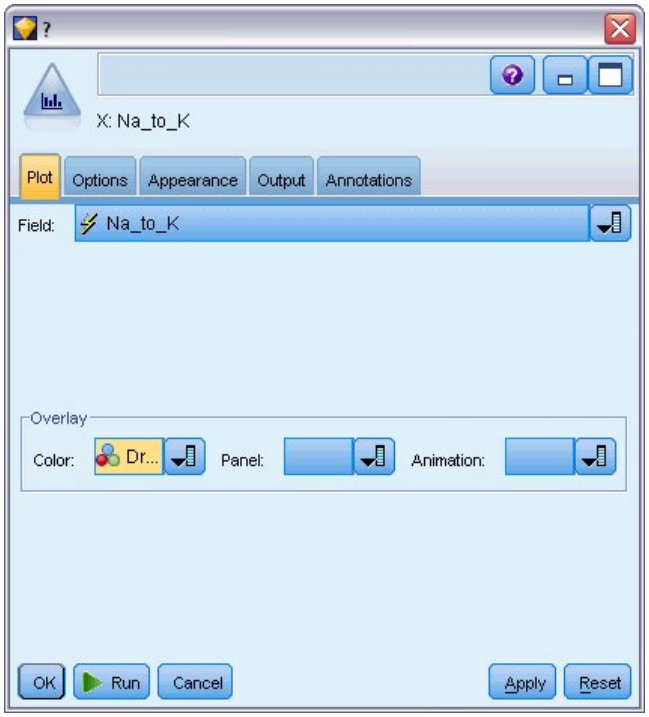
Ponieważ współczynnik sodu do potasu wydaje się przewidywać, kiedy użyć leku *Y*, można wyliczyć zmienną, która zawiera wartość tego współczynnika dla każdego rekordu. Ta zmienna może być przydatna później, gdy budowany jest model przewidujący, kiedy użyć każdego z pięciu leków. Aby uprościć układ strumienia, rozpocznij, usuwając wszystkie węzły z wyjątkiem węzła źródłowego DRUG1n. Do DRUG1n dołącz węzeł Wyliczanie (karta Zmienne), a następnie dwukrotnie kliknij węzeł Wyliczanie, aby go edytować.



Rysunek 90. Edytowanie węzła wyliczeń

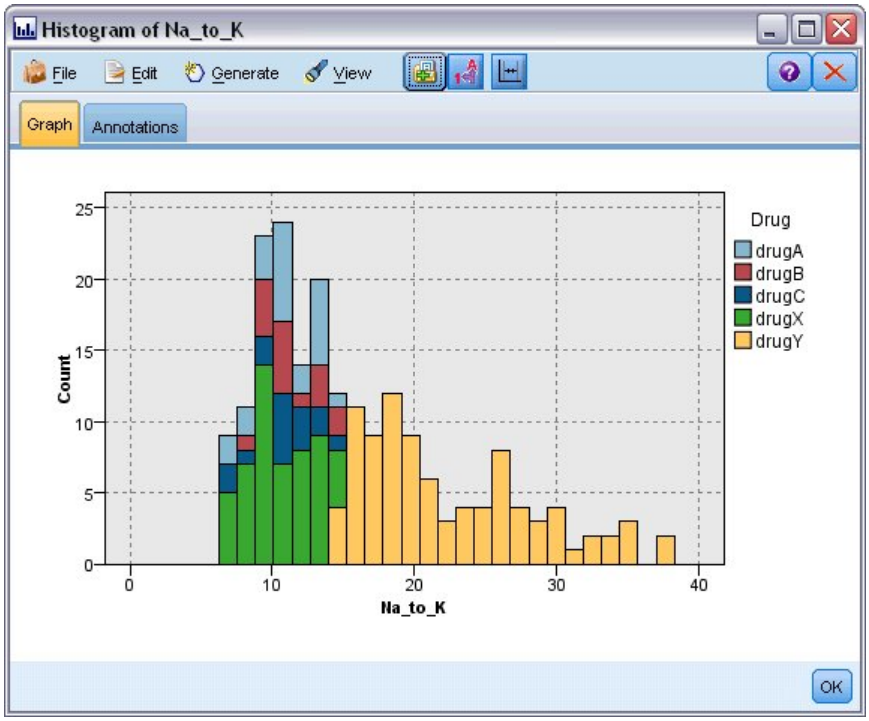
Nazwij nową zmienną *Na_to_K*. Ponieważ nowa zmienna jest uzyskiwana przez podzielenie wartości poziomu sodu przez wartość poziomu potasu, wprowadź formułę Na/K. Można również utworzyć formułę, klikając ikonę po prawej stronie zmiennej. Powoduje to otwarcie Konstruktora wyrażeń, który umożliwia interaktywne tworzenie wyrażeń, używając wbudowanej listy funkcji, operandów i zmiennych oraz ich wartości.

Można sprawdzić rozkład nowej zmiennej, załączając węzeł histogramu do węzła wyliczeń. W oknie dialogowym węzła histogramu określ *Na_to_K* jako zmienną wykresu i *Drug* jako zmienną nałożenia.



Rysunek 91. Edytowanie węzła histogramu

Po uruchomieniu strumienia otrzymywany jest wykres przedstawiony tutaj. Na podstawie wykresu można wyciągnąć wniosek, że gdy wartość *Na_to_K* wynosi 15 lub więcej, lek *Y* jest wybranym lekiem.

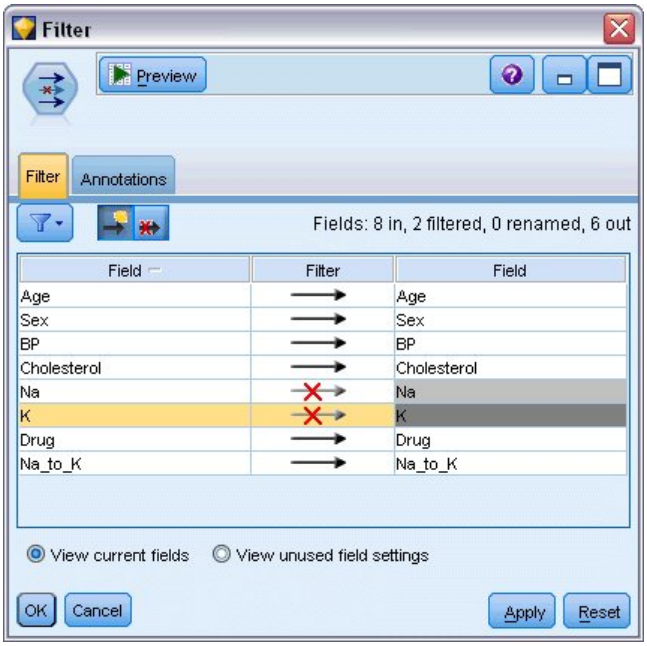


Rysunek 92. Wyświetlony histogram

Budowanie modelu

Eksplorując dane i manipulując nimi, można było sformułować pewne hipotezy. Współczynnik sodu do potasu we krwi wydaje się wpływać na wybór leku, tak samo jako ciśnienie krwi. Nie można jeszcze jednak w pełni wyjaśnić wszystkich relacji. Dlatego właśnie modelowanie może zapewnić niektóre odpowiedzi. W tym przypadku spróbujemy dopasować dane, używając modelu budującego reguły C5.0.

Ponieważ używana jest zmienna pochodna *Na_to_K*, można odfiltrować oryginalne zmienne *Na* i *K*, aby nie były używane dwukrotnie w algorytmie modelowania. Można to zrobić za pomocą węzła filtrowania.

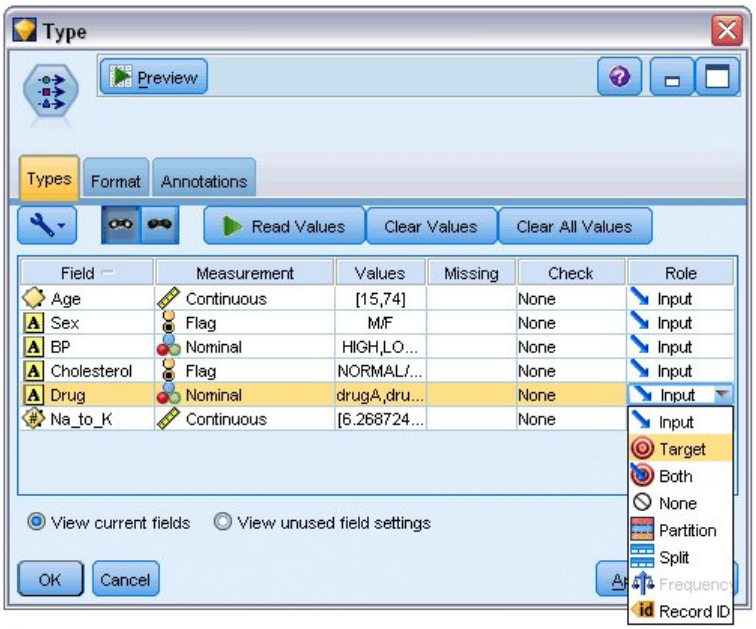


Rysunek 93. Edytowanie węzła filtrowania

Na karcie Filtrowanie kliknij strzałki obok zmiennych *Na* i *K*. Nad strzałkami pojawiają się czerwone znaki X wskazujące, że zmienne są teraz odfiltrowane.

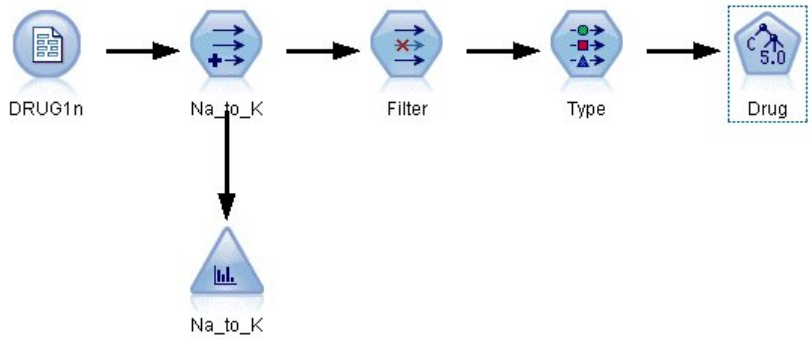
Następnie załącz węzeł typu podłączony do węzła filtrowania. Węzeł typu pozwala na wskazanie typów zmiennych, które są używane i określenie sposobu, w jaki są używane do przewidywania wyników.

Na karcie Typy dla zmiennej *Drug* ustaw rolę **Przewidywana**, wskazującą, że *Drug* jest zmienną, która ma zostać przewidziana. Pozostaw role dla innych zmiennych ustawione na wartość **Dane wejściowe**, aby były używane jako predyktory.



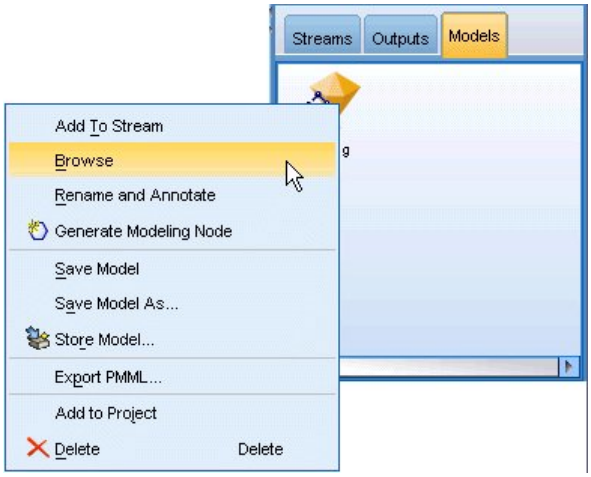
Rysunek 94. Edytowanie węzła typu

Aby oszacować model, umieść węzeł C5.0 w obszarze roboczym i załącz go na końcu strumienia, jak pokazano w przykładzie. Następnie kliknij zielony przycisk **Uruchom** na pasku narzędzi, aby uruchomić strumień.



Rysunek 95. Dodawanie węzła C5.0

Przeglądanie modelu



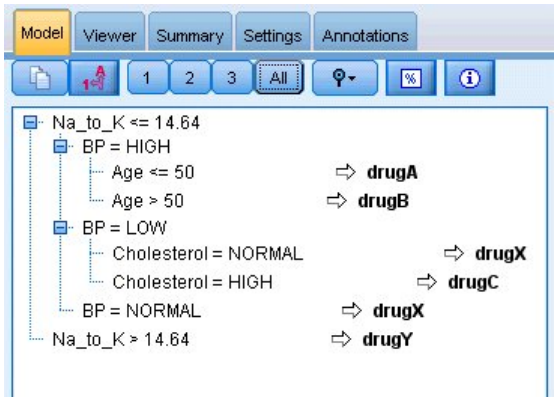
Rysunek 96. Przeglądanie modelu

Kiedy wykonywany jest węzeł C5.0, model użytkowy dodawany jest do strumienia i palety modeli w prawym górnym rogu okna. Aby przeglądać model, kliknij prawym przyciskiem myszy dowolną ikonę i z menu kontekstowego wybierz opcję **Edytuj** lub **Przeglądaj**.

Przeglądarka reguł wyświetla zestaw reguł wygenerowanych przez węzeł C5.0 w formacie drzewa decyzyjnego. Wstępnie drzewo jest zwinięte. Aby je rozwinąć, kliknij przycisk **Wszystkie**, aby wyświetlić wszystkie poziomy.

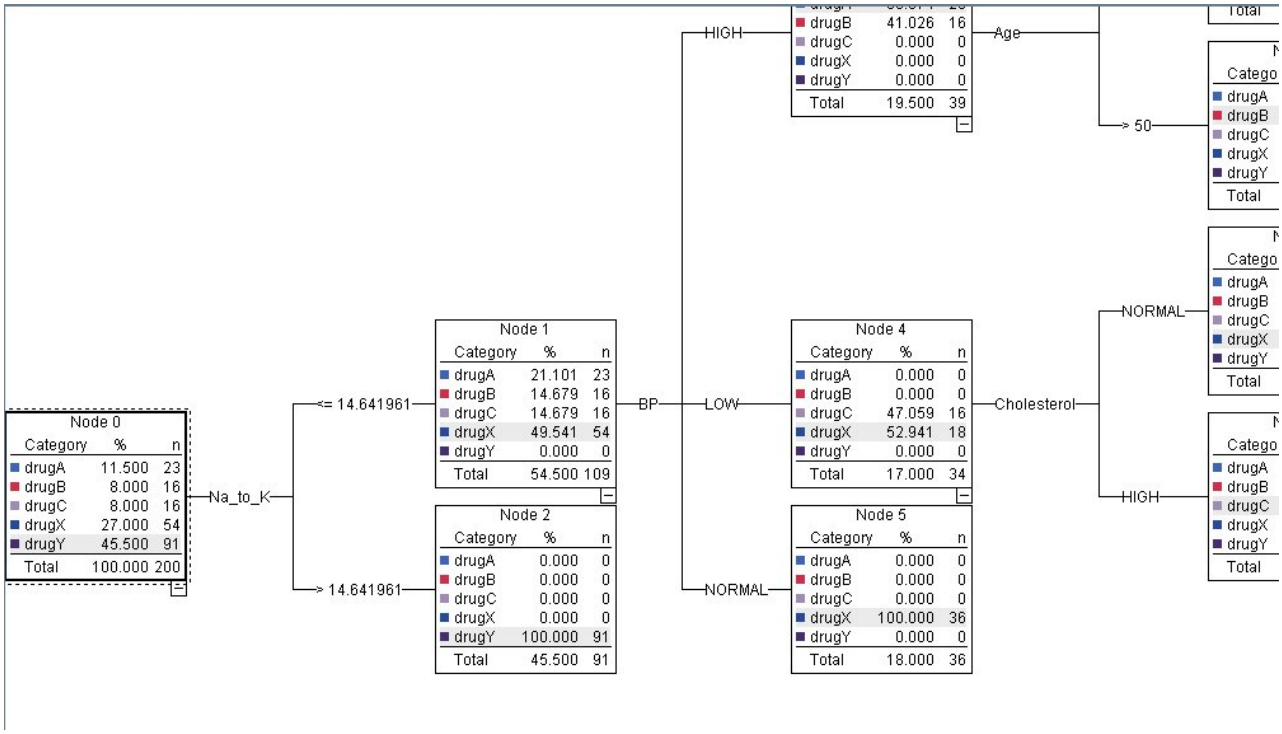
Rysunek 97. Przeglądarka reguł

Teraz można dostrzec brakujące elementy rozwiązania. Dla osób ze współczynnikiem *Na* do *K* niższym niż 14,64 i wysokim ciśnieniem krwi wiek określa najlepszy lek. Dla osób z niskim ciśnieniem krwi najlepszym predyktorem wydaje się poziom cholesterolu.



Rysunek 98. W pełni rozwinięta przeglądarka reguł

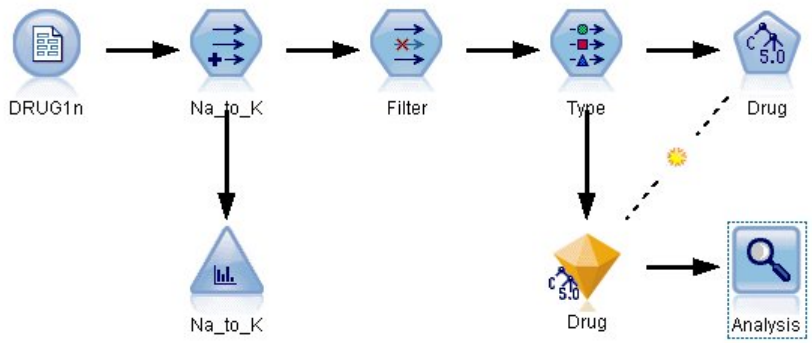
To samo drzewo decyzyjne można wyświetlić w bardziej rozbudowanym formacie graficznym, klikając kartę **Przegląd**. W tym obszarze można łatwiej zobaczyć liczbę przypadków dla każdej kategorii ciśnienia krwi, jak również udział procentowy przypadków.



Rysunek 99. Drzewo decyzyjne w formacie graficznym

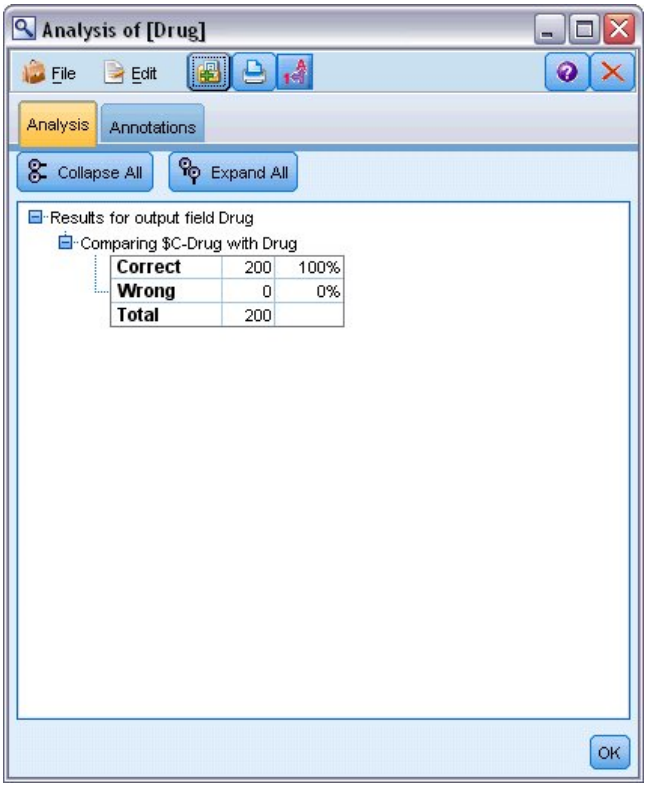
Używanie węzła analizy

Można ocenić dokładność modelu, używając węzła analizy. Załącz węzeł analizy (z palety węzłów wyników) do modelu użytkowego, otwórz węzeł analizy i kliknij pozycję **Uruchom**.



Rysunek 100. Dodawanie węzła analizy

Wyniki węzła analizy pokazują, że dla tego sztucznego zbioru danych model poprawnie przewidział wybór leku dla każdego rekordu w zbiorze danych. W przypadku rzeczywistego zbioru danych mało prawdopodobne jest uzyskanie 100-procentowej dokładności, ale można użyć węzła analizy, aby pomóc określić, czy model jest odpowiednio dokładny dla konkretnego zastosowania.



Rysunek 101. Wyniki węzła analizy

Rozdział 9. Monitorowanie predyktorów (Dobór predyktorów)

Węzeł Dobór predyktorów pomaga zidentyfikować pola, które są najważniejsze w przewidywaniu określonego wyniku. W zestawie setek lub nawet tysięcy predyktorów węzeł Dobór predyktorów monitoruje, szereguje i wybiera predyktory, które mogą być najważniejsze. Ostatecznie można uzyskać szybszy i wydajniejszy model — używający mniejszej liczby predyktorów, szybciej wykonywany i łatwiejszy do zrozumienia.

Dane użyte w tym przykładzie reprezentują hurtownię danych hipotetycznej firmy telefonicznej i zawierają informacje dotyczące odpowiedzi na specjalną promocję wśród 5000 klientów firmy. Dane obejmują dużą liczbę zmiennych dotyczących wieku, zatrudnienia, dochodów klientów oraz statystyki dot. korzystania z telefonu. Trzy zmienne przewidywane przedstawiają dane, czy klient odpowiedział na każdą z ofert. Firma chce użyć tych danych, aby lepiej przewidzieć, którzy klienci najprawdopodobniej odpowiedzą na podobne oferty w przyszłości.

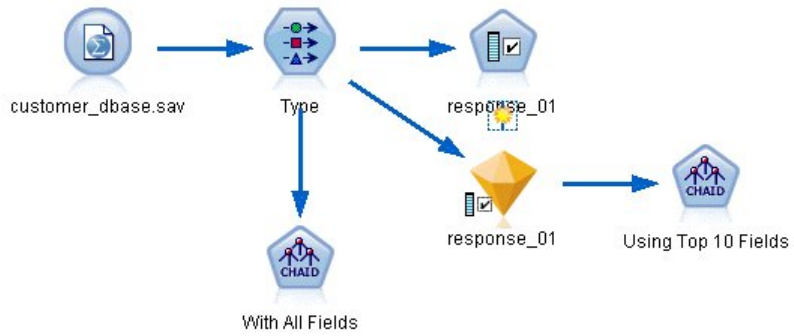
W tym przykładzie zastosowano strumień o nazwie *featureselection.str*, który odwołuje się do pliku danych o nazwie *customer_dbase.sav*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik o nazwie *featureselection.str* znajduje się w katalogu *streams*.

Ten przykład koncentruje się tylko na jednej ofercie jako zmiennej przewidywanej. Przykład używa węzła budowania drzewa CHAID do opracowania modelu opisującego, którzy klienci z największym prawdopodobieństwem odpowiedzą na promocję. Przykład porównuje dwa podejścia:

- Bez wyboru predyktorów. Wszystkie zmienne predyktorów w zbiorze danych są używane jako dane wejściowe dla drzewa CHAID.
- Z wyborem predyktorów. Węzeł Dobór predyktorów jest używany do wyboru 10 najważniejszych predyktorów. Są one następnie danymi wejściowymi drzewa CHAID.

Porównując dwa wynikowe modele drzew, widzimy, w jaki sposób dobór predyktorów zapewnia efektywne wyniki.

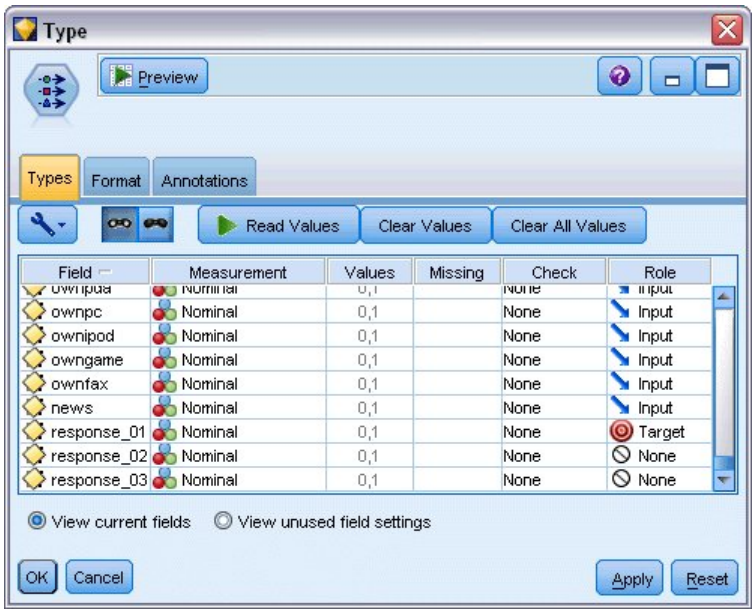
Tworzenie strumienia



Rysunek 102. Przykładowy strumień doboru predyktorów

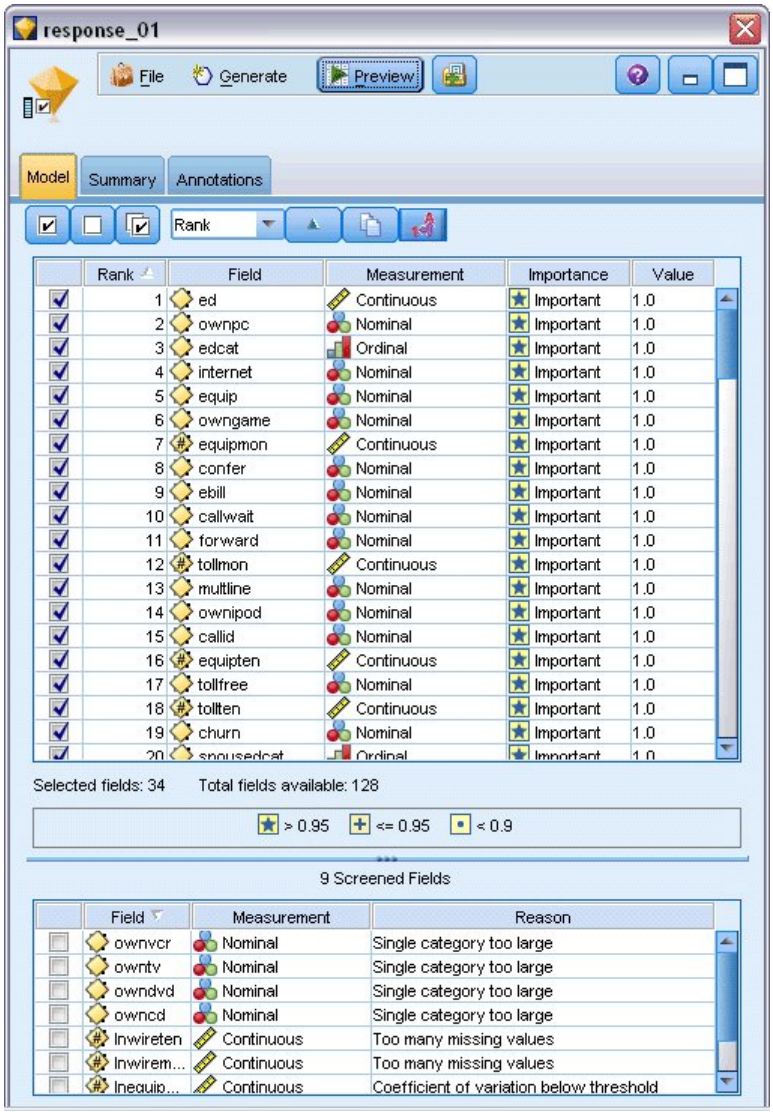
1. Umieść węzeł źródłowy Plik Statistics na pustym obszarze roboczym strumienia. Skieruj ten węzeł na plik danych przykładowych *customer_dbase.sav* dostępny w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. (Można też otworzyć plik przykładowego strumienia *featureselection.str* znajdujący się w katalogu *streams*).
2. Dodaj węzeł typu. Na karcie Typy przewiń w dół do samego końca i zmień rolę dla zmiennej *response_01* na *Przewidywana*. Zmień rolę na *Brak* dla innych zmiennych odpowiedzi (*response_02* i *response_03*), jak również dla identyfikatora klienta (*custid*) w górnej części listy. Pozostaw rolę ustawioną na *Dane wejściowe* dla

wszystkich pozostałych zmiennych i kliknij przycisk **Odczytaj wartości**, a następnie kliknij przycisk **OK**.



Rysunek 103. Dodawanie węzła typu

3. Do strumienia dodaj węzeł modelowania Dobór predyktorów. W tym węźle można określić reguły i kryteria monitorowania lub dyskwalifikacji zmiennych.
4. Uruchom strumień, aby utworzyć model użytkowy Dobór predyktorów.
5. Kliknij model użytkowy prawym przyciskiem myszy w strumieniu lub na palecie modeli i wybierz opcję **Edytuj** lub **Przeglądaj**, aby wyświetlić wyniki.

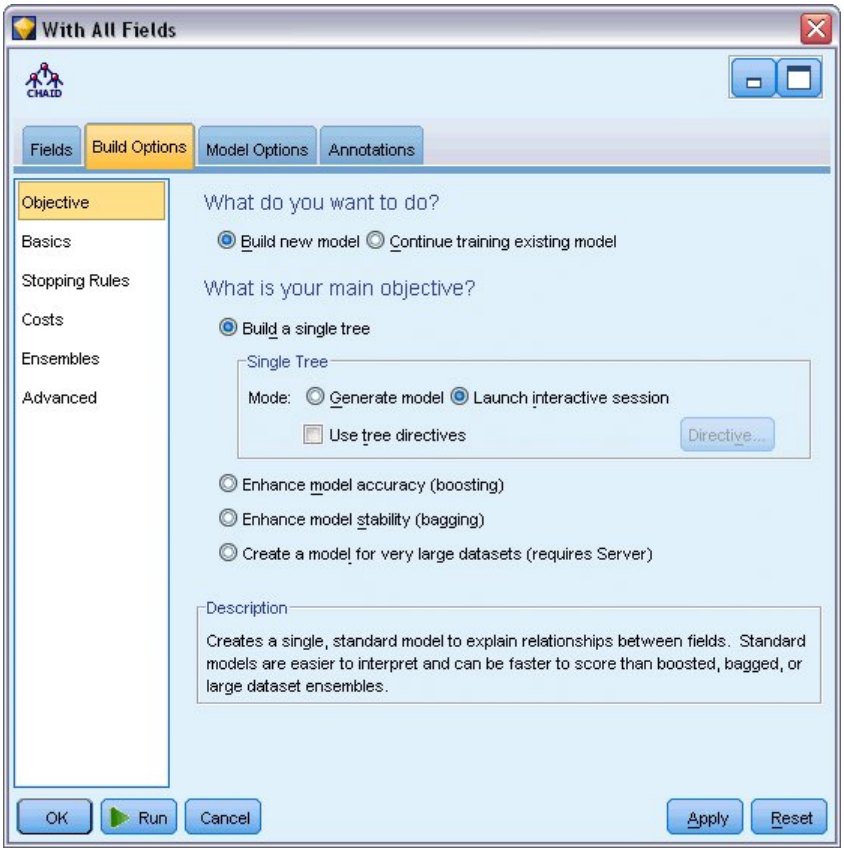


Rysunek 104. Karta Model w modelu użytkowym Dobór predyktorów

Górny panel przedstawia zmienne wykryte jako przydatne w predykcji. Są one uszeregowane na podstawie ważności. Dolny panel przedstawia zmienne, które są monitorowane w analizie i dlaczego. Badając zmienne w górnym panelu, można zadecydować, które zostaną użyte w kolejnych sesjach modelowania.

6. Teraz można wybrać zmienne do użycia w kolejnych węzłach. Mimo że wstępnie określono 34 zmienne jako ważne, jeszcze bardziej chcemy ograniczyć zbiór predyktorów.
7. Wybierz tylko 10 górnych predyktorów, używając znaczników wyboru w pierwszej kolumnie, aby usunąć zaznaczenie niechcianych predyktorów. (Kliknij znacznik wyboru w wierszu 11, przytrzymaj klawisz Shift i kliknij znacznik wyboru w wierszu 34). Zamknij model użytkowy.
8. Aby porównać wyniki bez wyboru predyktorów, należy dodać do strumienia dwa węzły modelowania CHAID: jeden używający doboru predyktorów i drugi, który nie używa tej metody.
9. Podłącz jeden węzeł CHAID do węzła typu i drugi do modelu użytkowego Dobór predyktorów.
10. Otwórz każdy węzeł CHAID, wybierz kartę Opcje budowania i upewnij się, że w oknie Cele zaznaczone są opcje **Zbuduj nowy model**, **Zbudować pojedyncze drzewo** i **Drzewo interakcyjne**.

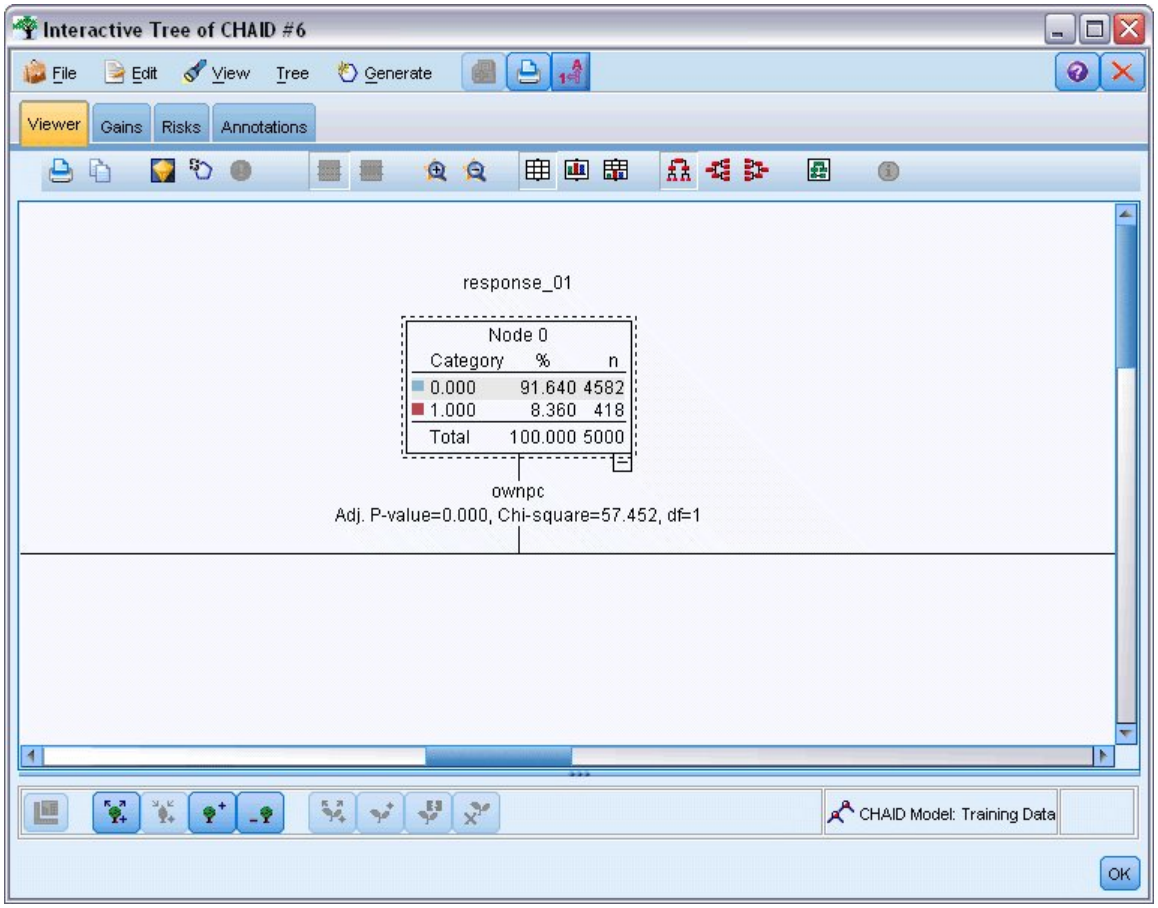
W oknie Podstawowe sprawdź, czy parametr **Maksymalna głębokość drzewa** jest ustawiony na 5.



Rysunek 105. Ustawienia celów dla węzła modelowania CHAID dla wszystkich zmiennych predykcyjnych

Budowanie modeli

1. Wykonaj węzeł CHAID, który używa wszystkich predyktorów w zbiorze danych (węzeł podłączony do węzła typu). Należy zwrócić uwagę, jak długo trwa jego wykonanie. Okno wyników wyświetla tabelę.
2. W menu wybierz kolejno **Drzewo > Rozwiń drzewo**, aby rozwinąć i wyświetlić rozwinięte drzewo.



Rysunek 106. Rozwijanie drzewa w konstruktorze drzewa

3. Teraz zrób to samo dla drugiego węzła CHAID, który używa tylko 10 predyktorów. Ponownie rozwiń drzewo po otwarciu konstruktora drzewa.

Drugi model powinien zostać wykonany szybciej niż pierwszy. Ponieważ ten zbiór danych jest raczej mały, różnica w czasie wykonania wynosi prawdopodobnie kilka sekund, ale dla większych rzeczywistych zbiorów danych, różnica może być bardzo zauważalna i może być mierzona w minutach lub nawet godzinach. Użycie wyboru predyktorów może znacznie przyspieszyć czas przetwarzania.

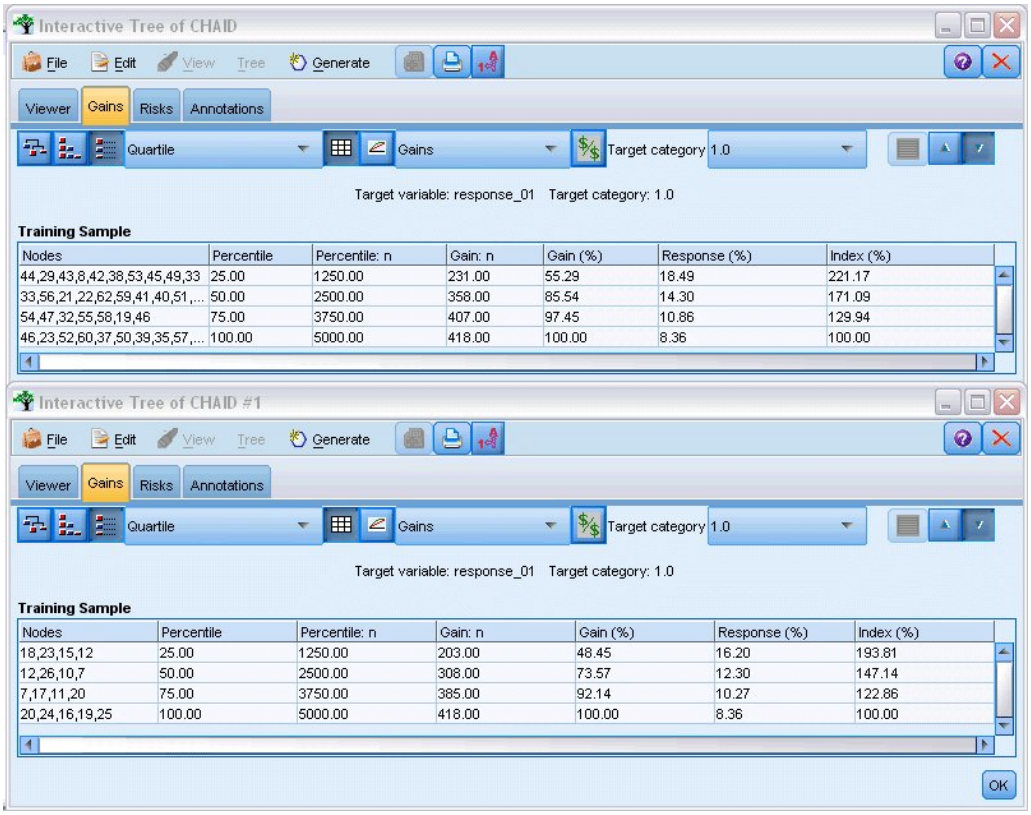
Drugie drzewo zawiera również mniej węzłów drzewa niż pierwsze. Łatwiej jest je zrozumieć. Przed zadecydowaniem o użyciu tej metody należy jednak sprawdzić, czy jest skuteczna i jak wypada w porównaniu z modelem, który używa wszystkich predyktorów.

Porównywanie wyników

Aby porównać dwa wyniki, potrzebna jest miara efektywności. W tym celu użyjemy karty Korzyści w konstruktorze drzewa. Przyjrzymy się wartości **wzrost**, która mierzy, o ile bardziej prawdopodobne jest, że rekordy w węźle przypadają w kategorii przewidywanej przy porównywaniu wszystkich rekordów w zbiorze danych. Na przykład: wartość wzrostu wynosząca 148% wskazuje, że istnieje 1,48 razy większe prawdopodobieństwo, że rekordy w węźle będą przypadać w kategorii docelowej w porównaniu do wszystkich rekordów w zbiorze danych.Wzrost jest określony w kolumnie *Indeks* na karcie Korzyści.

1. W konstruktorze drzewa dla pełnego zestawu predyktorów kliknij kartę Korzyści. Zmień kategorię przewidywaną na 1,0. Zmień wyświetlanie na kwartyle, klikając najpierw przycisk Kwantyle na pasku narzędzi. Następnie z listy rozwijanej wybierz pozycję **Kwartyl** po prawej stronie tego przycisku.

2. Powtórz tę procedurę w konstruktorze drzewa dla zbioru 10 predyktorów, aby dostępne były dwie podobne tabele Korzyści do porównania, jak pokazano to na poniższych rysunkach.



Rysunek 107. Wykresy korzyści dla dwóch modeli CHAID

Każda tabela Korzyści grupuje węzły końcowe swojego drzewa w kwartylach. Aby porównać skuteczność dwóch modeli, oceń wzrost (wartość *Indeks*) dla górnego kwartyla w każdej tabeli.

Kiedy uwzględnione są wszystkie predyktory, model wykazuje wzrost wynoszący 221%. Oznacza to, że jest 2,2 razy bardziej prawdopodobne, że przypadki z charakterystyką taką, jak w tych węzłach odpowiedzą na promocję docelową. Aby zobaczyć, jaka jest to charakterystyka, kliknij, aby wybrać górny wiersz. Następnie przełącz na kartę Przegląd, gdzie powiązane węzły są teraz zaznaczone na czarno. Przeanalizuj drzewo do każdego zaznaczonego węzła końcowego, aby zobaczyć, jak zostały podzielone predyktory. Tylko górny kwartyl zawiera 10 węzłów. Przy przeniesieniu na rzeczywiste modele oceniania 10 różnych profili klientów może sprawiać trudność w zarządzaniu.

Przy uwzględnieniu tylko 10 najważniejszych predyktorów (zidentyfikowanych przez dobór predyktorów) wartość wzrostu wynosi prawie 194%. Mimo że ten model nie jest tak dobry, jak model, który używa wszystkich predyktorów, jest na pewno przydatny. W tym przypadku górny kwartyl zawiera tylko cztery węzły, jest więc prostszy. Dlatego też można określić, że model z doborem predyktorów jest lepszy niż model z wszystkimi predyktorami.

Podsumowanie

Przeanalizujmy korzyści związane z doborem predyktorów. Użycie mniejszej liczby predyktorów jest mniej kosztowne. Oznacza to, że należy zgromadzić, przetworzyć i przekazać do modeli mniejsze ilości danych. Poprawia się czas przetwarzania. W tym przykładzie, nawet z dodatkowym krokiem wyboru predyktorów, budowanie modelu było zauważalnie szybsze dla mniejszego zbioru predyktorów. Dla większych rzeczywistych zbiorów danych oszczędności czasu powinny być znacznie większe.

Użycie mniejszej liczby predyktorów powoduje uproszczone ocenianie. Jak pokazuje przykład, można zidentyfikować tylko cztery profile klientów, którzy prawdopodobnie odpowiedzą na promocję. Należy zauważyć, że przy większej liczbie predyktorów istnieje ryzyko przeuczenia modelu. Prostszy model może lepiej generalizować inne zbiory danych (należy to jednak przetestować, aby upewnić się).

Użytkownik mógł użyć algorytmu budowania drzewa do automatycznego wyboru najważniejszych predyktorów. W tym celu często używany jest algorytm CHAID i możliwe jest rozbudowanie drzewa o kolejne poziomy, kontrolując jego głębokość i złożoność. Węzeł Dobór predyktorów jest jednak szybszy i łatwiejszy w użyciu. Węzeł szereguje wszystkie predyktory w jednym szybkim kroku, pozwalając na sprawną identyfikację najważniejszych zmiennych. Pozwala również na zmianę uwzględnianych predyktorów. Ten przykład można wykonać również dla najważniejszych 15 lub 20 predyktorów zamiast 10, porównując wyniki, aby określić optymalny model.

Rozdział 10. Skracanie łańcucha danych wejściowych (Węzeł rekodowania)

Skracanie łańcucha danych wejściowych (Rekodowanie)

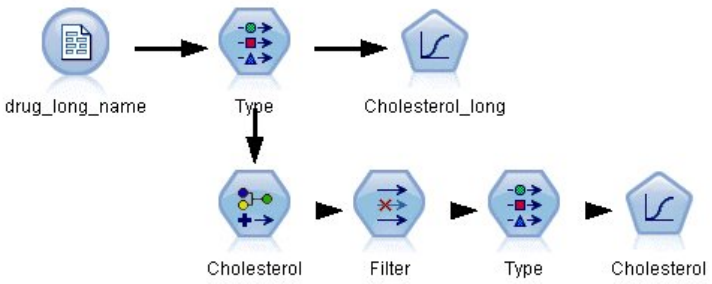
Dla dwumianowej regresji logistycznej i modeli automatycznej klasyfikacji, które zawierają model dwumianowej regresji logistycznej, zmienne łańcuchowe są ograniczone do maksymalnie ośmiu znaków. Kiedy łańcuchy mają więcej niż osiem znaków, można je rekodować, używając węzła rekodowania.

W tym przykładzie zastosowano strumień o nazwie *reclassify_strings.str*, który odwołuje się do pliku danych o nazwie *drug_long_name*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *reclassify_strings.str* znajduje się w katalogu *streams*.

Ten przykład koncentruje się na małej części strumienia, aby pokazać rodzaje błędów, które można wygenerować przez zbyt długie łańcuchy, i wyjaśnia, jak używać węzła rekodowania do zmiany szczegółów łańcucha do dopuszczalnej długości. Mimo że ten przykład używa dwumianowego węzła regresji logistycznej, w takim samym stopniu dotyczy przypadku użycia węzła automatycznej klasyfikacji do generowania modelu dwumianowej regresji logistycznej.

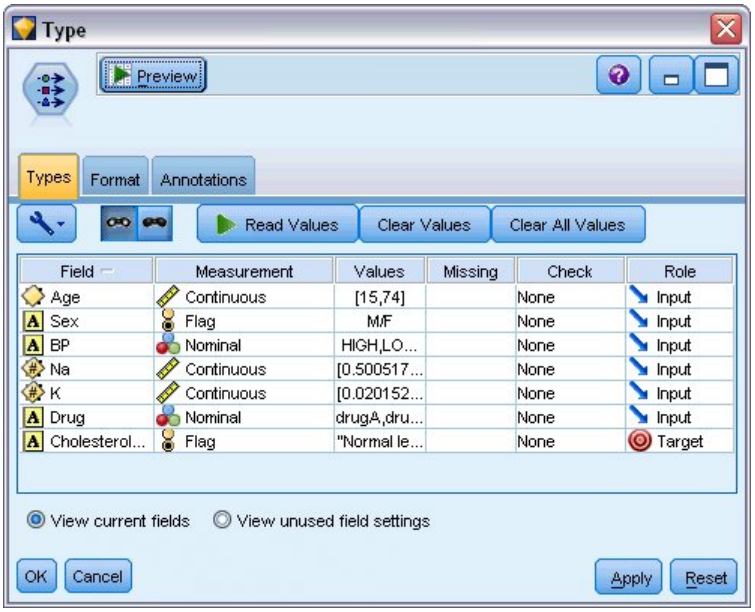
Rekodowanie danych

1. Używając węzła źródłowego plików zmiennych, podłącz zbiór danych *drug_long_name* w folderze *Demos*.



Rysunek 108. Przykładowy strumień przedstawiający rekodowanie dwumianowej regresji logistycznej

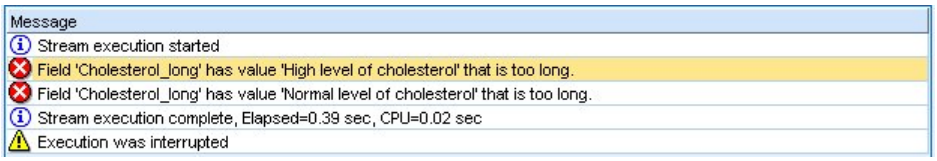
2. Dodaj węzeł typu do węzła źródłowego i wybierz zmienną **Cholesterol_long** jako zmienną przewidywaną.
3. Dołącz węzeł regresji logistycznej do węzła typu.
4. W węźle regresji logistycznej kliknij kartę Model i wybierz procedurę **Dwumianowa**.



Rysunek 109. Szczegóły długiego łańcucha zmiennej Cholesterol_long

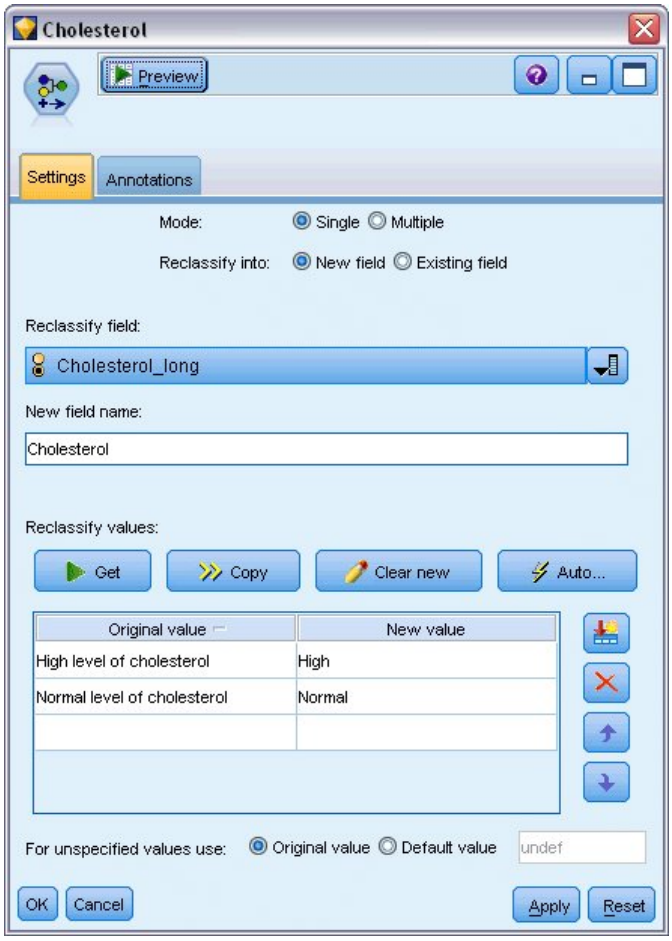
5. Po uruchomieniu węzła regresji logistycznej w strumieniu *reclassify_strings.str*, zostanie wyświetlony komunikat ostrzegający, że wartości łańcucha **Cholesterol_long** są zbyt długie.

   W przypadku napotkania tego typu komunikatu o błędzie należy postępować zgodnie z procedurą wyjaśnioną w pozostałej części tego przykładu, aby zmodyfikować dane.



Rysunek 110. Komunikat o błędzie wyświetlany podczas wykonywania dwumianowego węzła regresji logistycznej

6. Dołącz węzeł rekodowania do węzła typu.
7. W obszarze Rekoduj zmienną wybierz **Cholesterol_long**.
8. Wpisz **Cholesterol** jako nazwę nowej zmiennej.
9. Kliknij przycisk **Uzyskaj**, aby dodać wartości **Cholesterol_long** do oryginalnej kolumny wartości.
10. W kolumnie nowej wartości wpisz **High** obok oryginalnej wartości **High level of cholesterol** i **Normal** obok oryginalnej wartości **Normal level of cholesterol**.

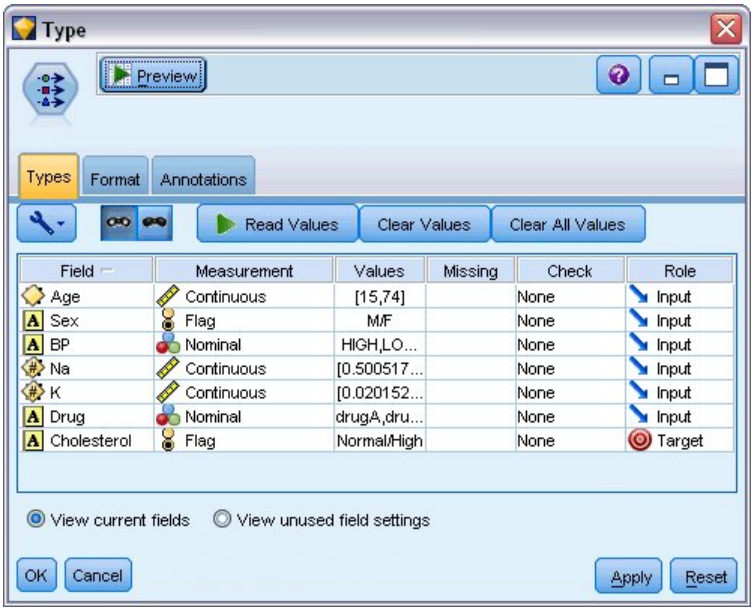


Rysunek 111. Rekodowanie długich łańcuchów

11. Dodaj węzeł filtrowania do węzła rekodowania.
12. W kolumnie Filtrowanie kliknij, aby usunąć zmienną **Cholesterol_long**.

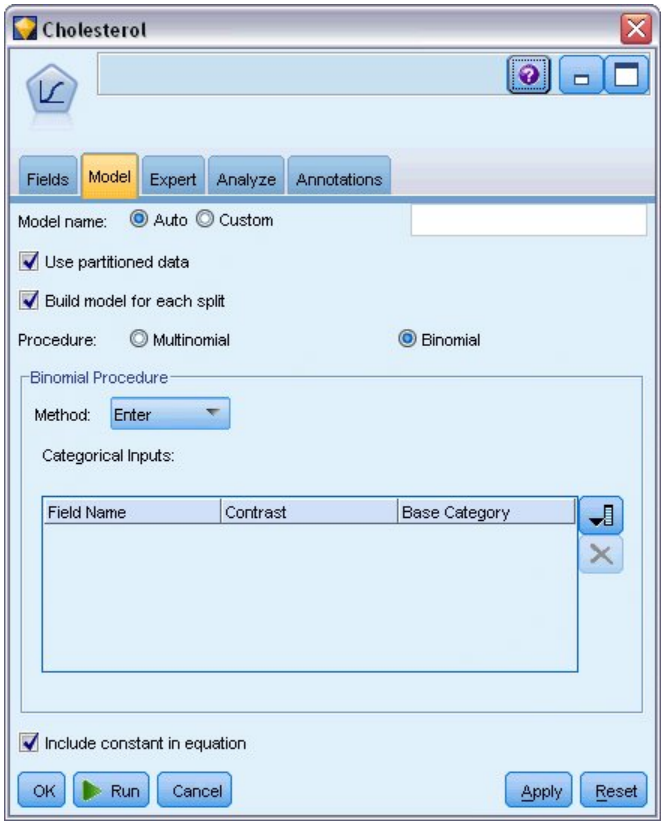
Rysunek 112. Filtrowanie zmiennej Cholesterol_long z danych

13. Dodaj węzeł typu do węzła filtrowania i wybierz zmienną **Cholesterol** jako zmienną przewidywaną.



Rysunek 113. Szczegóły krótkiego łańcucha zmiennej Cholesterol

14. Dołącz węzeł logistyczny do węzła typu.
15. W węźle logistycznym kliknij kartę Model i wybierz procedurę **Dwumianowa**.
16. Można teraz wykonać dwumianowy węzeł logistyczny i wygenerować model bez wyświetlania komunikatu o błędzie.



Rysunek 114. Wybieranie procedury dwumianowej

Ten przykład przedstawia tylko część strumienia. Jeśli potrzebujesz dalszych informacji o typach strumieni, w których potrzebne może być rekodowanie długich łańcuchów, dostępne są następujące przykłady:

- Węzeł Auto Klasyfikacja. Więcej informacji można znaleźć w temacie “Modelowanie odpowiedzi klienta (Auto Klasyfikacja)” na stronie 37.
- Dwumianowy węzeł regresji logistycznej. Więcej informacji można znaleźć w temacie Rozdział 13, “Poziom odejścia usług telekomunikacyjnych (Dwumianowa regresja logistyczna)”, na stronie 139.

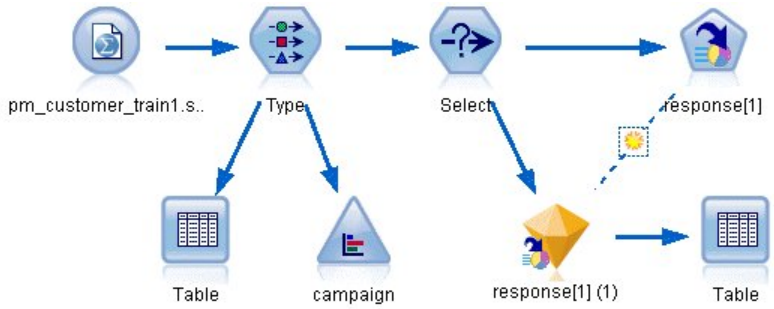
Więcej informacji dotyczących używania programu IBM SPSS Modeler, takich jak podręcznik użytkownika, informacje o węzłach i podręcznik dotyczący algorytmów można znaleźć w katalogu *\Documentation* na dysku instalacyjnym.

Rozdział 11. Modelowanie odpowiedzi klienta (Lista decyzyjna)

Algorytm Lista decyzyjna generuje reguły wskazujące niższe lub wyższe prawdopodobieństwo danego wyniku binarnego (tak lub nie). Modele Lista decyzyjna są szeroko stosowane w zarządzaniu relacjami z klientami, np. w telefonicznych centrach obsługi lub zastosowaniach marketingowych.

W tym przykładzie fikcyjna firma chce uzyskać wyższy zysk w przyszłych kampaniach marketingowych, dobierając właściwą ofertę dla każdego klienta. Przykład używa modelu Lista decyzyjna, aby na podstawie poprzednich promocji określić cechy klientów, którzy z największym prawdopodobieństwem pozytywnie zareagują na ofertę, i wygenerować listę mailingową na podstawie wyników.

Modele list decyzyjnych są wyjątkowo dobrze przystosowane do modelowania interaktywnego, pozwalając na dostosowanie parametrów w modelu i natychmiastowe wyświetlenie wyników. Inne podejście, które pozwala na automatyczne tworzenie różnych modeli i szeregowanie wyników to węzeł Auto Klasyfikacja.

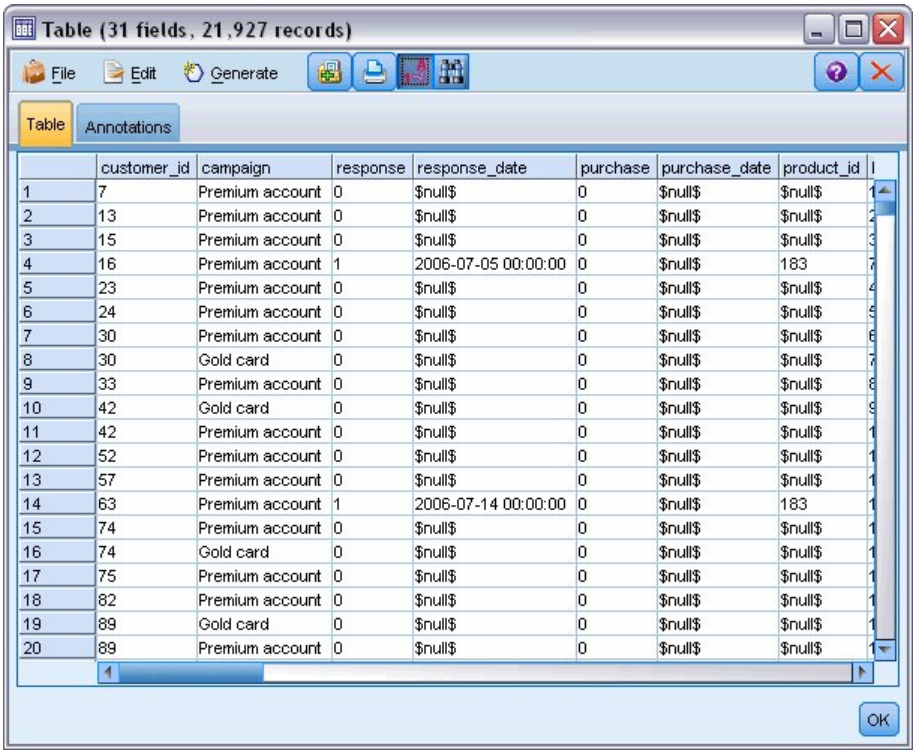


Rysunek 115. Strumień przykładowy Lista decyzyjna

W tym przykładzie zastosowano strumień o nazwie *pm_decisionlist.str*, który odwołuje się do pliku danych o nazwie *pm_customer_train1.sav*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *pm_decisionlist.str* znajduje się w katalogu *streams*.

Dane historyczne

Plik *pm_customer_train1.sav* zawiera dane historyczne śledzące oferty złożone określonym klientom w minionych kampaniach, zgodnie ze wskazaniem wartości w polu *campaign*. Największa liczba rekordów przypada w kampanii *Premium account*.



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

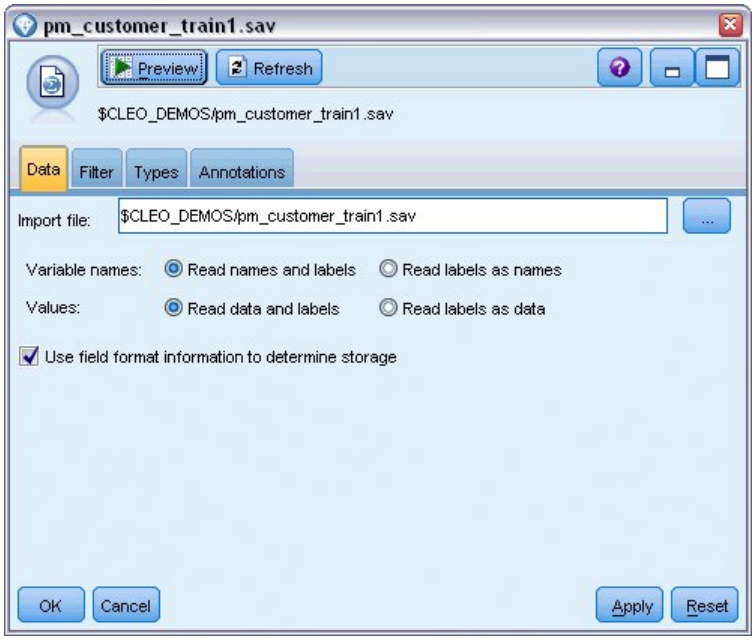
Rysunek 116. Dane dotyczące wcześniejszych promocji

Wartości zmiennej *campaign* są w rzeczywistości zakodowane jako liczby całkowite w danych z etykietami zdefiniowanymi w węźle Typ (na przykład *2 = Premium account*). Można przełączyć wyświetlanie etykiet wartości w tabeli za pomocą paska narzędzi.

Plik zawiera również różne zmienne zawierające informacje demograficzne i finansowe o każdym z klientów, których można użyć do zbudowania lub nauki modelu, który przewiduje współczynniki odpowiedzi dla różnych grup na podstawie określonych cech.

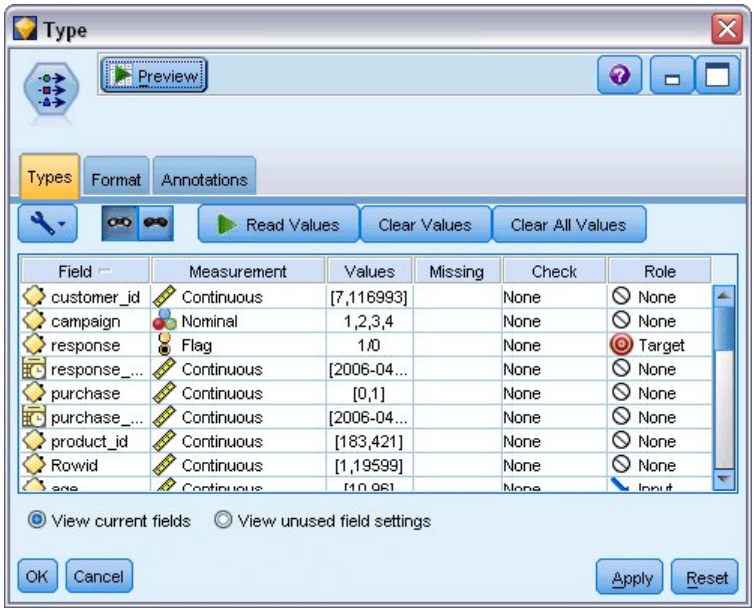
Tworzenie strumienia

1. Dodaj węzeł Plik Statistics wskazujący na plik *pm_customer_train1.sav* znajdujący się w folderze *Demos* w folderze instalacji IBM SPSS Modeler. (W ścieżce do pliku można określić parametr $CLEO_DEMOS/ jako skrót do tego folderu).



Rysunek 117. Odczytywanie danych

2. Dodaj węzeł typu i wybierz zmienną *response* jako zmienną przewidywaną (Rola = **Przewidywana**). Poziom pomiaru dla tej zmiennej ustaw na **Flaga**.



Rysunek 118. Ustawianie poziomu pomiaru i roli

3. Ustaw rolę na **Brak** dla następujących zmiennych: *customer_id*, *campaign*, *response_date*, *purchase*, *purchase_date*, *product_id*, *Rowid* i *X_random*. Wszystkie te zmienne mają zastosowanie w danych, ale nie będą używane w budowaniu rzeczywistego modelu.
4. Kliknij przycisk **Odczytaj wartości** w węźle Typ, aby zapewnić, że wartości są określone.

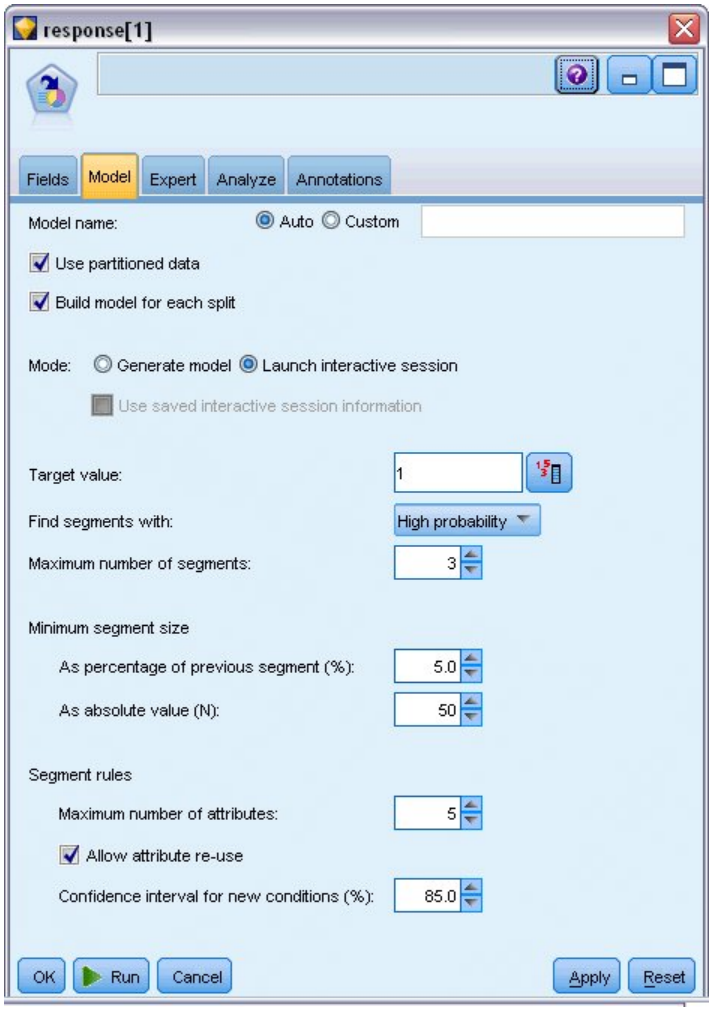
Mimo że dane zawierają informacje o czterech różnych kampaniach, każdorazowo skoncentrujemy analizę na jednej kampanii. Ponieważ największa liczba rekordów przypada dla kampanii Premium (z kodem w danych *campaign=2*),

można użyć węzła selekcji, aby uwzględnić tylko te rekordy w strumieniu.

Rysunek 119. Wybieranie rekordów dla pojedynczej kampanii

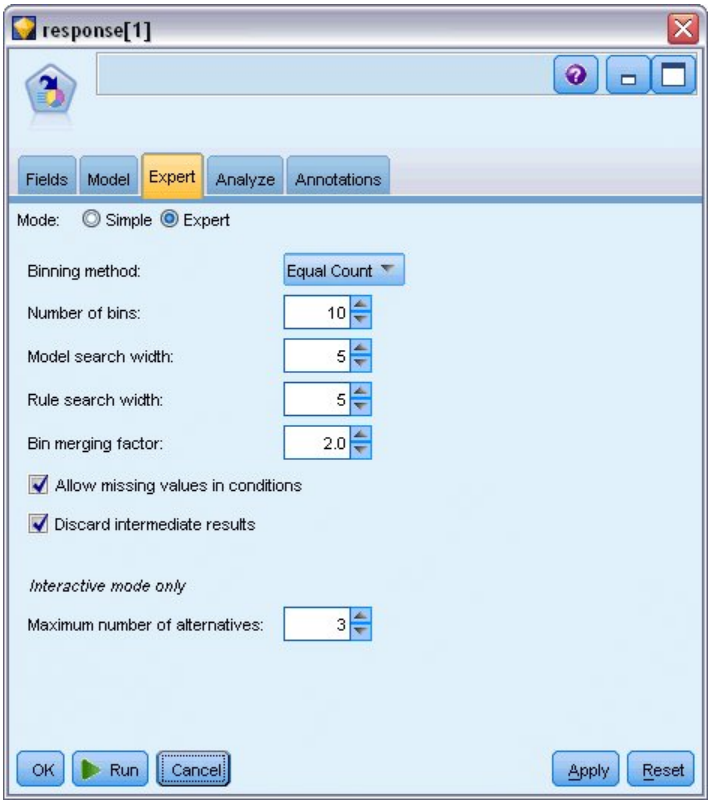
Tworzenie modelu

1. Załącz węzeł Lista decyzyjna do strumienia. Na karcie Model ustaw pozycję **Wartość przewidywana** na 1, aby wskazać wynik, którego poszukujesz. W tym przypadku szukamy klientów, którzy odpowiedzieli *Tak* na poprzednią ofertę.



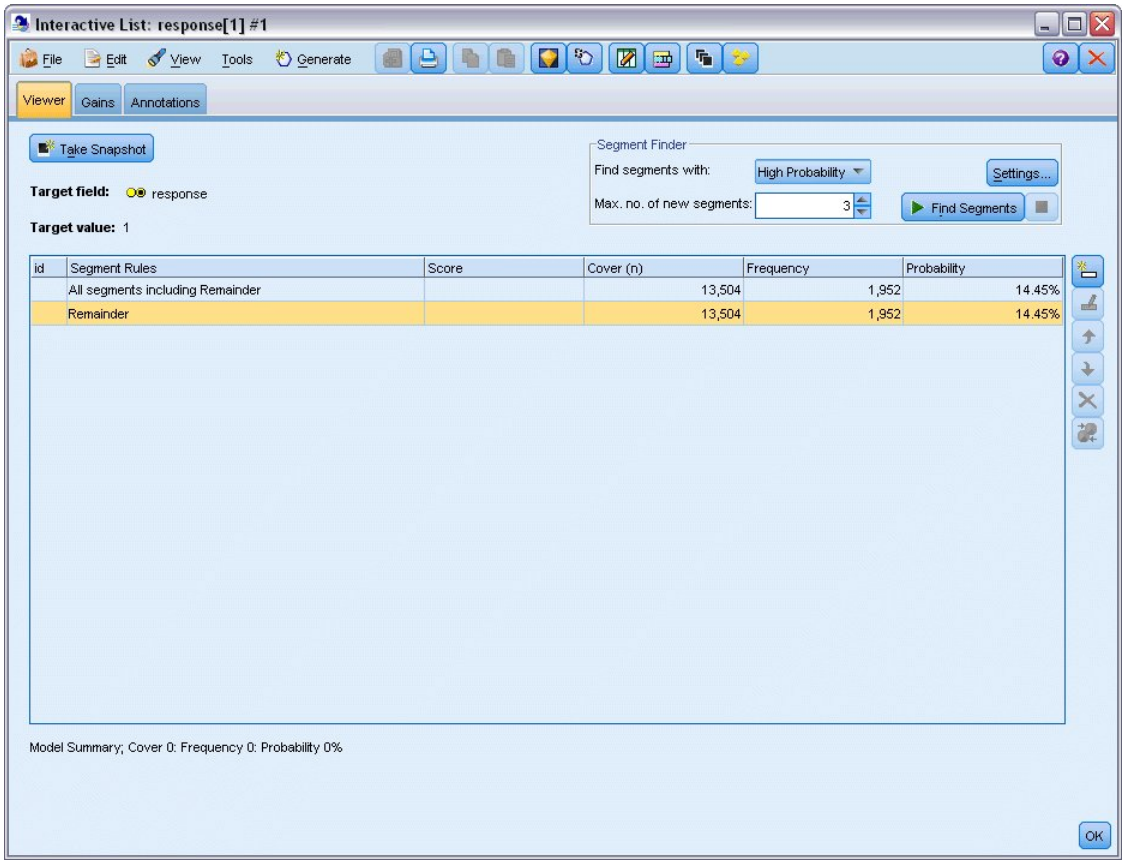
Rysunek 120. Karta Model węzła Lista decyzyjna

2. Wybierz opcję **Drzewo interakcyjne**.
3. Aby zachować prostotę modelu dla celów tego przykładu, ustaw maksymalną liczbę segmentów na 3.
4. Zmień przedział ufności dla nowych warunków na 85%.
5. Na karcie Zaawansowany ustaw **Tryb** na **Zaawansowany**.



Rysunek 121. Karta Zaawansowany węzła Lista decyzyjna

6. Zwiększ wartość parametru **Maksymalna liczba alternatyw** do 3. Ta opcja działa łącznie z ustawieniem **Drzewo interakcyjne** wybraną na karcie Model.
7. Kliknij przycisk **Wykonaj**, aby wyświetlić przeglądarkę Lista interaktywna.

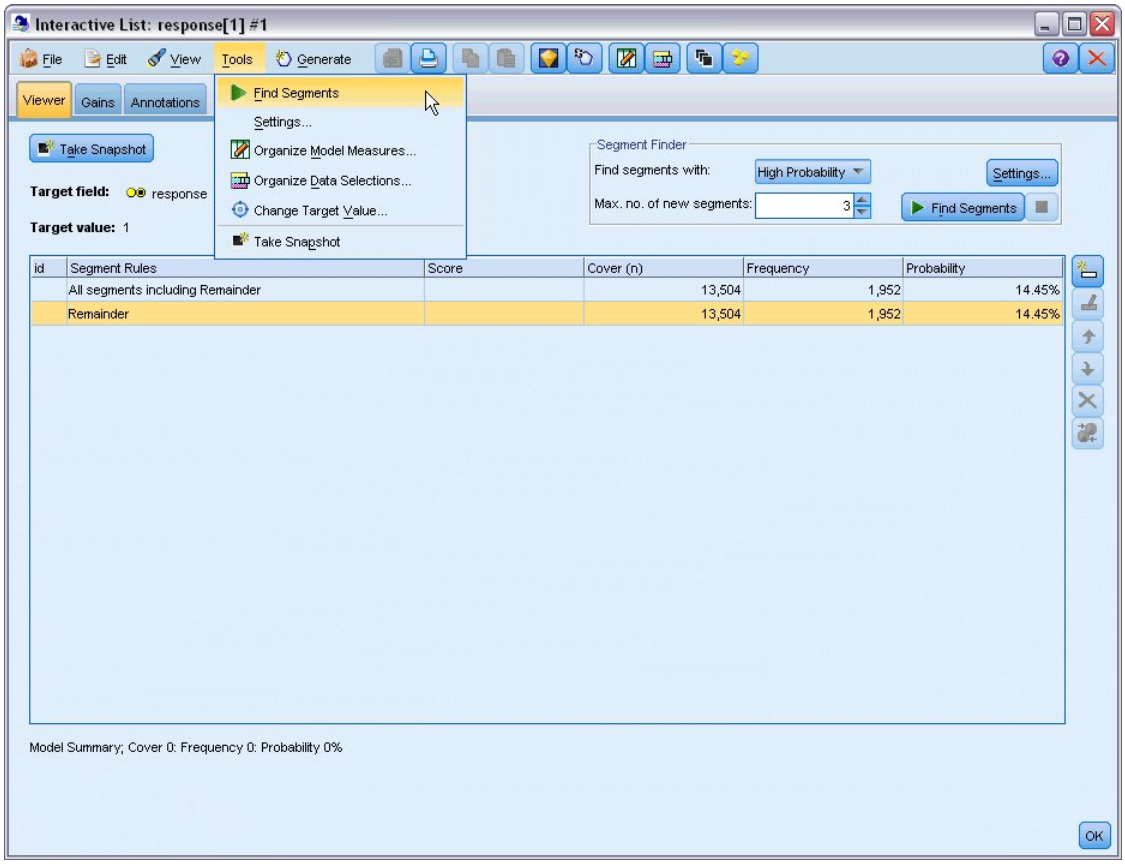


Rysunek 122. Przeglądarka Lista interaktywna

Ponieważ żadne segmenty nie zostały jeszcze zdefiniowane, wszystkie rekordy są zakwalifikowane jako pozostałość. Z 13 504 rekordów w próbce 1952 odpowiedziało *Tak*, co daje ogólny współczynnik trafień wynoszący 14,45%. Użytkownik chce poprawić ten współczynnik, identyfikując segmenty klientów, którzy z większym (lub mniejszym) prawdopodobieństwem odpowiedzą pozytywnie.

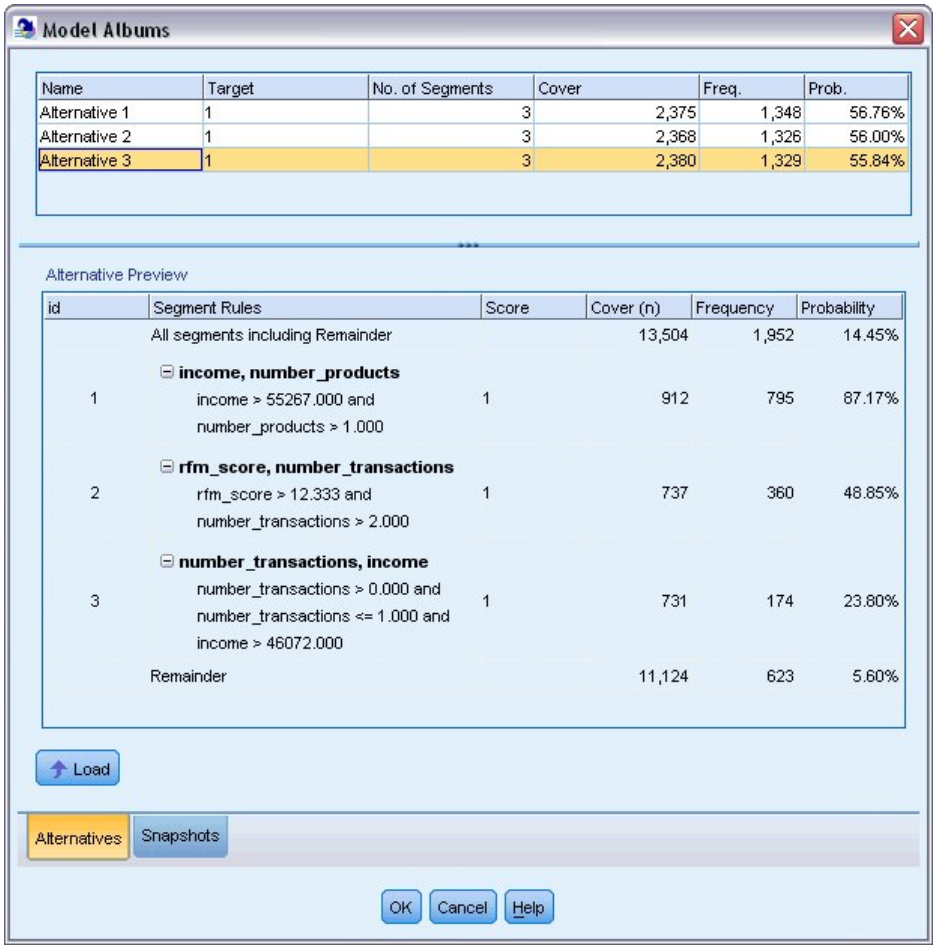
8. W przeglądarce Lista interaktywna wybierz z menu:

   Narzędzia > Znajdź segmenty



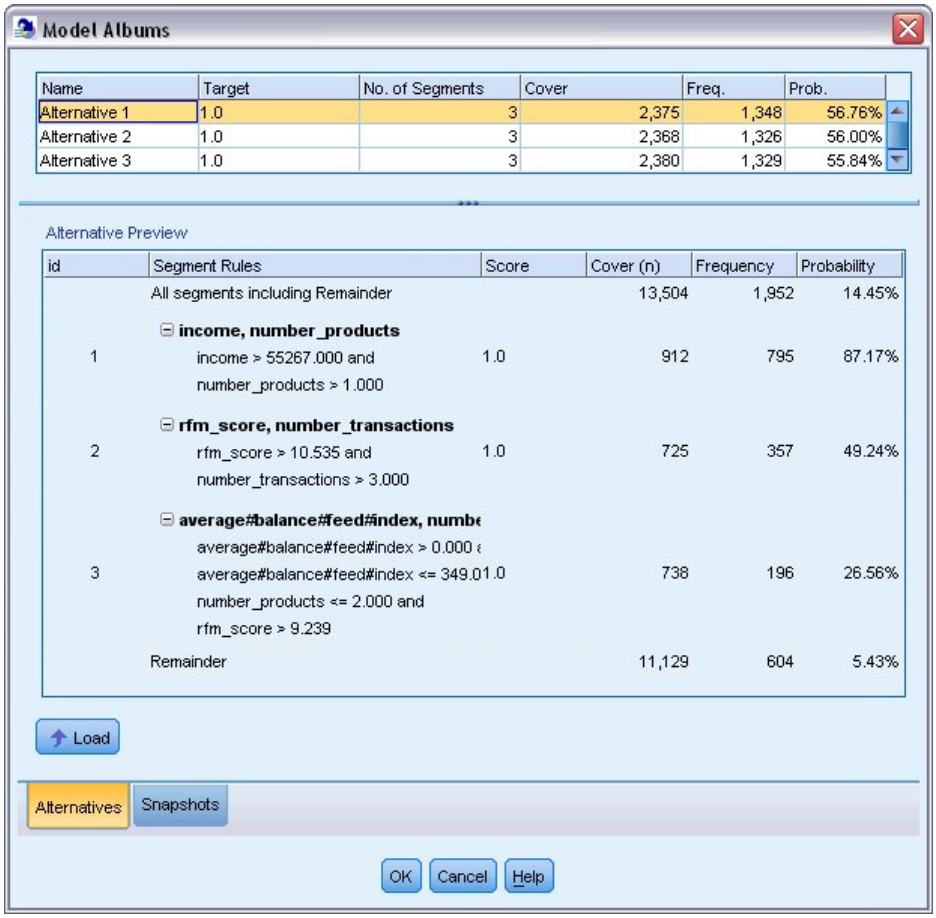
Rysunek 123. Przeglądarka Lista interaktywna

Powoduje to uruchomienie domyślnego zadania eksploracji na podstawie ustawień określonych w węźle Lista decyzyjna. Ukończone zadanie zwraca trzy alternatywne modele, które są wymienione na karcie Alternatywne modele okna dialogowego Albumy modelu.



Rysunek 124. Dostępne modele alternatywne

9. Wybierz pierwszą alternatywę z listy — jej szczegóły są przedstawione w panelu Podgląd alternatywny.



Rysunek 125. Wybrany model alternatywny

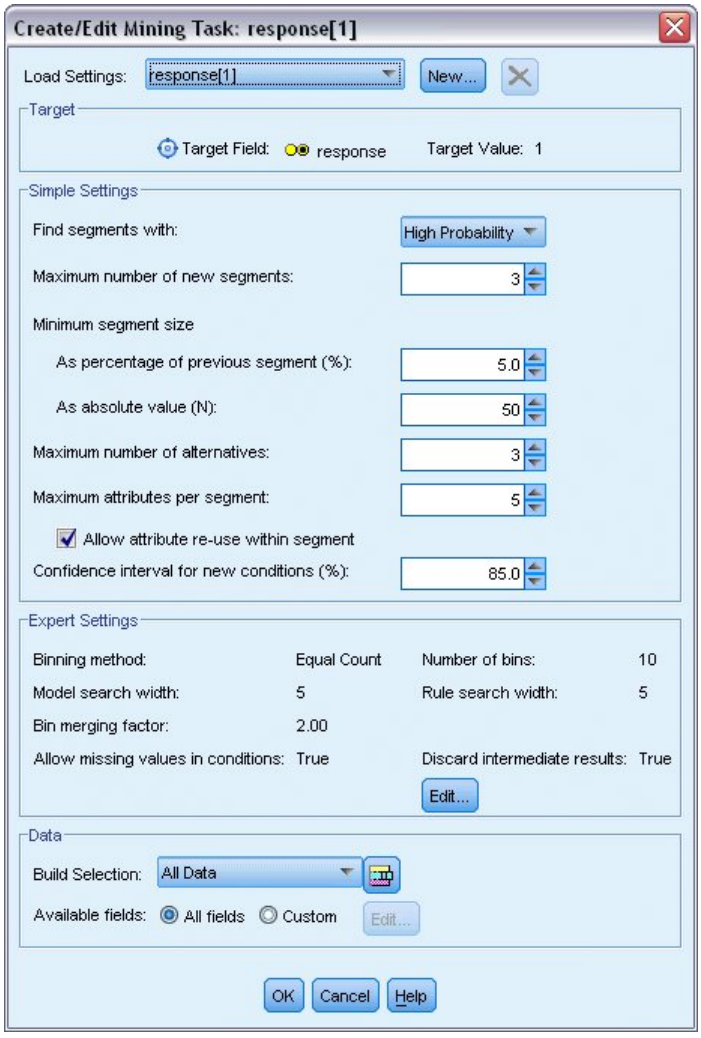
Panel Podgląd alternatywny pozwala na szybkie przeglądanie wielu alternatyw bez zmiany modelu roboczego, ułatwiając eksperymentowanie z różnymi podejściami.

Uwaga: Aby model był lepiej widoczny, możesz zmaksymalizować panel Podgląd alternatywny w oknie dialogowym, tak jak pokazano na przykładzie. Można to zrobić, przeciągając krawędź panelu.

Używając reguł opartych na predyktorach, takich jak dochód, liczba transakcji miesięcznie i ocena RFM, model identyfikuje segmenty ze wskaźnikami odpowiedzi wyższymi niż dla ogólnej próbki. Po połączeniu segmentów model sugeruje, że można poprawić współczynnik trafień do poziomu 56,76%. Model obejmuje jednak tylko małą część ogólnej próbki, pozostawiając ponad 11 000 rekordów zakwalifikowanych jako pozostałość przy kilkuset trafieniach.Użytkownik poszukuje modelu, który zapewni więcej trafień, wciąż eliminując segmenty o niskiej efektywności.

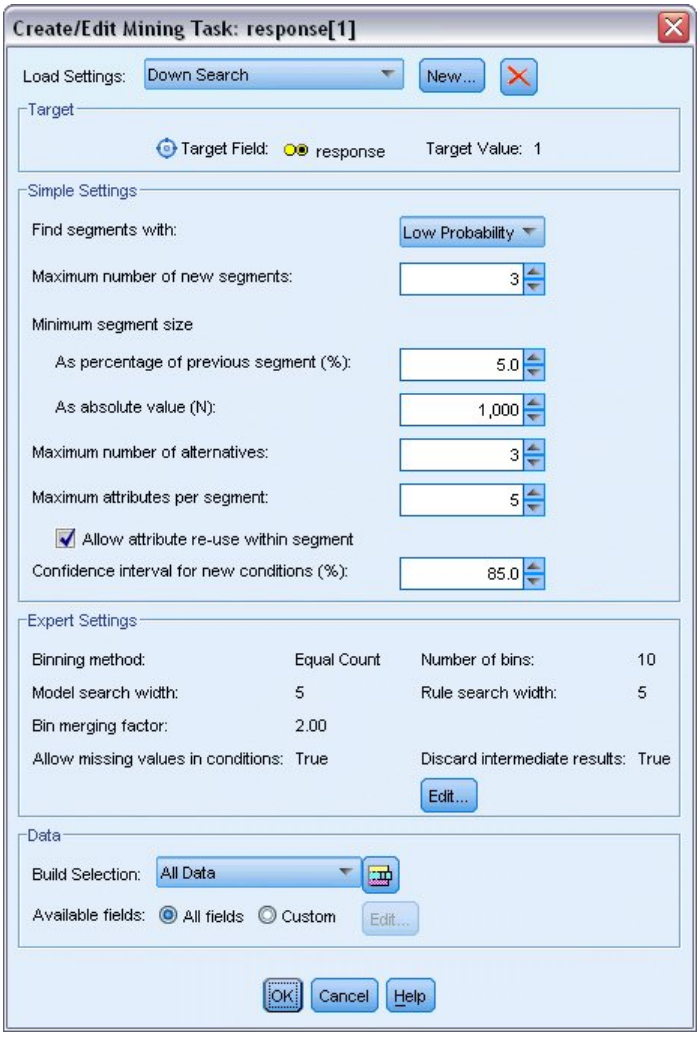
10. Aby spróbować innego podejścia w modelowaniu, wybierz kolejno następujące pozycje menu:

    Narzędzia > Ustawienia



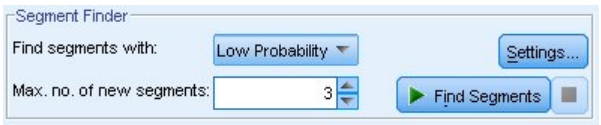
Rysunek 126. Okno dialogowe tworzenia/edycji zadania eksploracji

11. Kliknij przycisk **Nowy** (w prawym górnym rogu), aby utworzyć drugie zadanie eksploracji, i określ *wyszukiwanie w dół* jako nazwę zadania w oknie dialogowym Nowe ustawienia.



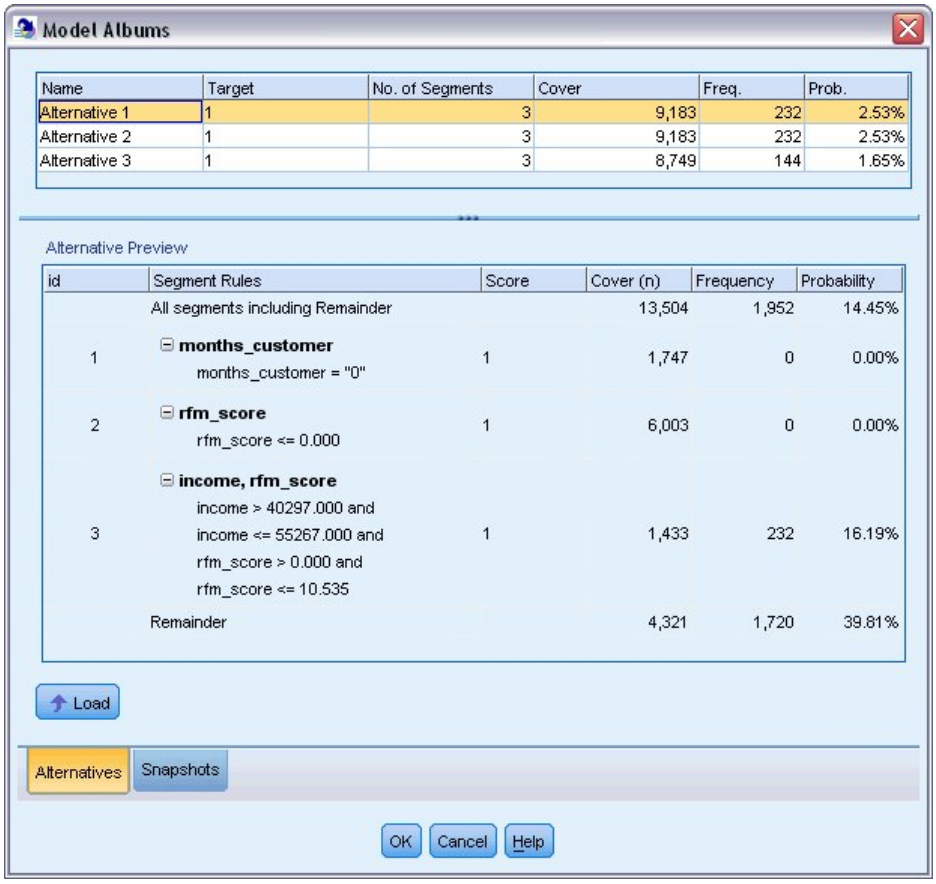
Rysunek 127. Okno dialogowe tworzenia/edycji zadania eksploracji

12. Zmień kierunek wyszukiwania dla zadania na **Niskim prawdopodobieństwem**. Spowoduje to, że algorytm wyszuka segmenty z *najniższym* wskaźnikiem odpowiedzi zamiast z najwyższym.
13. Zwiększ minimalną wielkość segmentu do 1000. Kliknij przycisk **OK**, aby powrócić do przeglądarki Lista interaktywna.
14. W przeglądarce Lista interaktywna upewnij się, że panel *Wyszukiwanie segmentu* wyświetla szczegóły nowego zadania, i kliknij opcję **Znajdź segmenty**.



Rysunek 128. Znajdowanie segmentów w nowym zadaniu eksploracji

Zadanie zwraca nowy zbiór alternatywnych modeli, które są wyświetlane na karcie Alternatywne modele okna dialogowego Albumy modelu i można je przeglądać w taki sam sposób, jak poprzednie wyniki.

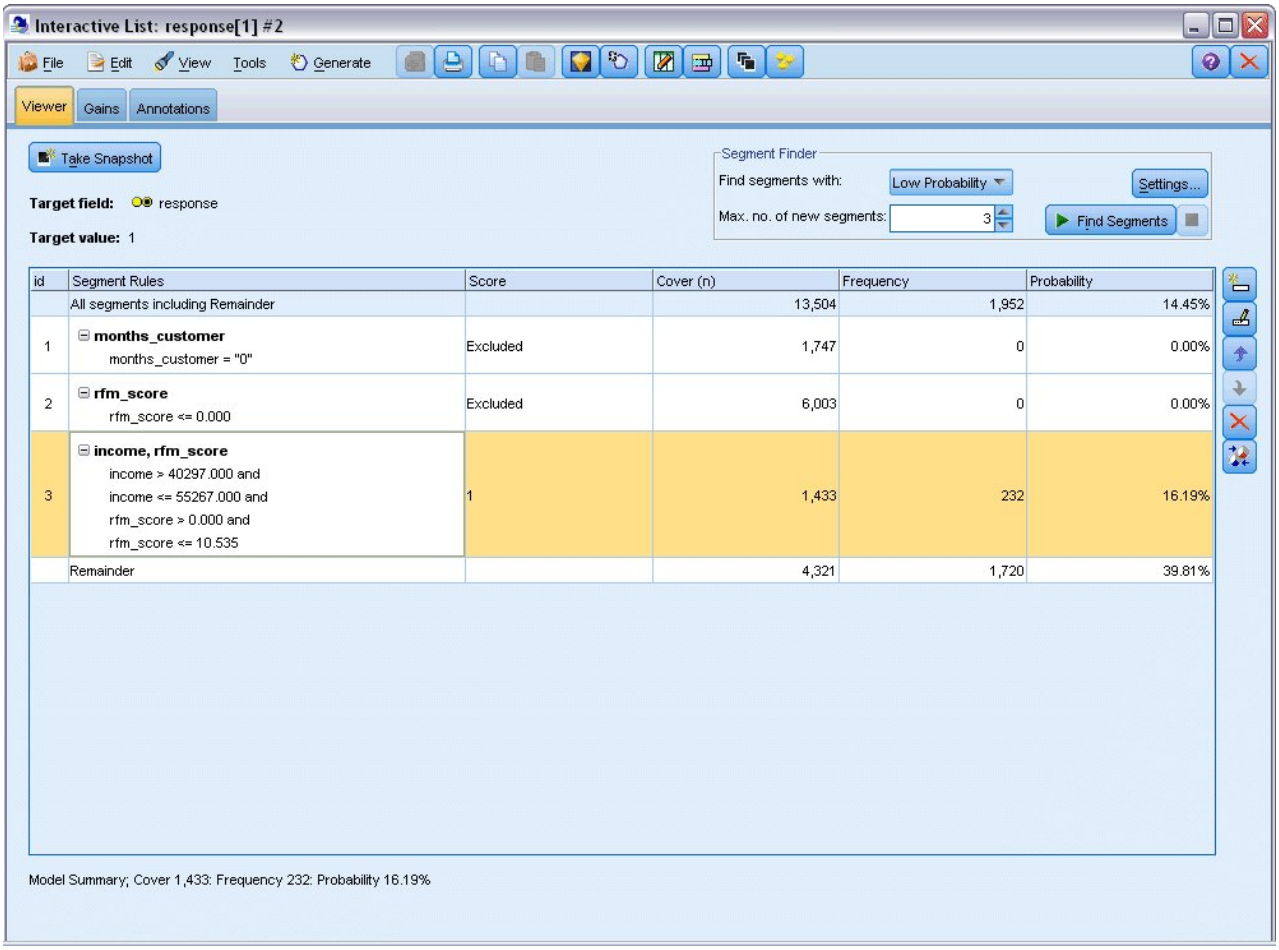


Rysunek 129. Wyniki modelu wyszukiwania w dół

Tym razem każdy model identyfikuje segmenty z niskimi wskaźnikami odpowiedzi zamiast z wysokimi. Patrząc na pierwszą alternatywę, tylko wyłączenie tych segmentów spowoduje zwiększenie współczynnika trafień dla pozostałości do poziomu 39,81%. Jest to niższa wartość niż we wcześniejszym modelu, ale ma większy zasięg (co oznacza większą ilość trafień).

Łącząc te dwa podejścia (używając wyszukiwania niskiego prawdopodobieństwa do usunięcia nieużytecznych rekordów, a następnie używając wyszukiwania wysokiego prawdopodobieństwa), możliwe może być poprawienie tego wyniku.

15. Kliknij opcję **Wczytaj**, aby był to model roboczy (pierwsza alternatywa wyszukiwania w dół), a następnie kliknij przycisk **OK**, aby zamknąć okno dialogowe Albumy modelu.

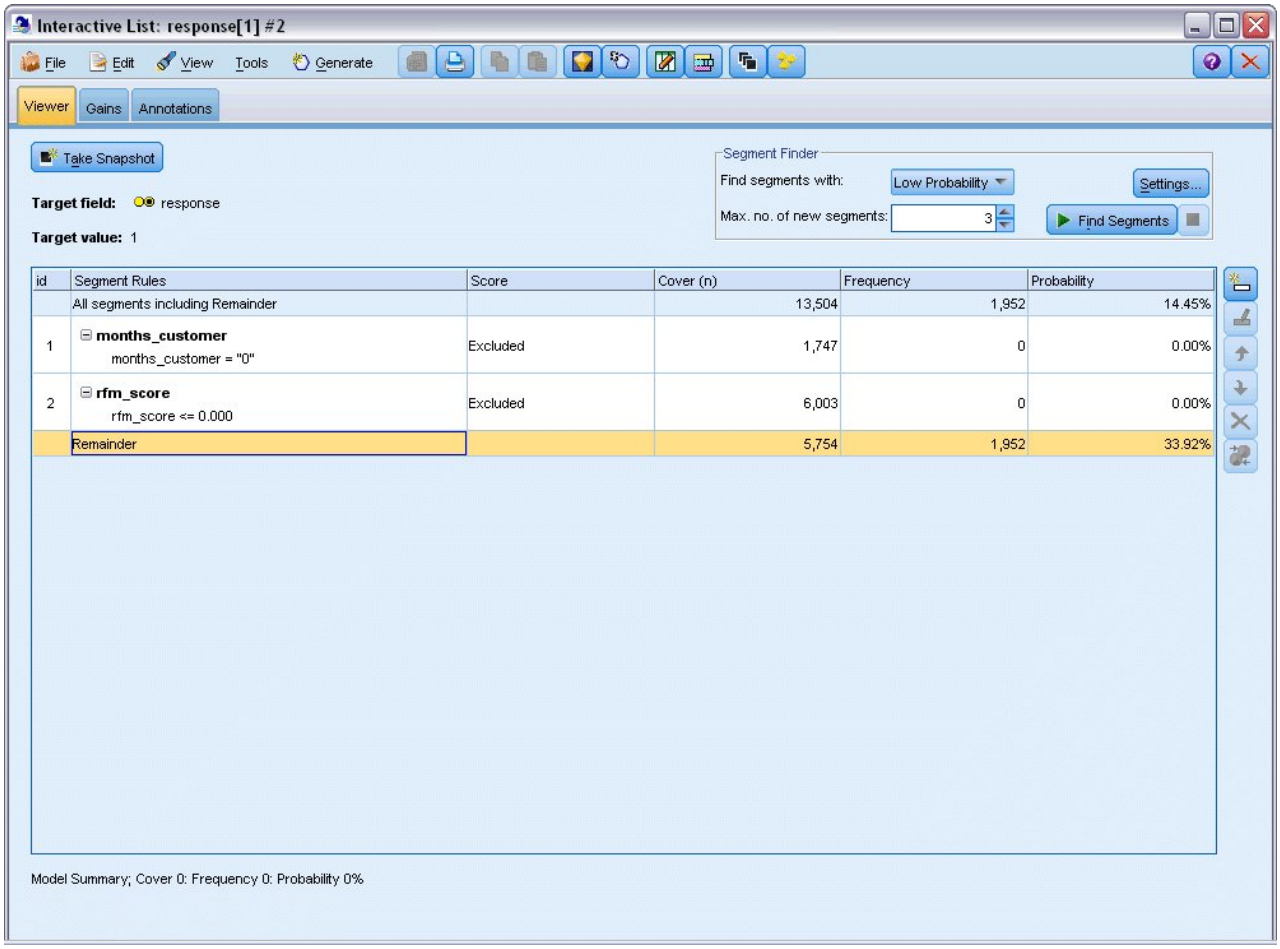


Rysunek 130. Wyłączanie segmentu

16. Kliknij prawym przyciskiem dwa pierwsze segmenty i wybierz opcję **Wyklucz segment**. Razem te segmenty obejmują prawie 8000 rekordów zawierających zero trafień, więc uzasadnione jest wyłączenie ich z przyszłych ofert. (Wyłączone segmenty będą oceniane jako wartości null, aby to wskazać).
17. Kliknij trzeci segment prawym przyciskiem myszy i wybierz opcję **Usuń segment**. Przy 16,19% współczynnik trafień dla tego segmentu nie różni się wiele od współczynnika bazowego 14,45%, nie dodaje więc wystarczających informacji, aby uzasadnić zachowanie go.

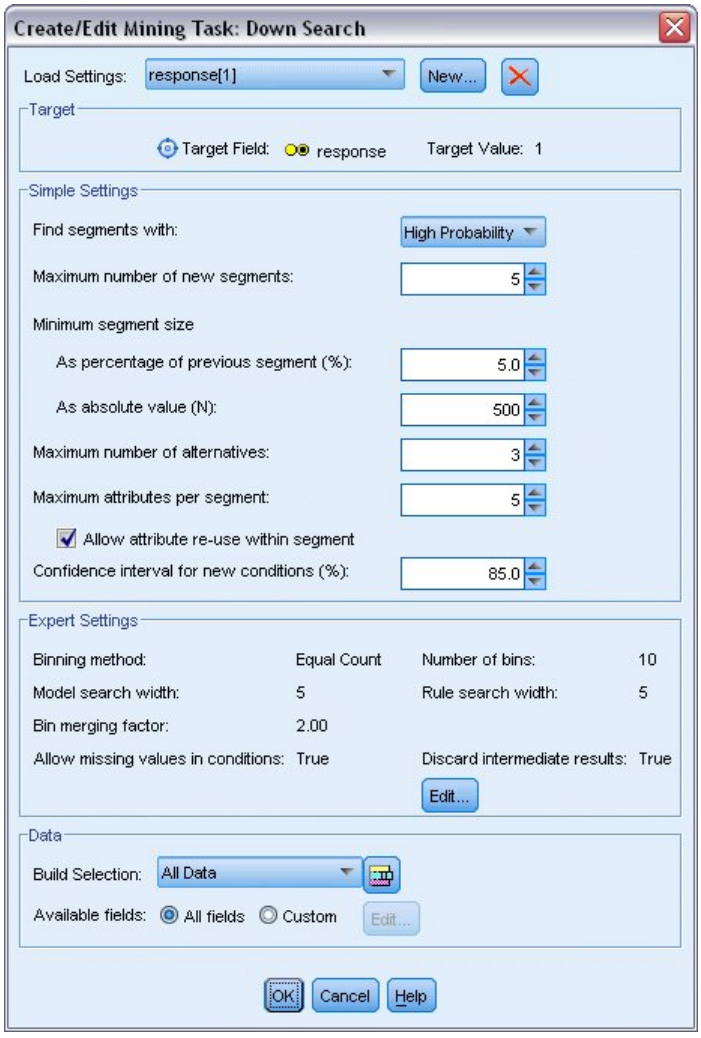
    Uwaga: Usunięcie segmentu nie oznacza tego samego, co jego wyłączenie. Wyłączenie segmentu po prostu zmienia sposób oceniania, podczas gdy usunięcie go całkowicie usuwa segment z modelu.

    Wyłączając segmenty z najgorszymi wynikami, można wyszukać w pozostałości wysoko wydajne segmenty.
18. Kliknij wiersz pozostałości w tabeli, aby go wybrać i aby następne zadanie eksploracji dotyczyło tylko pozostałości.



Rysunek 131. Wybieranie segmentu

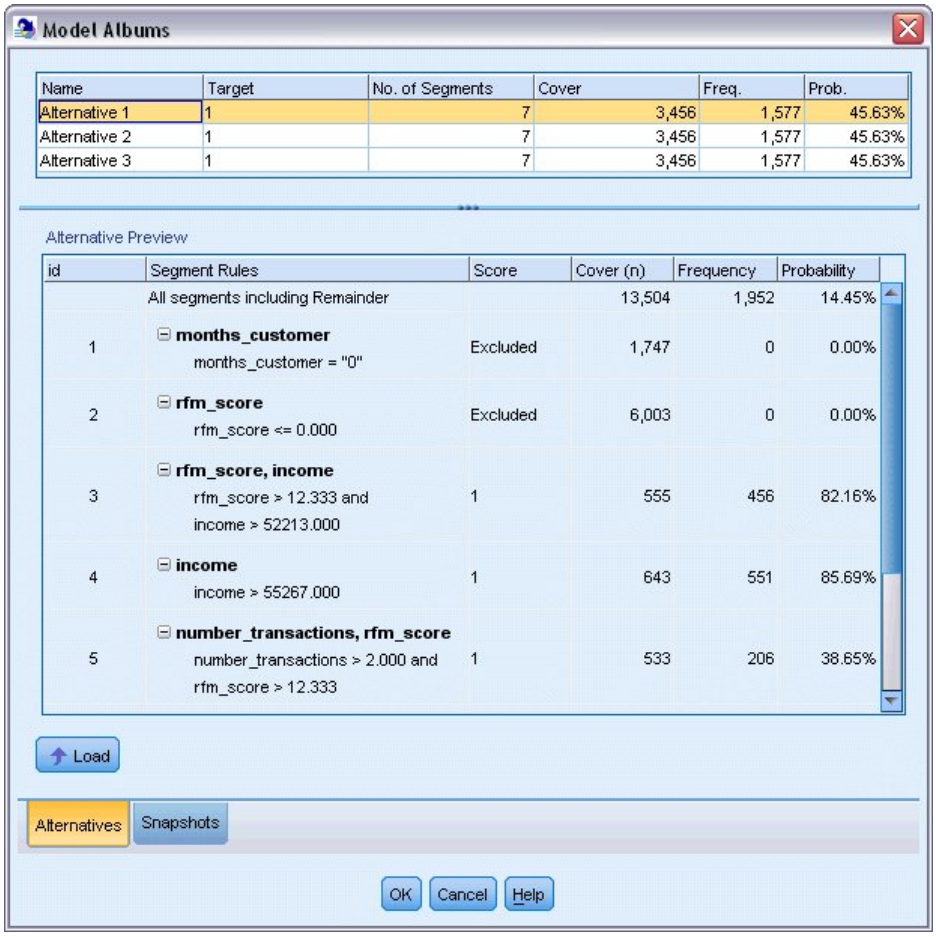
19. Przy wybranej pozostałości kliknij opcję **Ustawienia**, aby otworzyć ponownie okno dialogowe Utwórz/Edytuj zadanie eksploracji.
20. W górnej części, w obszarze **Ustawienia czytania**, wybierz domyślne zadanie eksploracji: **response[1]**.
21. Edytuj obszar **Ustawienia podstawowe**, aby zwiększyć liczbę nowych segmentów do 5 i minimalną wielkość segmentu do 500.
22. Kliknij przycisk **OK**, aby powrócić do przeglądarki Lista interaktywna.



Rysunek 132. Wybieranie domyślnego zadania eksploracji

23. Kliknij opcję **Znajdź segmenty**.

    Powoduje to wyświetlenie kolejnego zestawu alternatywnych modeli. Przekazywanie wyników z jednego zadania eksploracji do innego sprawia, że te ostatnie modele zawierają mieszankę segmentów o wysokiej i niskiej efektywności. Segmenty z niskimi wskaźnikami odpowiedzi zostały wykluczone, co oznacza, że zostaną ocenione jako wartości null, podczas gdy uwzględnione segmenty zostaną ocenione jako 1. Ogólne statystyki odzwierciedlają te wyłączenia, przy pierwszym modelu alternatywnym wykazującym współczynnik trafień 45,63% z wyższym pokryciem (1577 trafień z 3456 rekordów) niż jakikolwiek poprzedni model.

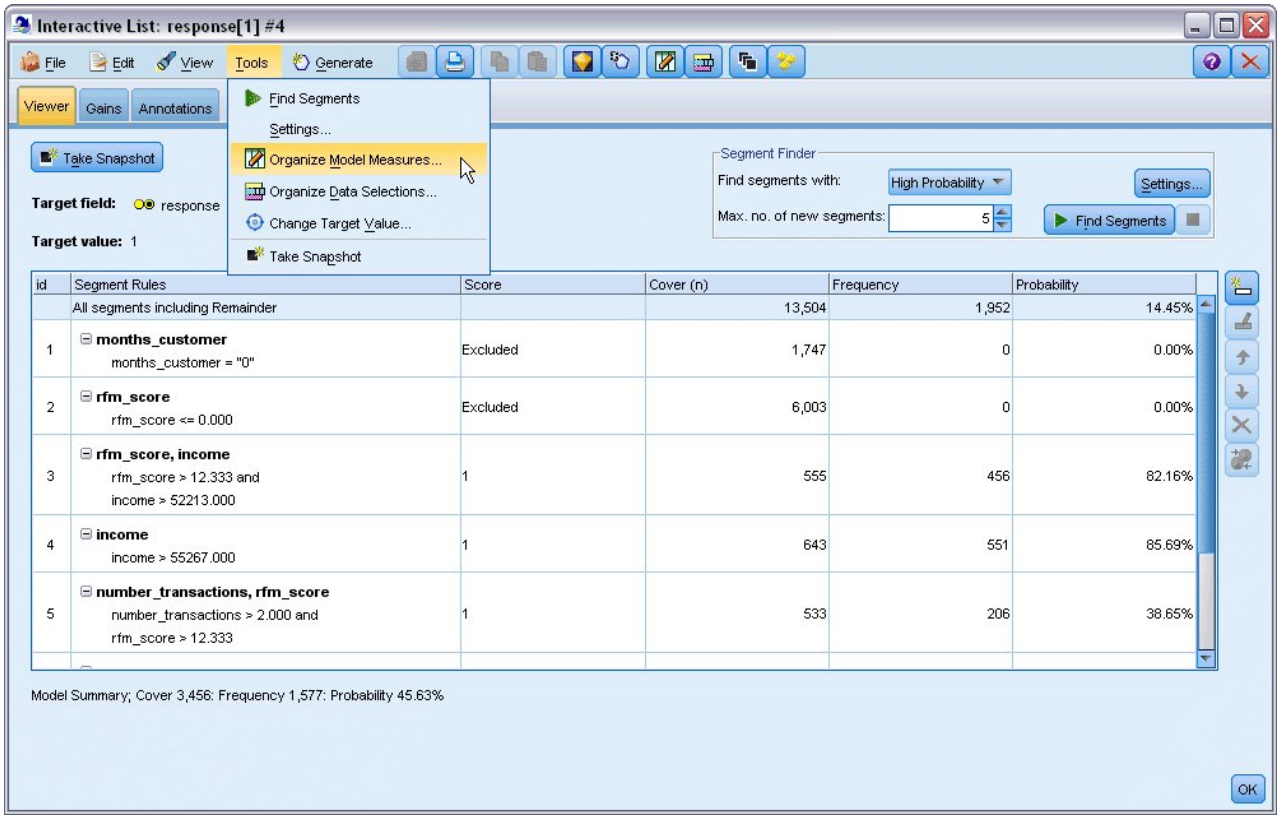


Rysunek 133. Alternatywy dla modelu połączonego

24. Przejrzyj pierwszą alternatywę, a następnie kliknij przycisk **Wczytaj**, aby ustawić ją jako model roboczy.

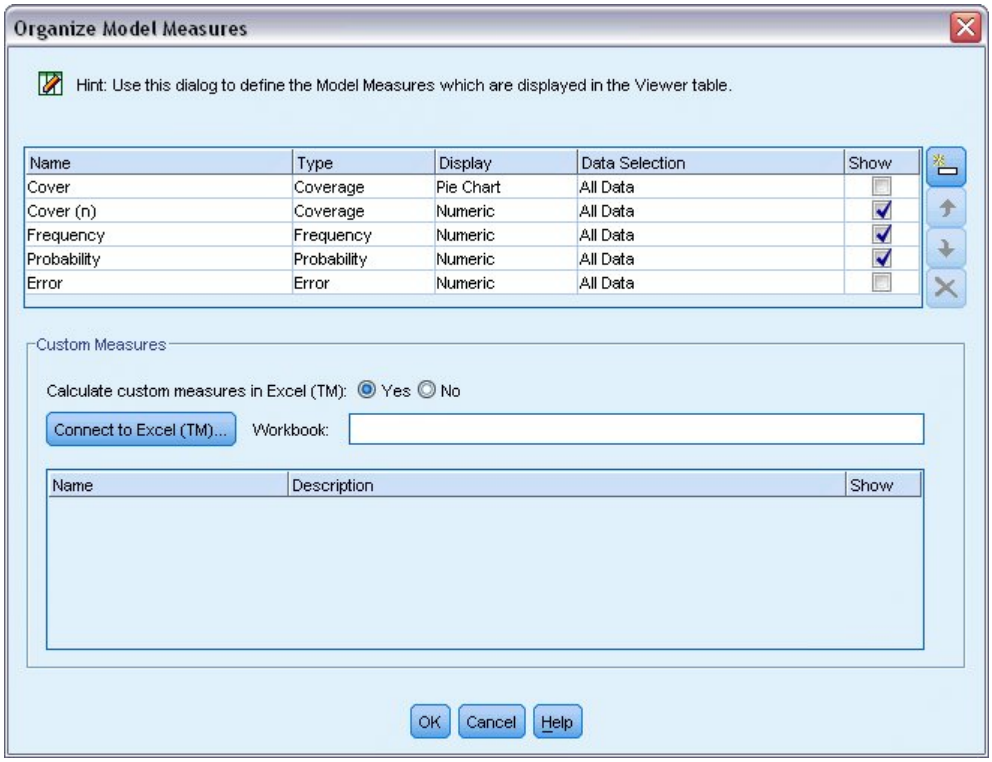
Obliczanie miar użytkownika za pomocą programu Excel

1. Aby uzyskać większy wgląd w to, jak model działa pod względem praktycznym, wybierz opcję **Organizuj miary modelu** z menu Narzędzia.



Rysunek 134. Organizacja miar modelu

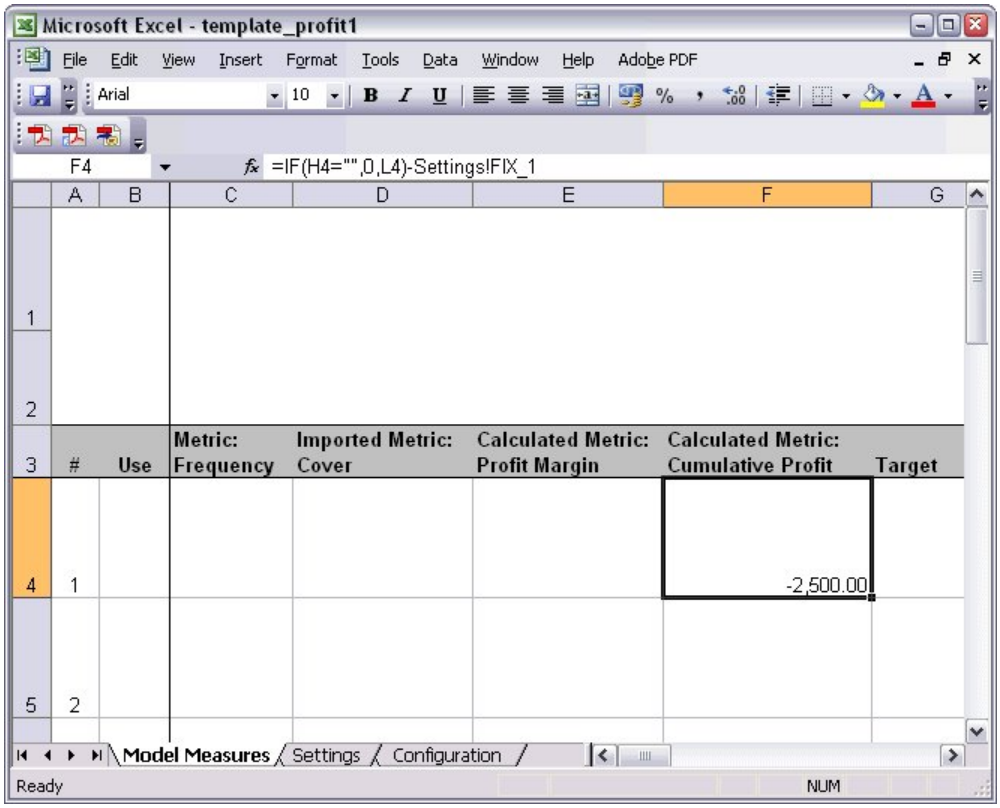
Okno dialogowe Organizuj miary modelu pozwala na wybranie miar (lub kolumn) wyświetlanych w przeglądarce Lista interaktywna. Można również określić, czy miary są obliczane dla wszystkich rekordów, czy dla wybranego podzbioru, i można zadecydować, aby w stosownych przypadkach wyświetlany był wykres kołowy zamiast liczby.



Rysunek 135. Okno dialogowe Organizuj miary modelu

Dodatkowo, jeśli zainstalowana jest aplikacja Microsoft Excel, można utworzyć łącze do szablonu Excel, który obliczy miary użytkownika i doda je do prezentacji interaktywnej.

2. W oknie dialogowym Organizuj miary modelu ustaw opcję **Oblicz niestandardowe miary w programie Excel (TM)** na **Tak**.
3. Kliknij przycisk **Połącz z programem Excel (TM)**
4. Wybierz arkusz kalkulacyjny *template_profit.xlt* znajdujący się w folderze *streams* w folderze *Demos* instalacji programu IBM SPSS Modeler i kliknij przycisk **Otwórz**, aby uruchomić arkusz.



Rysunek 136. Arkusz kalkulacyjny miar modelu programu Excel

Szablon programu Excel zawiera trzy arkusze kalkulacyjne:

- **Model Measures** wyświetla miary modelu importowane z modelu i oblicza miary użytkownika do eksportowania z powrotem do modelu.
- **Settings** zawiera parametry używane w obliczeniach miar użytkownika.
- **Configuration** definiuje miary importowane z modelu i eksportowane do modelu.

Metryki eksportowane z powrotem do modelu to:

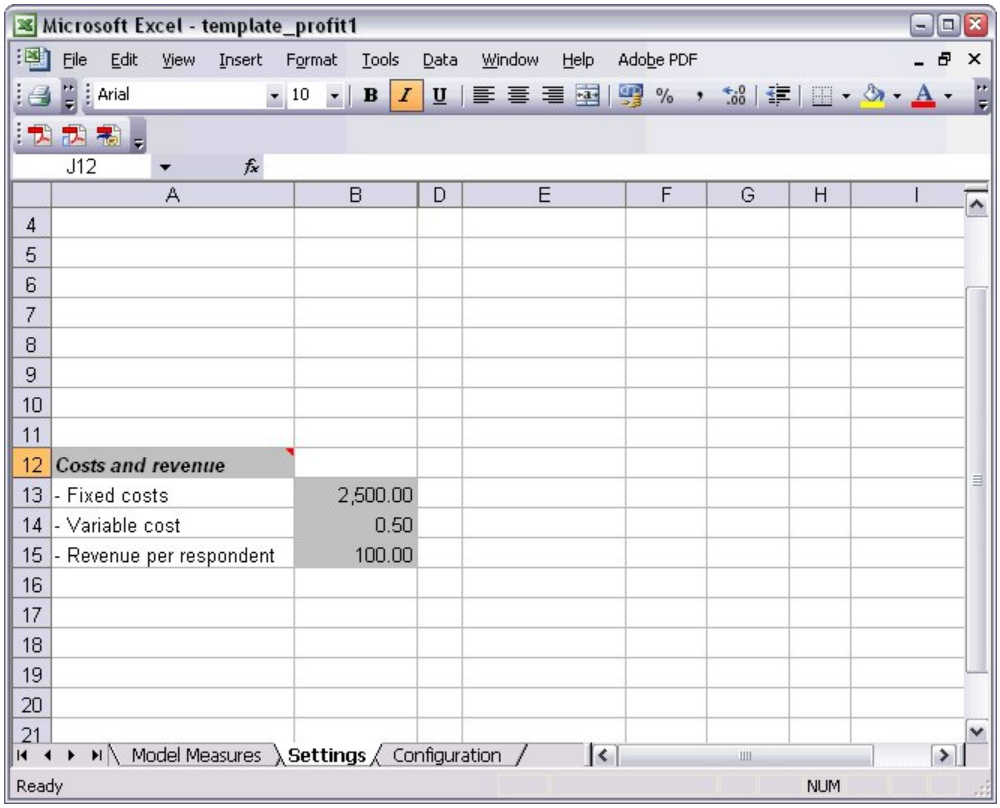
- **Profit Margin.** Przychód netto z segmentu
- **Cumulative Profit.** Całkowity zysk z kampanii

Zgodnie z definicją w następujących formułach:

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

Należy zauważyć, że wartości Frequency i Cover są importowane z modelu.

Parametry kosztu i przychodów są określane przez użytkownika na arkuszu Settings.



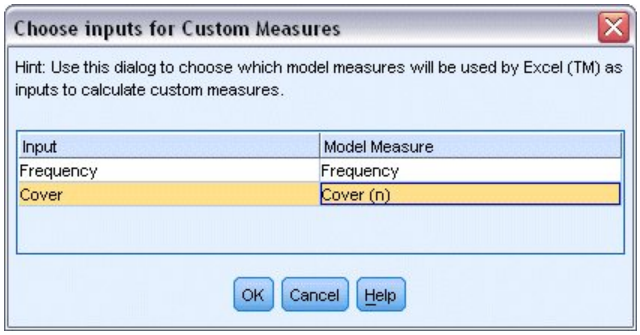
Rysunek 137. Arkusz kalkulacyjny Settings programu Excel

Fixed cost to koszt konfiguracji kampanii, taki jak projektowanie i planowanie.

Variable cost to koszt przekazania oferty każdemu klientowi, taki jak koperty i znaczki.

Revenue per respondent to przychód netto z każdego klienta, który odpowie na ofertę.

5. Aby zakończyć łącze do modelu, użyj paska zadań systemu Windows (lub naciśnij klawisze Alt+Tab), aby przejść z powrotem do przeglądarki Lista interaktywna.



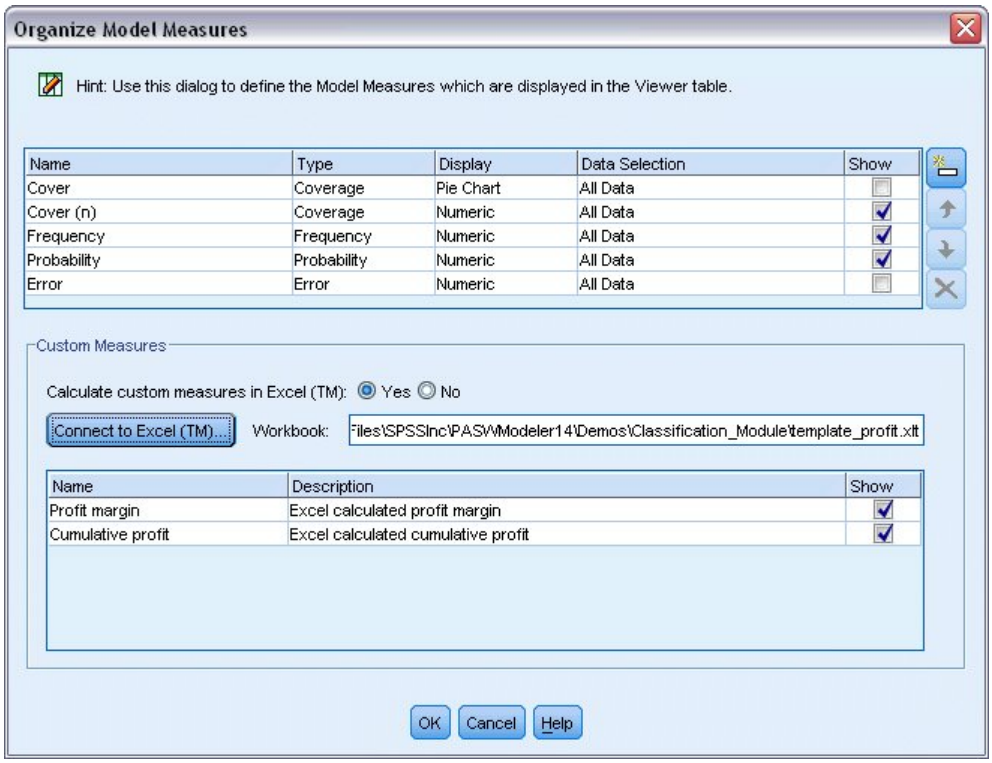
Rysunek 138. Wybieranie danych wejściowych dla miar użytkownika

Wyświetlone zostaje okno dialogowe Choose Inputs for Custom Measures, które pozwala na odwzorowanie danych wejściowych z modelu na konkretnych parametrach zdefiniowanych w szablonie. W lewej kolumnie przedstawiono listę dostępnych miar, a prawa kolumna odwzorowuje je na parametrach arkusza kalkulacyjnego zdefiniowanych w arkuszu Configuration.

6. W kolumnie **Model Measures** wybierz pozycję **Frequency** i **Cover (n)** dla odpowiednich danych wejściowych i kliknij przycisk **OK**.

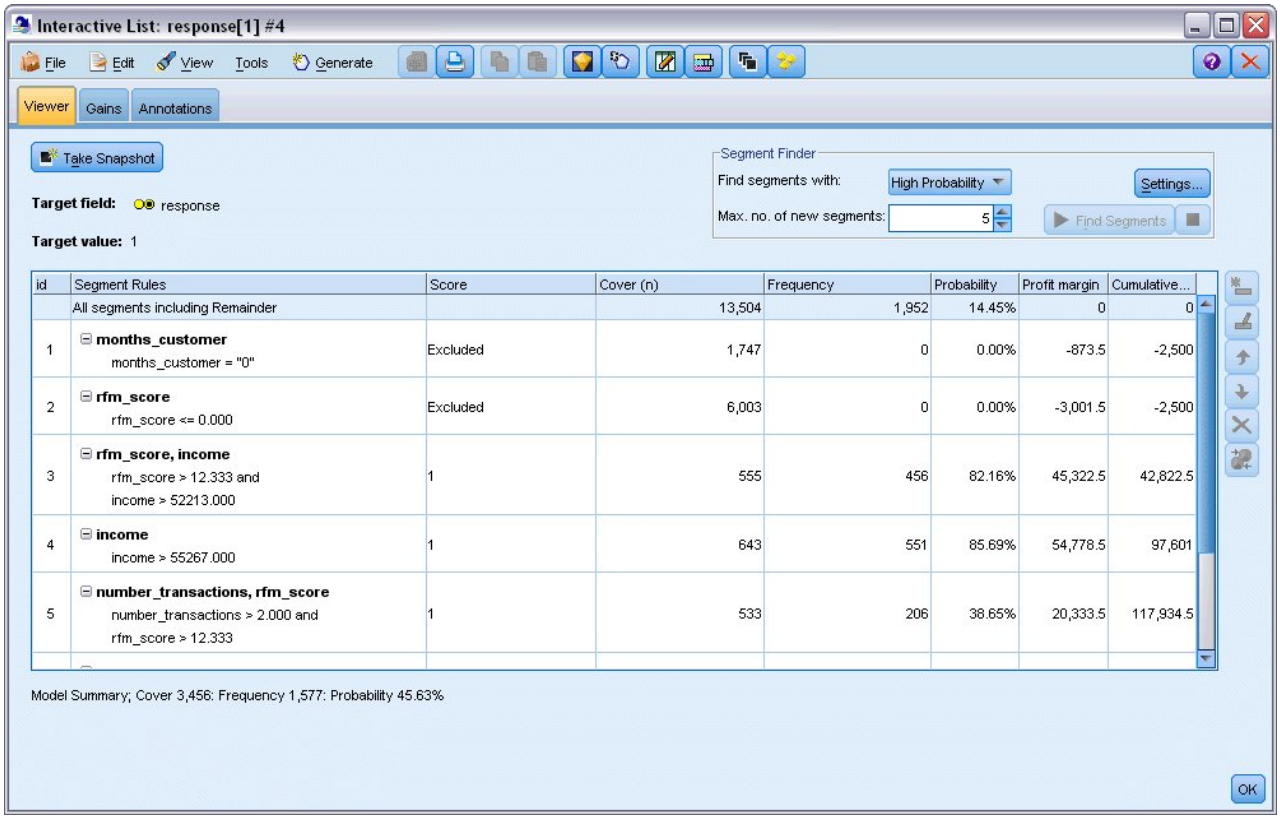
W tym przypadku nazwy parametrów w szablonie — Frequency i Cover (n) — odpowiadają danym wejściowym, ale można również użyć innych nazw.

7. Kliknij przycisk **OK** w oknie dialogowym Organizuj miary modelu, aby aktualizować przeglądarkę Lista interaktywna.



Rysunek 139. Okno dialogowe Organizuj miary modelu przedstawiające miary użytkownika z programu Excel

Nowe miary są teraz dodawane jako nowe kolumny w oknie, ale będą przeliczane za każdym razem, gdy model jest aktualizowany.



Rysunek 140. Miary użytkownika z programu Excel wyświetlane w przeglądarce Lista interaktywna

Edytując szablon programu Excel, można utworzyć dowolną liczbę miar.

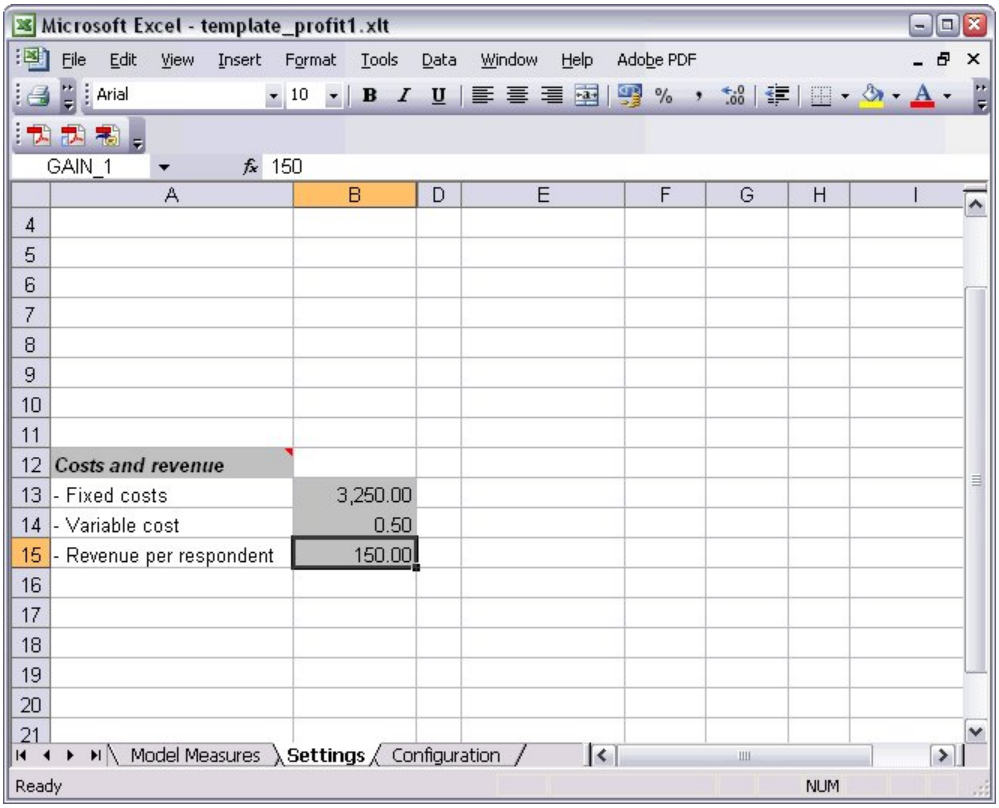
Modyfikowanie szablonu programu Excel

Mimo że do programu IBM SPSS Modeler dodawany jest domyślny szablon programu Excel do użycia z przeglądarką Lista interaktywna, można zmienić ustawienia lub dodać własne. Na przykład koszty w szablonie mogą być niepoprawne dla organizacji użytkownika lub wymagają poprawienia.

Uwaga: Jeśli istniejący szablon zostanie zmodyfikowany lub użytkownik utworzy własny, należy zapisać plik z rozszerzeniem programu Excel 2003 *.xlt*.

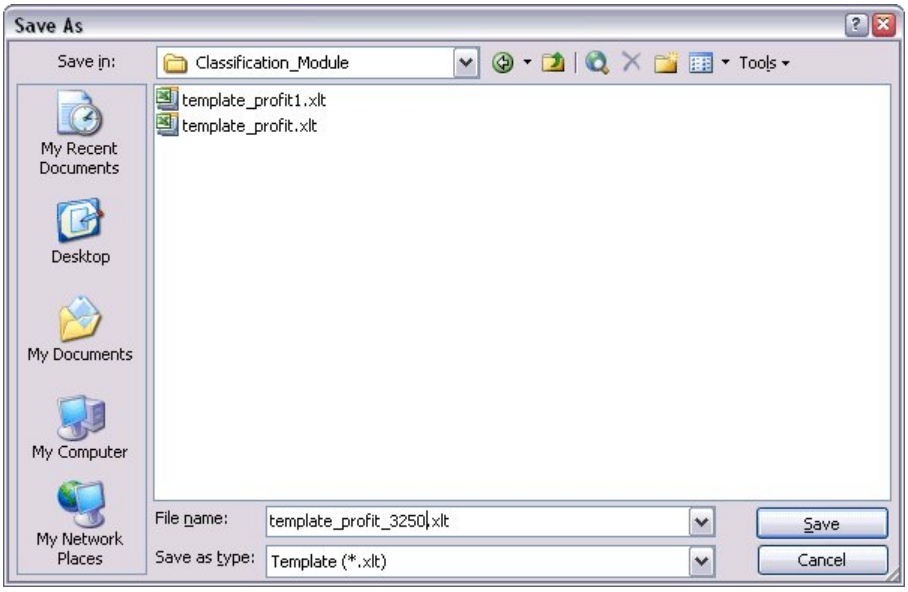
Aby zmodyfikować domyślny szablon nowymi szczegółami dotyczącymi kosztów i przychodów oraz aktualizować nowe wartości w przeglądarce Lista interaktywna:

1. W przeglądarce Lista interaktywna z menu Narzędzia wybierz polecenie **Organizuj miary modelu**.
2. W oknie dialogowym Organizuj miary modelu kliknij przycisk **Połącz z programem Excel™.**
3. Wybierz arkusz *template_profit.xlt* i kliknij przycisk **Otwórz**, aby uruchomić arkusz kalkulacyjny.
4. Wybierz arkusz Settings.
5. Zmień wartość **Fixed costs** na 3250,00, i **Revenue per respondent** na 150,00.



Rysunek 141. Zmodyfikowane wartości na arkuszu kalkulacyjnym Settings programu Excel

6. Zapisz zmodyfikowany szablon z unikalną, powiązaną nazwą pliku. Upewnij się, że plik ma rozszerzenie programu Excel 2003 *.xlt*.



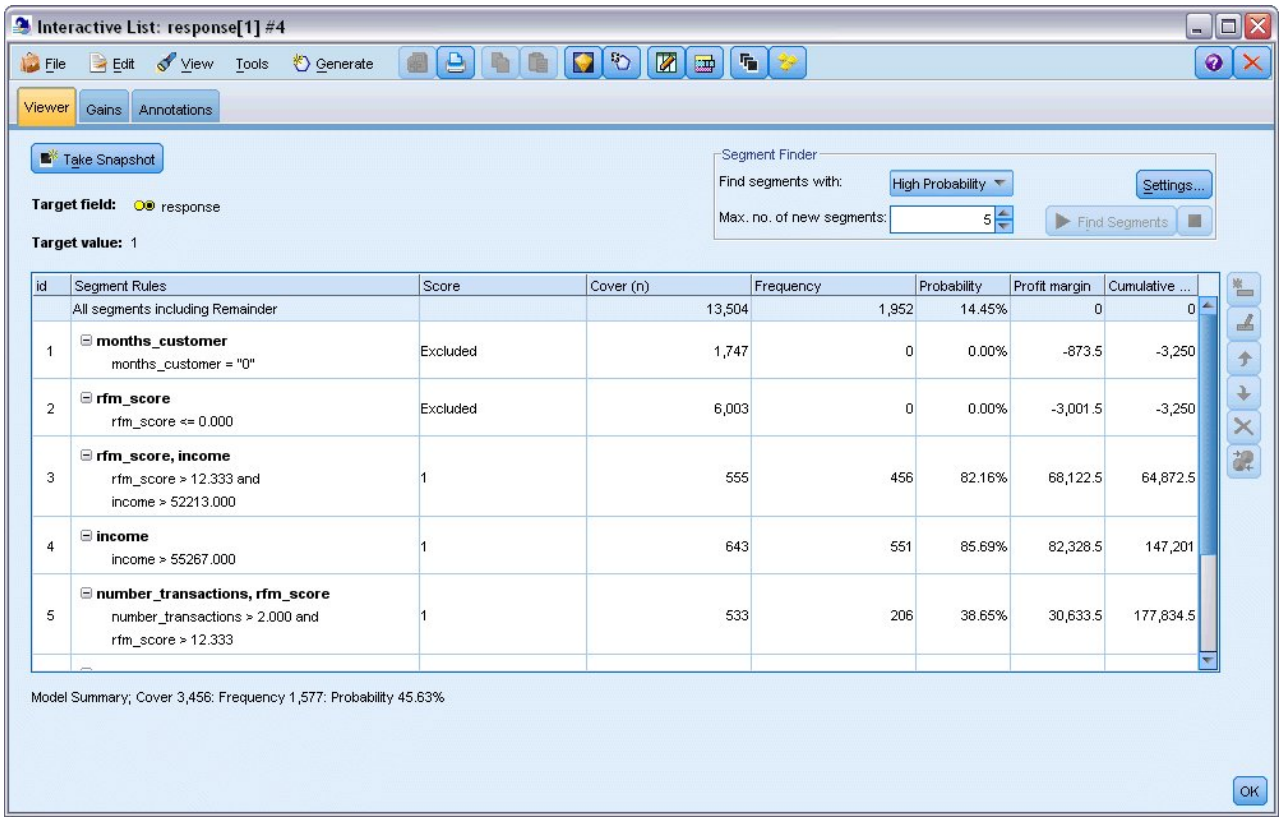
Rysunek 142. Zapisywanie zmodyfikowanego szablonu programu Excel

7. Użyj paska zadań systemu Windows (lub naciśnij klawisze Alt+Tab), aby przejść z powrotem do przeglądarki Lista interaktywna.

W oknie dialogowym Choose Inputs for Custom Measures wybierz miary, które chcesz wyświetlić, i kliknij przycisk **OK**.

8. W oknie dialogowym Organizuj miary modelu kliknij przycisk **OK**, aby aktualizować przeglądarkę Lista interaktywna.

Oczywiście w tym przykładzie przedstawiono tylko jeden prosty sposób modyfikowania szablonu programu Excel. Można wprowadzić dalsze zmiany, które pobierają dane lub przekazują dane do przeglądarki Lista interaktywna lub działają w programie Excel, generując inne dane wyjściowe, takie jak np. wykresy.



id	Segment Rules	Score	Cover (n)	Frequency	Probability	Profit margin	Cumulative ...
	All segments including Remainder		13,504	1,952	14.45%	0	0
1	**months_customer** months_customer = "0"	Excluded	1,747	0	0.00%	-873.5	-3,250
2	**rfm_score** rfm_score <= 0.000	Excluded	6,003	0	0.00%	-3,001.5	-3,250
3	**rfm_score, income** rfm_score > 12.333 and income > 52213.000	1	555	456	82.16%	68,122.5	64,872.5
4	**income** income > 55267.000	1	643	551	85.69%	82,328.5	147,201
5	**number_transactions, rfm_score** number_transactions > 2.000 and rfm_score > 12.333	1	533	206	38.65%	30,633.5	177,834.5

Rysunek 143. Zmodyfikowane miary użytkownika z programu Excel wyświetlane w przeglądarce Lista interaktywna

Zapisywanie wyników

Aby zapisać model do późniejszego użycia podczas sesji interaktywnej, można wykonać obraz stanu modelu, który będzie wymieniony na karcie Obrazy stanu. Można powrócić do dowolnego zapisanego obrazu stanu w dowolnym momencie podczas sesji interaktywnej.

Kontynuując pracę w ten sposób, można eksperymentować z dodatkowymi zadaniami eksploracji, aby wyszukać dodatkowe segmenty. Można również edytować istniejące segmenty, wstawiać segmenty użytkownika oparte na własnych regułach biznesowych, tworzyć wybory danych, aby zoptymalizować model dla konkretnych grup i dostosować model na wiele innych sposobów. Można też wyraźnie uwzględnić lub wyłączyć każdy segment, aby określić sposób, w jaki będzie oceniany.

Kiedy użytkownik jest zadowolony z wyników, może użyć menu Utwórz, aby wygenerować model, który można dodać do strumieni lub wdrożyć do celów oceniania.

Aby zapisać bieżący stan sesji interaktywnej na kolejny dzień, można też wybrać opcję **Aktualizacja węzła modelowania** z menu Plik. Spowoduje to zaktualizowanie węzła modelowania Lista decyzyjna z uwzględnieniem bieżących ustawień, w tym zadań eksploracji, obrazów stanu modelu, wyborów danych i miar użytkownika. Przy

następnym uruchomieniu strumienia należy upewnić się, że w węźle modelowania Lista decyzyjna zaznaczona jest opcja **użycia informacji o zapisanej sesji interaktywnej**, aby przywrócić sesję do bieżącego stanu.

Rozdział 12. Klasyfikowanie klientów usług telekomunikacyjnych (Wielomianowa regresja logistyczna)

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

Na przykład załóżmy, że operator telekomunikacyjny pogrupował bazę klientów wg wzorców korzystania z usług, tworząc cztery kategorie. Jeśli można użyć danych demograficznych do przewidywania członkostwa w grupie, można dostosować oferty dla indywidualnych potencjalnych klientów.

W tym przykładzie zastosowano strumień o nazwie *telco_custcat.str*, który odwołuje się do pliku danych o nazwie *telco.sav*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *telco_custcat.str* znajduje się w katalogu *streams*.

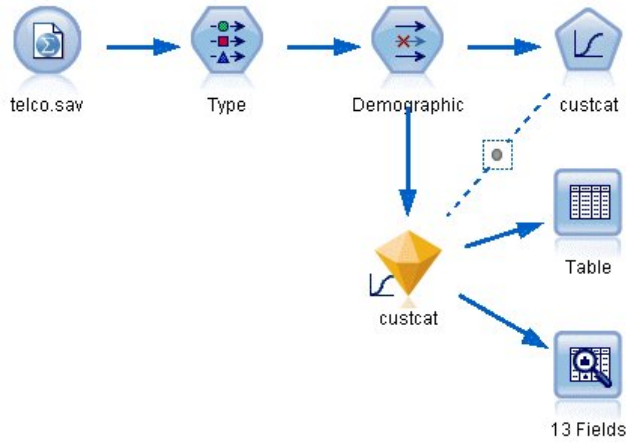
Ten przykład koncentruje się na użyciu danych demograficznych do przewidywania wzorców używania. Zmienna przewidywana *custcat* ma cztery możliwe wartości, które odpowiadają czterem grupom klientów w następujący sposób:

Wartość	Etykieta
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

Ponieważ zmienna przewidywana ma wiele kategorii, używany jest model wielomianowy. W przypadku, gdy zmienna przewidywana ma dwie różne kategorie, takie jak tak/nie, prawda/fałsz lub odejście/brak odejścia, można zamiast tego utworzyć model dwumianowy. Więcej informacji można znaleźć w temacie Rozdział 13, “Poziom odejścia usług telekomunikacyjnych (Dwumianowa regresja logistyczna)”, na stronie 139.

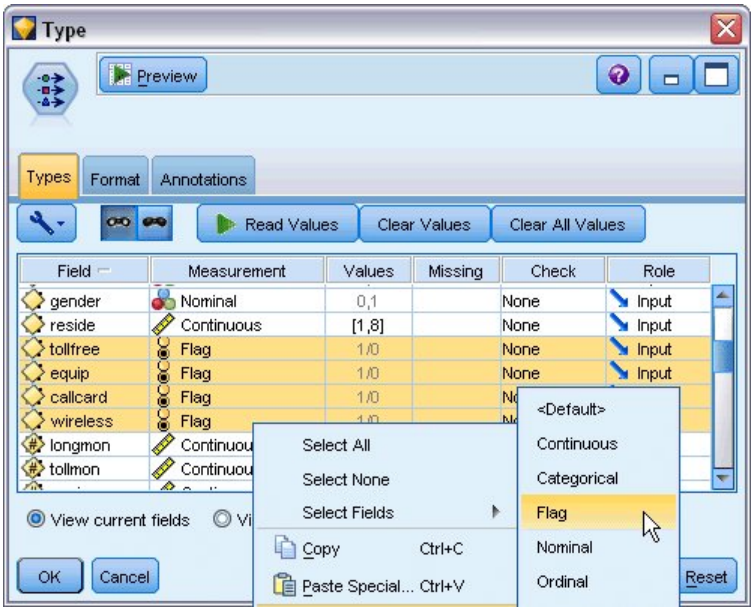
Tworzenie strumienia

1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *telco.sav* znajdujący się w folderze *Demos*.



Rysunek 144. Przykładowy strumień klasyfikujący klientów za pomocą wielomianowej regresji logistycznej

a. Dodaj węzeł typu i kliknij przycisk **Odczytaj wartości**, upewniając się, że ustawiono poprawnie wszystkie poziomy pomiaru. Na przykład większość zmiennych z wartościami 0 i 1 można traktować jako flagi.

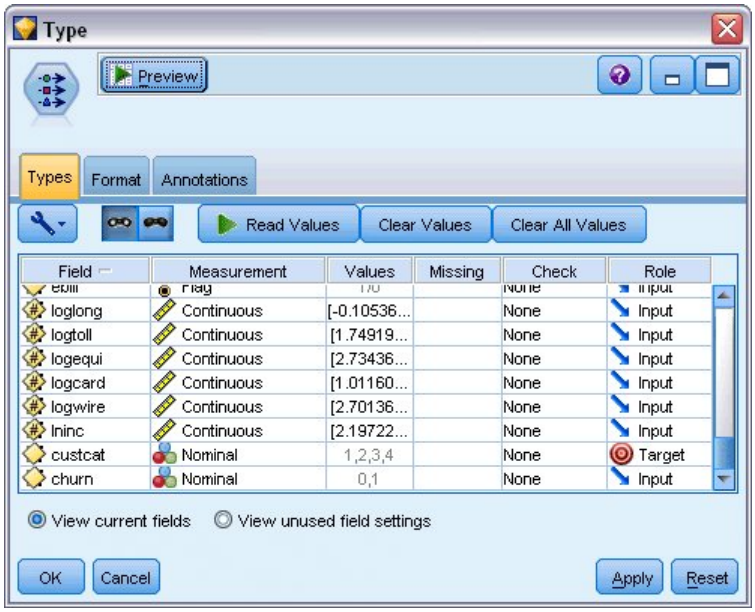


Rysunek 145. Ustawianie poziomu pomiaru dla wielu zmiennych

Wskazówka: Aby zmienić właściwości dla wielu zmiennych z podobnymi wartościami (takimi jak 0/1), kliknij nagłówek kolumny *Wartości*, aby posortować zmienne według wartości, a następnie przytrzymaj naciśnięty klawisz Shift, używając myszy lub klawiszy strzałek, aby wybrać wszystkie zmienne, które chcesz zmienić. Możesz następnie kliknąć wybrany zakres prawym klawiszem myszy, aby zmienić poziom pomiaru lub inne atrybuty dla wybranych zmiennych.

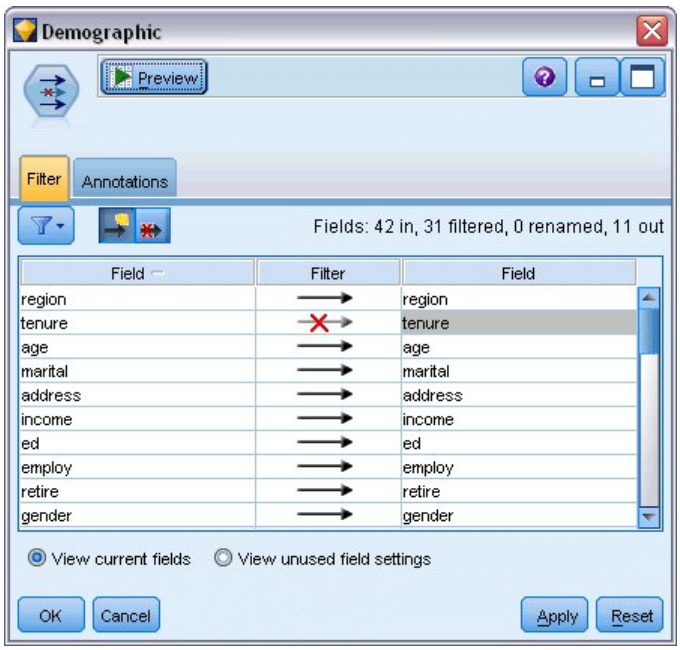
Należy zauważyć, że zmienna *gender* (płeć) bardziej prawidłowo jest uważana za zmienną z dwiema wartościami niż za flagę, więc należy pozostawić wartość kolumny Poziom pomiaru jako **Nominalna**.

b. Ustaw rolę zmiennej *custcat* na **Przewidywana**. Wszystkie pozostałe zmienne powinny mieć ustawioną rolę na **Dane wejściowe**.



Rysunek 146. Ustawianie roli zmiennej

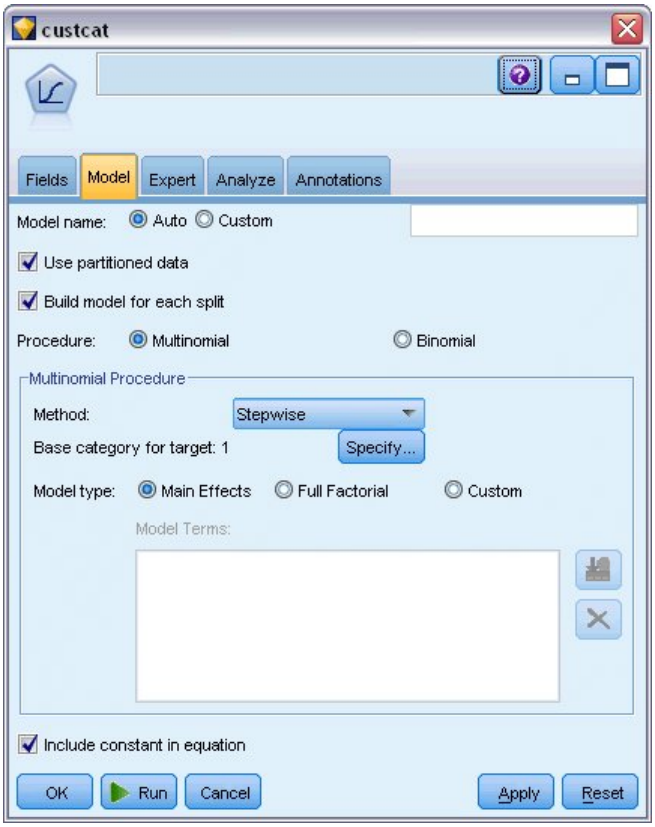
Ponieważ ten przykład koncentruje się na demografii, użyj węzła filtrowania do uwzględnienia tylko istotnych zmiennych (*region*, *age*, *marital*, *address*, *income*, *ed*, *employ*, *retire*, *gender*, *reside* i *custcat*). Inne zmienne można wyłączyć na potrzeby tej analizy.



Rysunek 147. Filtrowanie zmiennych demograficznych

(Można również zmienić rolę tych zmiennych na **Brak**, zamiast je wykluczać lub wybrać zmienne, których chcesz użyć w węźle modelowania).

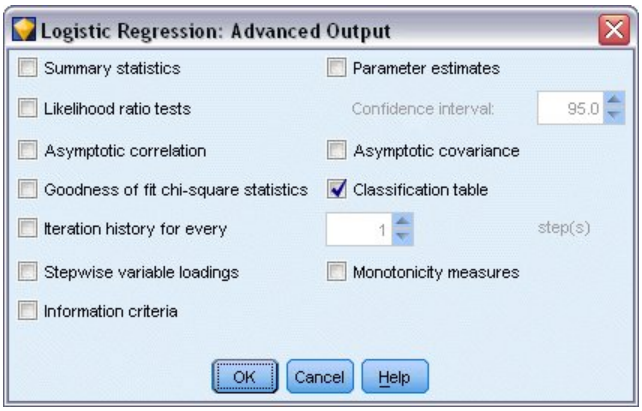
2. W węźle logistycznym kliknij kartę **Model** i wybierz metodę **Krokowa**. Zaznacz również opcje **Wielomianowa**, **Efekty główne** i **Uwzględnij stałą w równaniu**.



Rysunek 148. Wybieranie opcji modelu

Pozostaw wartość 1 dla parametru Kategoria odniesienia dla przewidywanej. Model porówna innych klientów z tymi, którzy subskrybują usługę Basic Service.

3. Na karcie Zaawansowany wybierz tryb **Zaawansowany**, wybierz **Wynik** i w oknie dialogowym Zaawansowane dane wyjściowe wybierz opcję **Tabela klasyfikacji**.

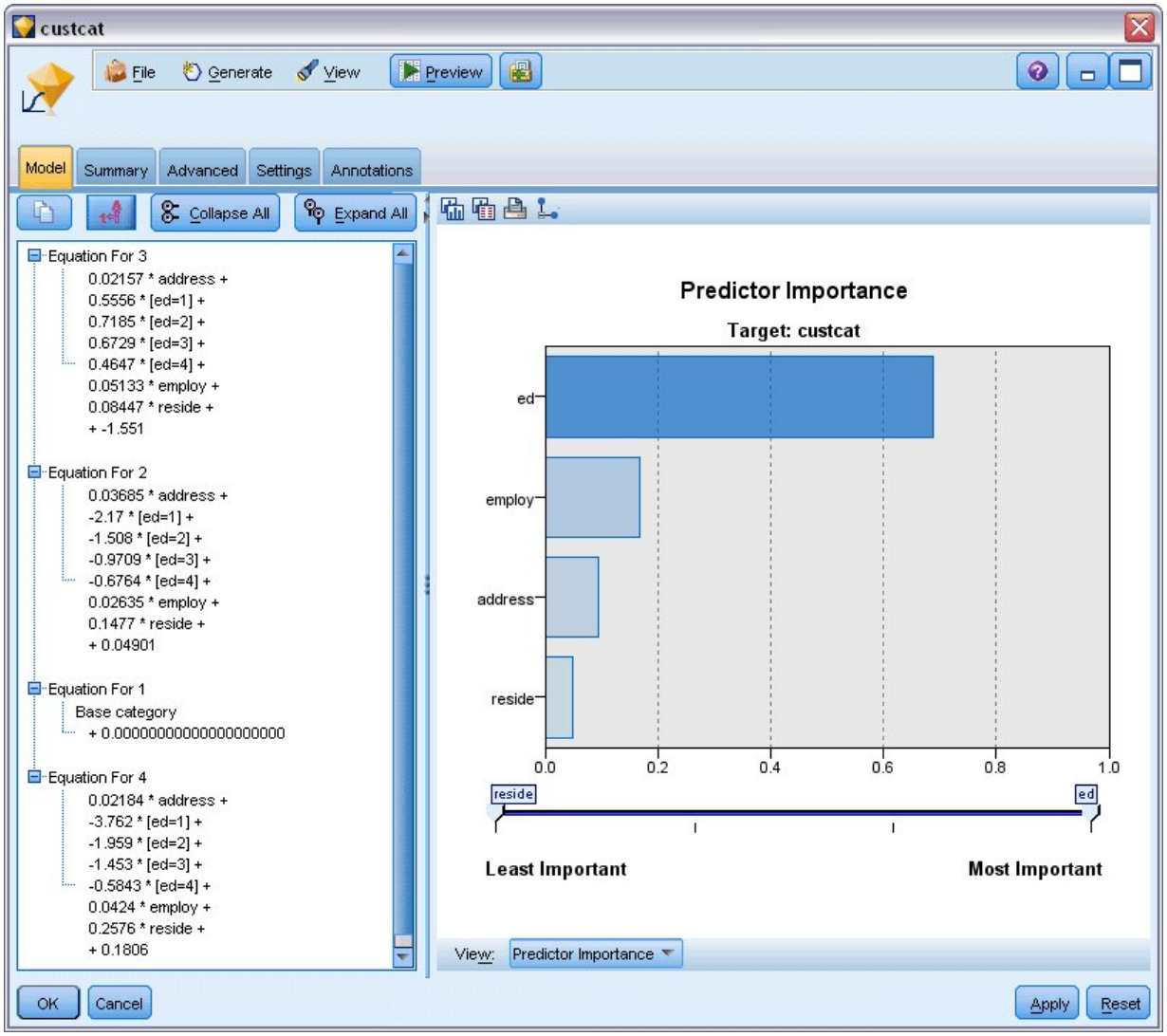


Rysunek 149. Wybieranie opcji wyników

Przeglądanie modelu

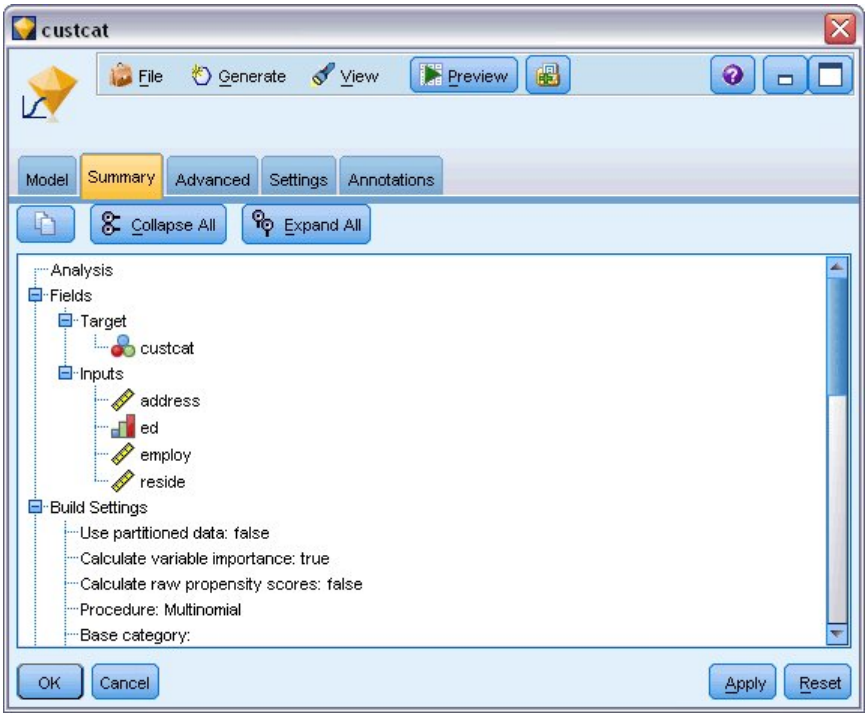
1. Wykonaj węzeł, aby wygenerować model, który zostanie dodany do palety modeli w prawym górnym rogu. Aby przeglądać szczegóły modelu, kliknij wygenerowany węzeł modelu prawym przyciskiem myszy i wybierz opcję **Przeglądaj**.

Karta modelu wyświetla równania używane do przypisywania rekordów do każdej kategorii zmiennej przewidywanej. Istnieją cztery możliwe kategorie, z których jedna jest kategorią odniesienia, dla której nie ma wyświetlonych szczegółów równania. Szczegóły są wyświetlane dla pozostałych trzech równań, gdzie kategoria 3 reprezentuje usługę Plus Service itd.



Rysunek 150. Przeglądanie wyników modelu

Karta Podsumowanie przedstawia (między innymi) zmienną przewidywaną i dane wejściowe (zmienne predykcyjne) używane przez model. Należy pamiętać, że są to zmienne, które zostały wybrane na podstawie metody krokowej, a nie pełna lista przekazana do analizy.

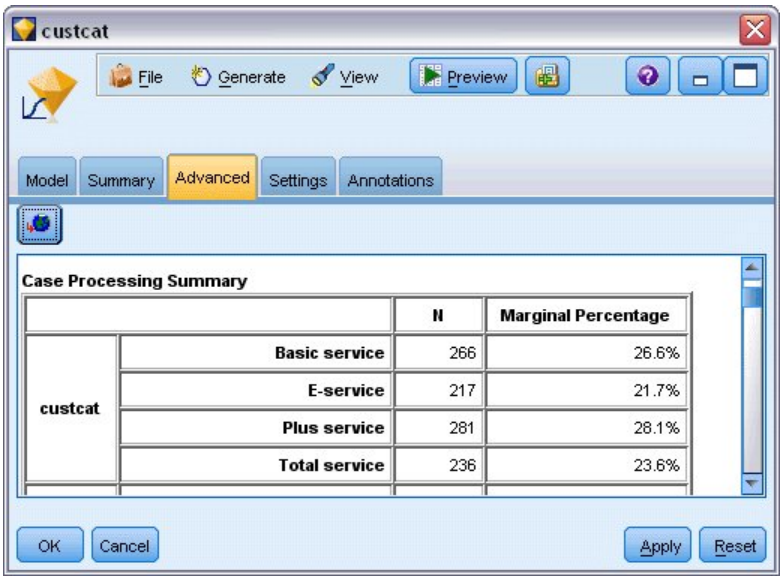


Rysunek 151. Podsumowanie modelu przedstawiające zmienną przewidywaną i zmienne wejściowe

Pozycje wyświetlane na karcie Zaawansowane zależą od opcji wybranych w oknie dialogowym Zaawansowane dane wyjściowe w węźle modelowania.

Jedna pozycja, która jest zawsze wyświetlana, to Informacja o analizowanych danych przedstawiająca udział procentowy rekordów przypadających w każdej kategorii zmiennej przewidywanej. Pozwala to użyć modelu zerowego jako podstawy porównania.

Bez budowania modelu używającego predyktorów, najlepszym rozwiązaniem byłoby przypisanie wszystkich klientów do najbardziej powszechnej grupy, czyli dla usługi Plus service.

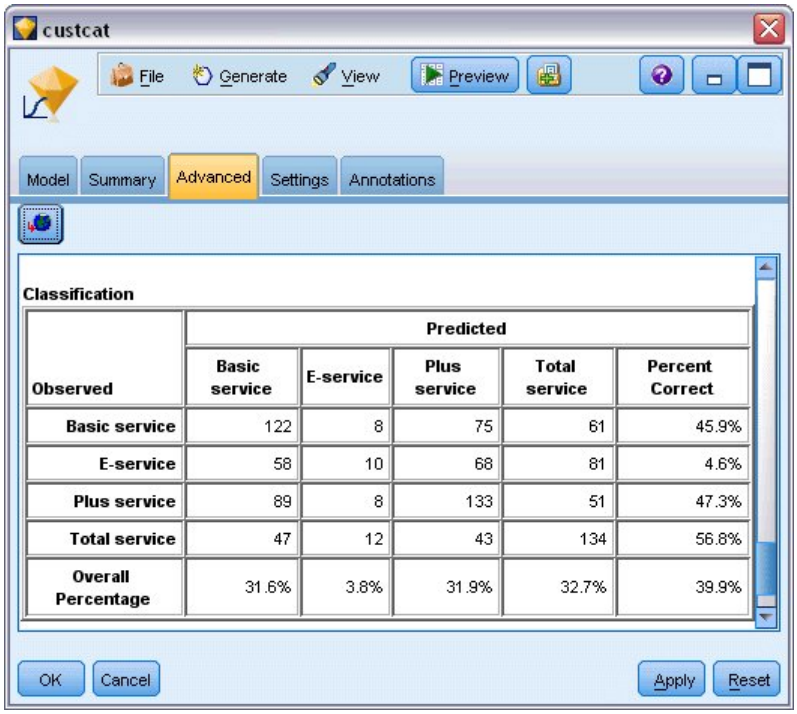


Rysunek 152. Informacja o analizowanych danych

Na podstawie danych uczących, jeśli przypisano by wszystkich klientów do modelu zerowego, takie rozwiązanie byłoby poprawne w 281/1000 = 28,1% przypadków. Karta Zaawansowany zawiera dalsze informacje, które pozwalają na zbadanie predykcji modelu. Następnie można porównać predykcje z wynikami modelu zerowego, aby jak dobrze model działa dla konkretnych danych.

W dolnej części karty Zaawansowany tabela Klasyfikacja przedstawia wyniki dla modelu, które są poprawne w 39,9% przypadków.

Model bardzo dobrze identyfikuje klientów usługi Total Service (kategoria 4), ale bardzo słabo identyfikuje klientów usługi E-service (kategoria 2). Jeśli potrzebna jest większa dokładność dla klientów w kategorii 2, konieczne może być znalezienie innego predyktora do ich identyfikacji.



Rysunek 153. Tabela klasyfikacji

W zależności od tego, co ma być przewidywane, model może być idealnie dopasowany do potrzeb użytkownika. Na przykład, jeśli użytkownikowi nie zależy na identyfikacji klientów w kategorii 2, model może być wystarczająco dokładny. Może być tak w przypadku, gdy usługa E-service ma największe straty i daje mały zysk.

Jeśli najwyższy zwrot z inwestycji pochodzi od klientów z kategorii 3 lub 4, model może zapewnić wymagane informacje.

Aby ocenić, jak dobrze model jest dopasowany do danych, podczas budowania modelu w oknie dialogowym Zaawansowane dane wyjściowe dostępne są narzędzia diagnostyczne. Wyjaśnienia podstaw matematycznych metod modelowania używanych w programie IBM SPSS Modeler przedstawiono w publikacji *IBM SPSS Modeler Algorithms Guide* dostępnej w katalogu *\Documentation* na dysku instalacyjnym.

Należy zauważyć, że wyniki opierają się tylko na danych uczących. Aby ocenić, jak dobrze model uogólnia inne dane w rzeczywistych przypadkach, należałoby użyć węzła podziału na podzbiory, aby zatrzymać podzbiór rekordów do celów testowania i walidacji.

Rozdział 13. Poziom odejścia usług telekomunikacyjnych (Dwumianowa regresja logistyczna)

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

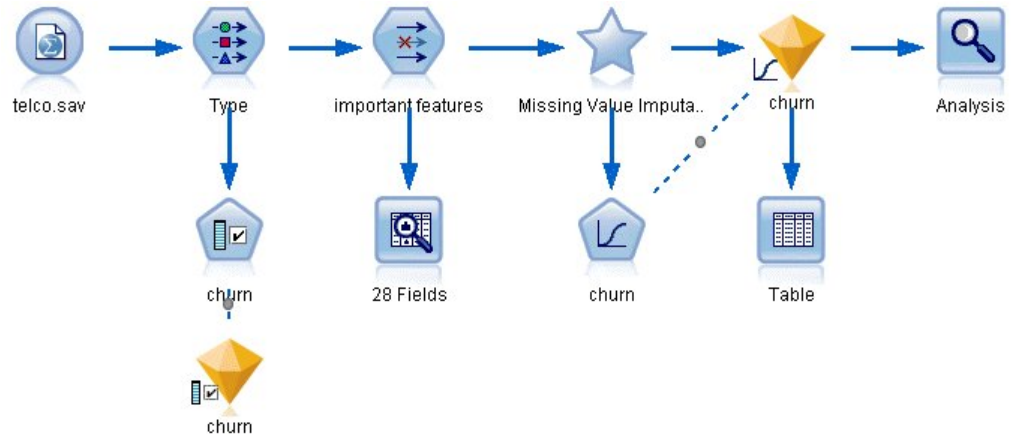
W tym przykładzie zastosowano strumień o nazwie *telco_churn.str*, który odwołuje się do pliku danych o nazwie *telco.sav*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *telco_churn.str* znajduje się w katalogu *streams*.

Załóżmy na przykład, że operator telekomunikacyjny jest zaniepokojony liczbą klientów odchodzących do konkurencji. Jeśli danych dotyczących wykorzystania usług można użyć do przewidzenia, którzy klienci są skłonni przenieść się do innego operatora, można przedstawiać tym klientom bardziej zindywidualizowaną ofertę w celu zatrzymania możliwie największej ich liczby.

Ten przykład koncentruje się na wykorzystaniu danych używania, aby przewidzieć utratę klientów (odchodzenie). Ponieważ zmienna przewidywana ma dwie kategorie, używany jest model dwumianowy. W przypadku zmiennej przewidywanej z wieloma kategoriami, można zamiast tego utworzyć model wielomianowy. Więcej informacji można znaleźć w temacie Rozdział 12, “Klasyfikowanie klientów usług telekomunikacyjnych (Wielomianowa regresja logistyczna)”, na stronie 131.

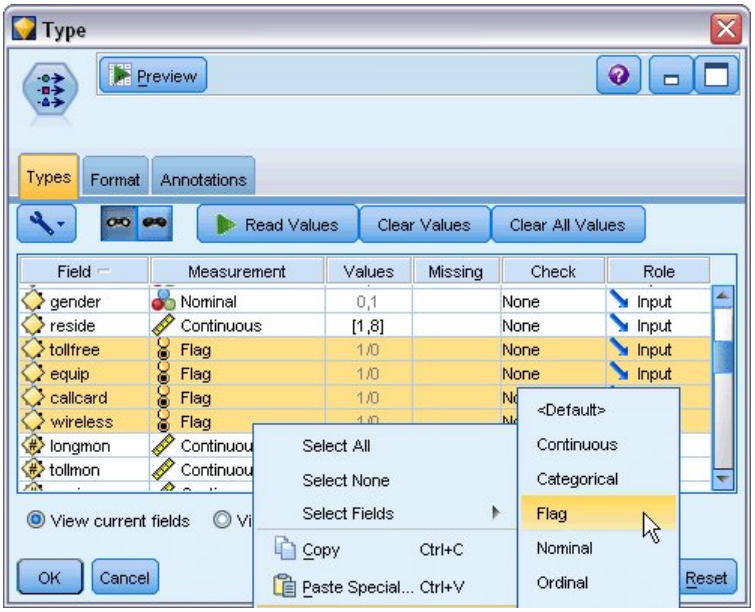
Tworzenie strumienia

1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *telco.sav* znajdujący się w folderze *Demos*.



Rysunek 154. Przykładowy strumień klasyfikujący klientów za pomocą dwumianowej regresji logistycznej

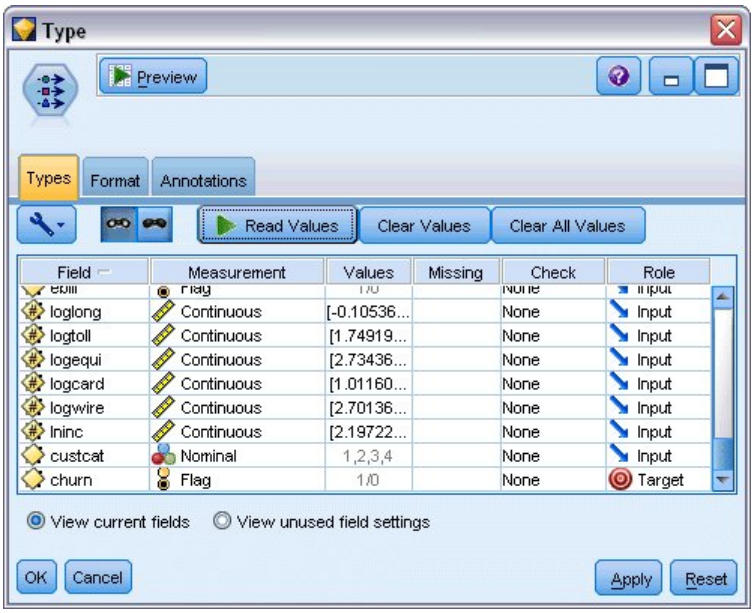
2. Dodaj węzeł typu, aby zdefiniować zmienne, upewniając się, że ustawiono poprawnie wszystkie poziomy pomiaru. Na przykład większość zmiennych z wartościami 0 i 1 można traktować jako flagi, ale niektóre zmienne, takie jak np. płeć, lepiej są przedstawiane jako zmienna nominalna z dwiema wartościami.



Rysunek 155. Ustawianie poziomu pomiaru dla wielu zmiennych

Wskazówka: Aby zmienić właściwości dla wielu zmiennych z podobnymi wartościami (takimi jak 0/1), kliknij nagłówek kolumny *Wartości*, aby posortować zmienne według wartości, a następnie przytrzymaj naciśnięty klawisz Shift, używając myszy lub klawiszy strzałek, aby wybrać wszystkie zmienne, które chcesz zmienić. Możesz następnie kliknąć wybrany zakres prawym klawiszem myszy, aby zmienić poziom pomiaru lub inne atrybuty dla wybranych zmiennych.

3. Ustaw poziom pomiaru dla zmiennej *odchodzenie* na wartość **Flaga** i ustaw rolę na **Przewidywana**. Wszystkie pozostałe zmienne powinny mieć ustawioną rolę na **Dane wejściowe**.

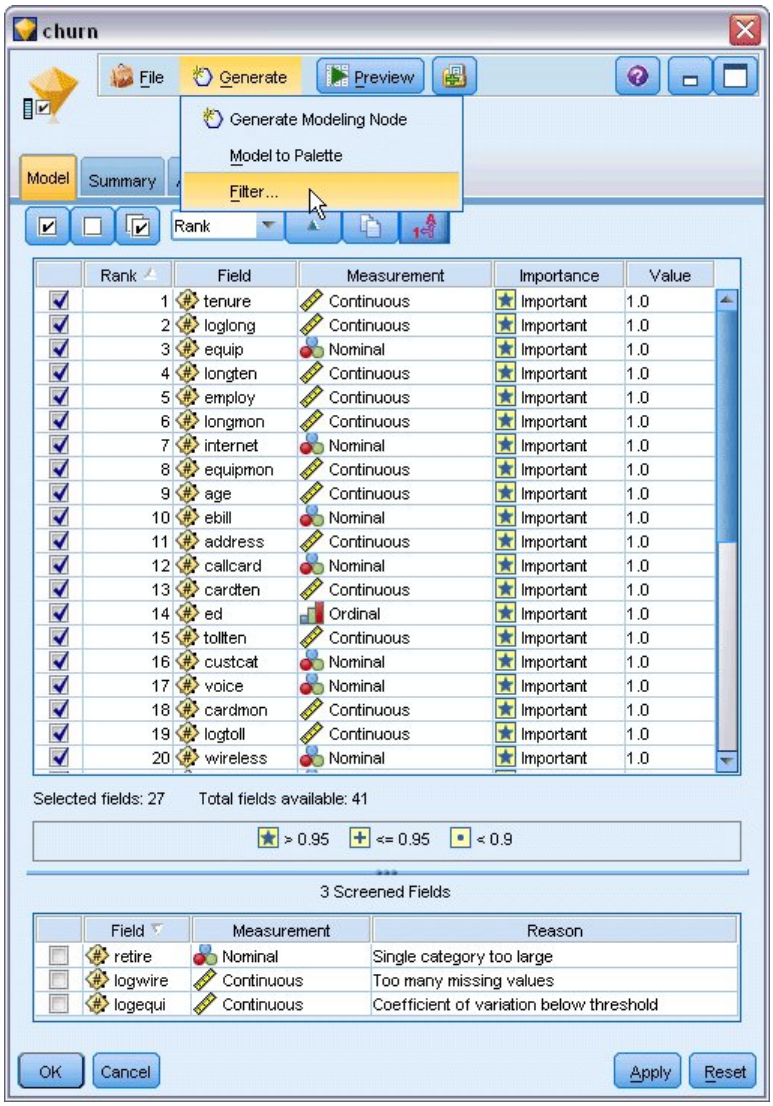


Rysunek 156. Ustawianie poziomu pomiaru i roli dla zmiennej odchodzenie

4. Do węzła typu dodaj węzeł modelowania Dobór predyktorów.

   Użycie węzła Dobór predyktorów pozwala na usunięcie predyktorów lub danych, które nie dodają żadnych przydatnych informacji pod względem relacji predyktor/zmienna przewidywana.

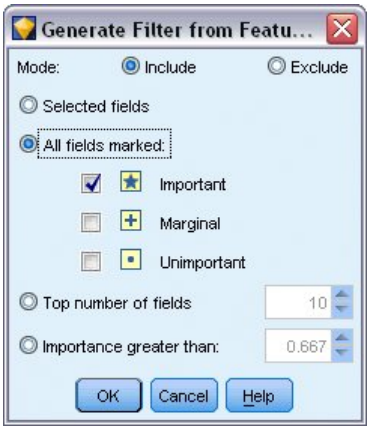
5. Uruchom strumień.
6. Otwórz wynikowy model użytkowy i w menu **Utwórz** wybierz pozycję **Filtrowanie**, aby utworzyć węzeł filtrowania.



Rysunek 157. Generowanie węzła filtrowania z węzła Dobór predyktorów

Nie wszystkie dane w pliku *telco.sav* będą przydatne w przewidywaniu odchodzenia. Można użyć filtra do wybrania tylko danych, które są uważane za istotne przy użyciu jako predyktor.

7. W oknie dialogowym Generuj filtr zaznacz opcję **Wszystkie zaznaczone zmienne: Ważne** i kliknij przycisk **OK**.
8. Dołącz wygenerowany węzeł filtrowania do węzła typu.

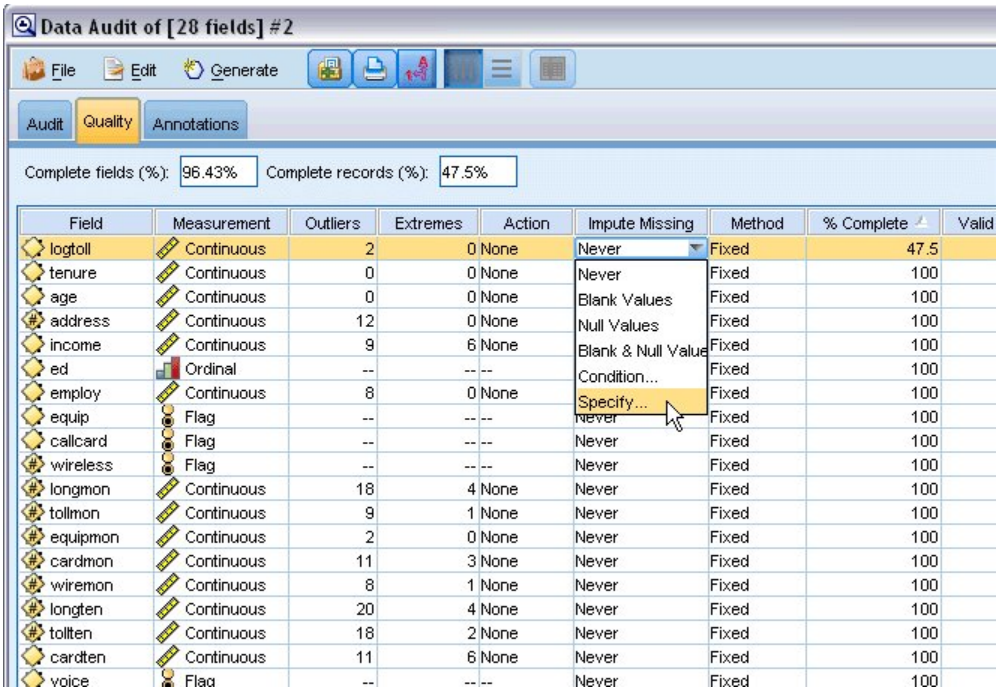


Rysunek 158. Wybieranie ważnych zmiennych

9. Załącz węzeł audytu danych do wygenerowanego węzła filtrowania.

   Otwórz węzeł audytu danych i kliknij przycisk **Uruchom**.

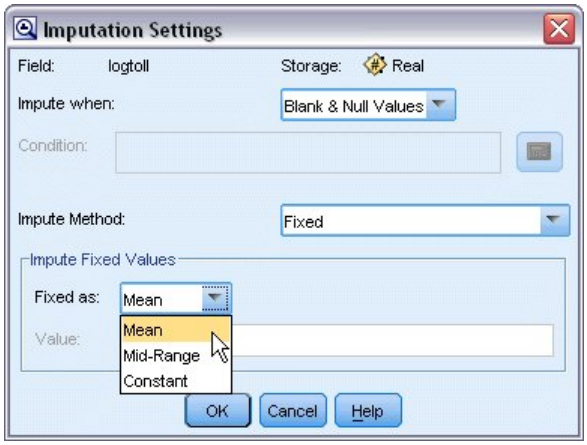
10. Na karcie Jakość przeglądarki audytu danych kliknij kolumnę *Ukończono %*, aby posortować kolumnę w rosnącej kolejności numerycznej. Pozwala to zidentyfikować zmienne z dużą ilością brakujących danych. W tym przypadku jedyną zmienną, którą należy zmienić jest *logtoll*, która jest kompletna w mniej niż 50%.
11. W kolumnie *Podstawianie braków* dla zmiennej *logtoll* kliknij opcję **Określ**.



Rysunek 159. Podstawianie brakujących wartości dla zmiennej logtoll

12. W pozycji **Podstaw gdy** wybierz opcję **Puste i wartości null**. W obszarze **Ustalone jako** wybierz pozycję **Średnia** i kliknij przycisk **OK**.

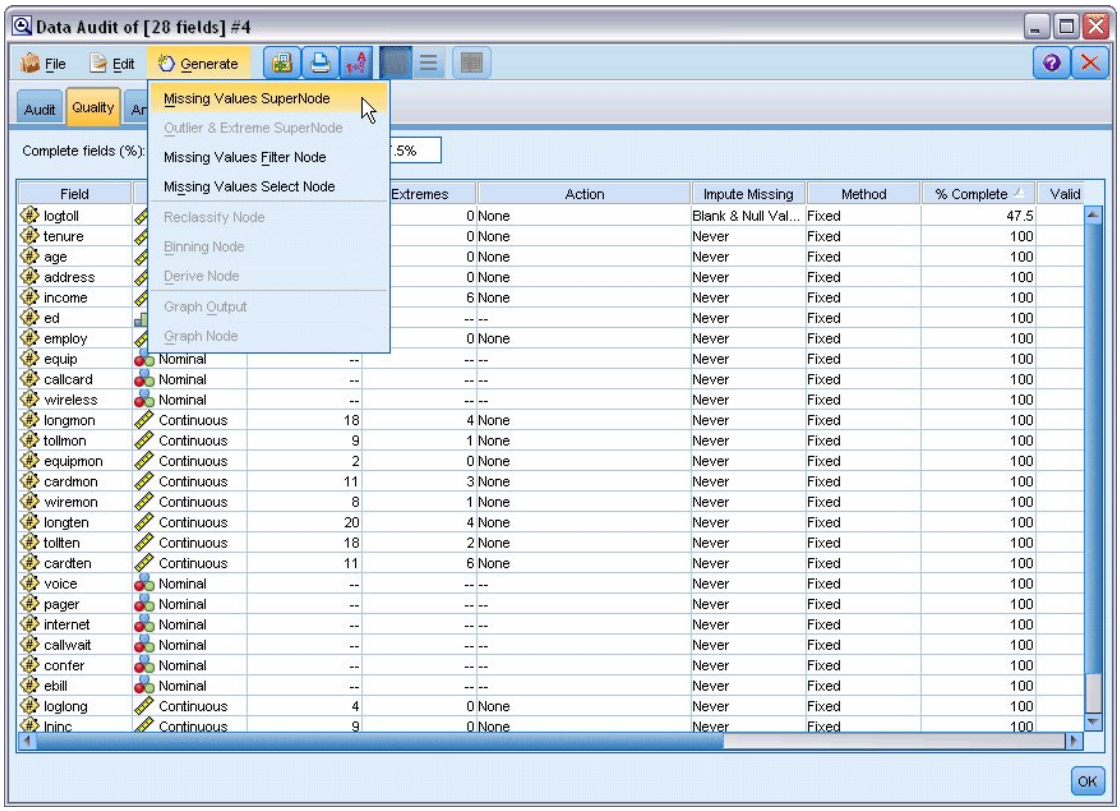
    Wybranie pozycji **Średnia** zapewnia, że podstawione wartości nie wpływają negatywnie na średnią wszystkich wartości w ogólnych danych.



Rysunek 160. Wybieranie ustawień podstawiania

13. Na karcie Jakość przeglądarki audytu danych wygeneruj Superwęzeł braków danych. W tym celu wybierz z menu:

    Utwórz > Superwęzeł braków danych

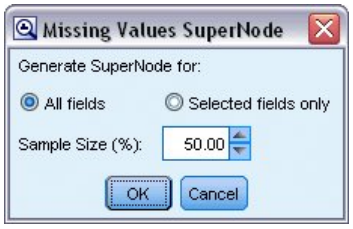


Rysunek 161. Generowanie superwęzła braków danych

W oknie dialogowym Superwęzeł braków danych zwiększ wartość **Wielkość próby** do 50% i kliknij przycisk **OK**.

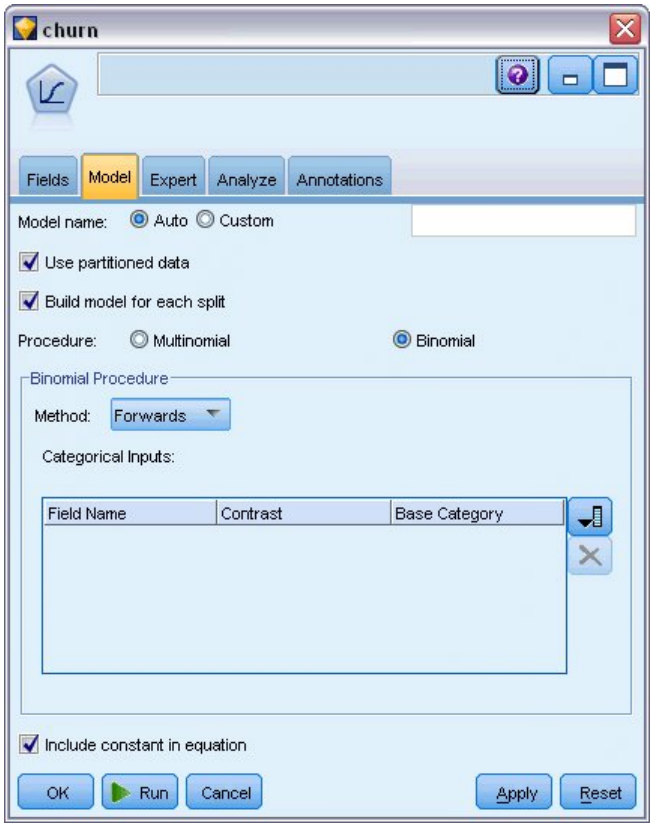
Superwęzeł jest wyświetlany w obszarze roboczym strumienia z tytułem: *Podstawianie brakujących wartości*.

14. Załącz superwęzeł do węzła filtrowania.



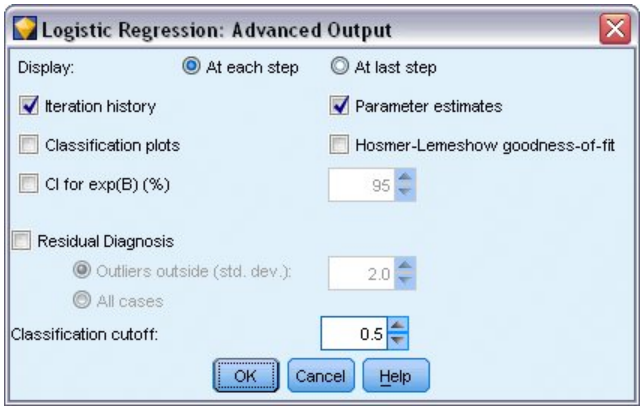
Rysunek 162. Określanie wielkości próby

15. Dodaj węzeł logistyczny do superwęzła.
16. W węźle logistycznym kliknij kartę Model i wybierz procedurę **Dwumianowa**. W obszarze *Procedura dwumianowa* wybierz metodę **Postępująca**.



Rysunek 163. Wybieranie opcji modelu

17. Na karcie Zaawansowany wybierz tryb **Zaawansowany**, a następnie kliknij opcje **Wynik**. Zostanie wyświetlone okno dialogowe Zaawansowane dane wyjściowe.
18. W oknie dialogowym Zaawansowane dane wyjściowe wybierz pozycję **W każdym kroku** jako typ opcji *Wyświetl*. Zaznacz pozycje **Przebieg iteracji** i **Oszacowania parametrów** i kliknij przycisk **OK**.



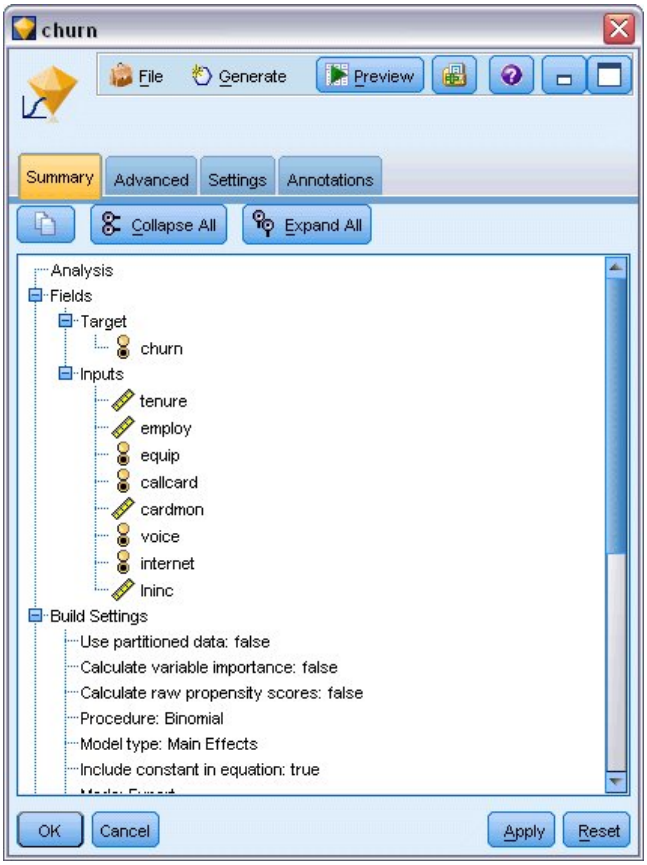
Rysunek 164. Wybieranie opcji wyników

Przeglądanie modelu

1. W węźle logistycznym kliknij przycisk **Uruchom**, aby utworzyć model.

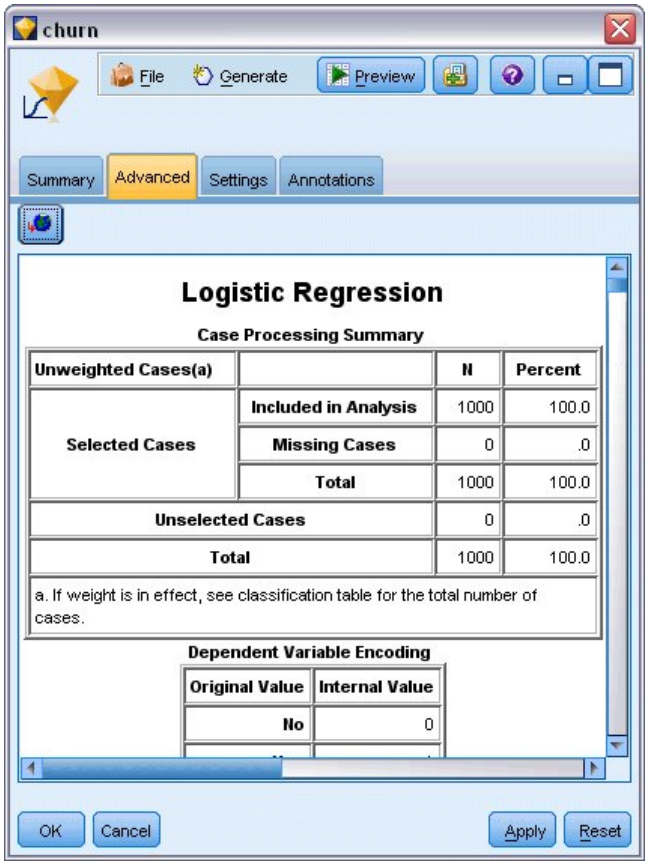
   Model użytkowy zostaje dodany do obszaru roboczego strumienia oraz na palecie modeli w prawym górnym rogu. Aby wyświetlić szczegóły modelu użytkowego, kliknij go prawym przyciskiem myszy i wybierz opcję **Edytuj** lub **Przeglądaj**.

   Karta Podsumowanie przedstawia (między innymi) zmienną przewidywaną i dane wejściowe (zmienne predykcyjne) używane przez model. Należy pamiętać, że są to zmienne, które zostały wybrane na podstawie metody postępującej, a nie pełna lista przekazana do analizy.



Rysunek 165. Podsumowanie modelu przedstawiające zmienną przewidywaną i zmienne wejściowe

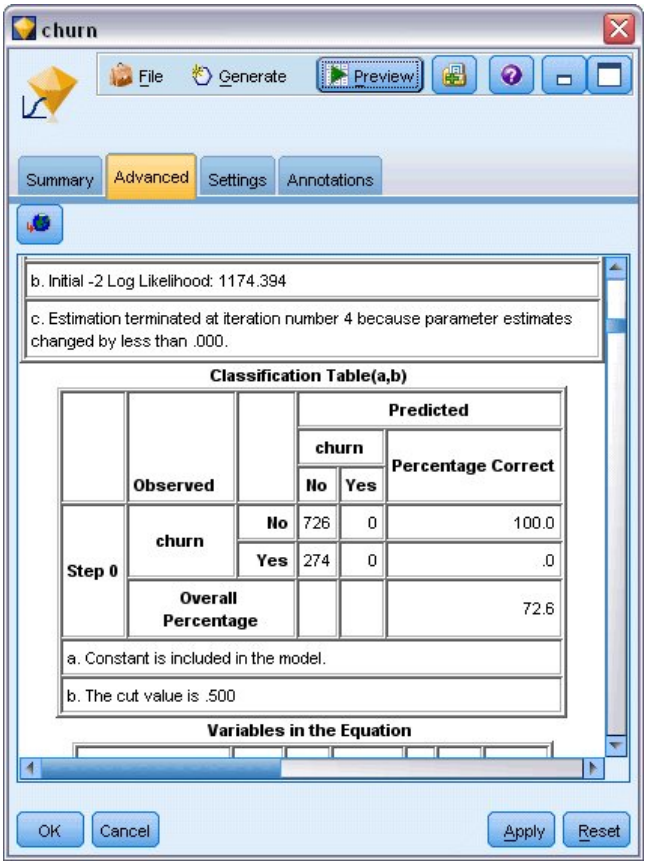
Pozycje wyświetlane na karcie Zaawansowany zależą od opcji wybranych w oknie dialogowym Zaawansowane dane wyjściowe w węźle logistycznym. Jedna pozycja, która jest zawsze wyświetlana, to Informacja o analizowanych danych przedstawiająca liczbę i udział procentowy rekordów uwzględnionych w analizie. Dodatkowo pozycja wymienia brakujące obserwacje (jeśli istnieją), gdzie co najmniej jedna zmienna nie jest dostępna oraz obserwacje, które nie zostały wybrane.



Rysunek 166. Informacja o analizowanych danych

2. Przewiń w dół z obszaru Informacja o analizowanych danych, aby wyświetlić tabelę klasyfikacji w obszarze Blok 0: Blok początkowy.

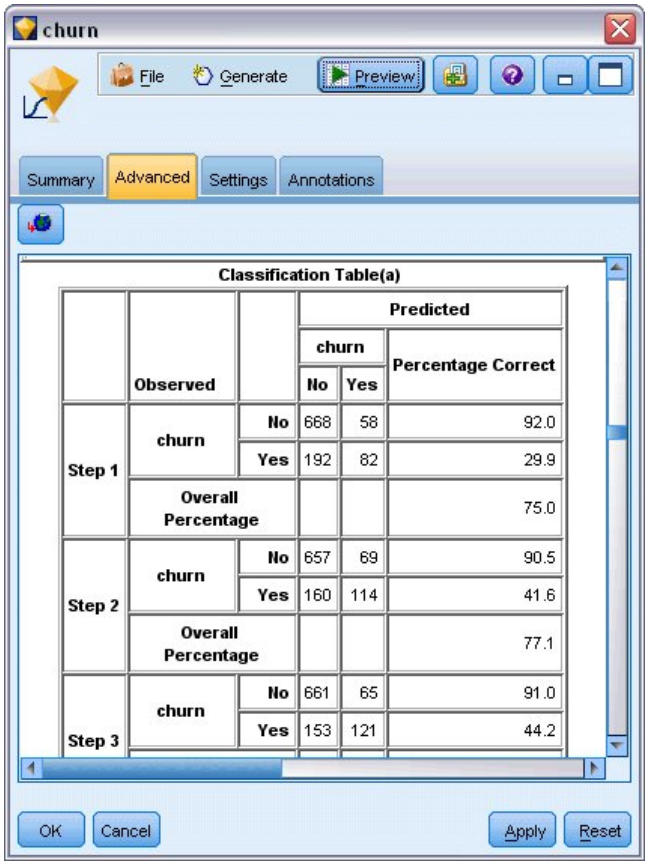
   Metoda Krokowa postępująca rozpoczyna działanie od modelu zerowego, czyli modelu bez predyktorów, którego można użyć jako podstawy do porównania z końcowym zbudowanym modelem. Model zerowy według konwencji przewiduje wszystko jako 0, więc model ten ma 72,6% dokładności, ponieważ przewidziano poprawnie 726 klientów, którzy nie odeszli. Jednak nie przewidziano poprawnie klientów, którzy odeszli.



Rysunek 167. Początkowa tabela klasyfikacji — Blok 0

3. Następnie przewiń w dół, aby wyświetlić tabelę klasyfikacji w obszarze Blok 1: Metoda = selekcja postępująca.

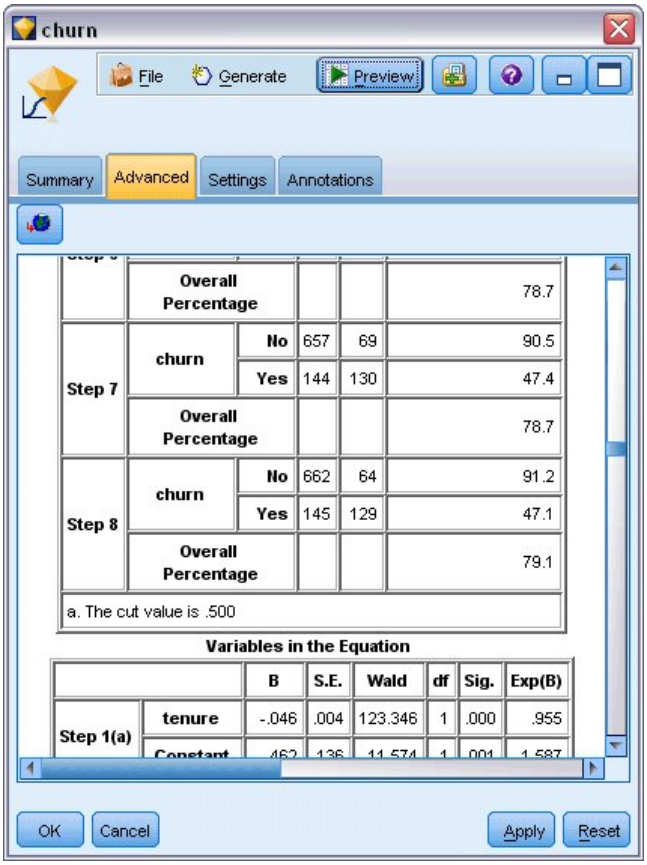
   Ta tabela klasyfikacji przedstawia wyniki dla modelu, gdy dodawany jest predyktor w każdym z kroków. Już w pierwszym kroku, po użyciu tylko jednego predyktora, model zwiększył dokładność predykcji odejścia z 0,0% do 29,9%.



Rysunek 168. Tabela klasyfikacji — Blok 1

4. Przewiń w dół do samego końca tej tabeli klasyfikacji.

Tabela klasyfikacji pokazuje, że ostatni krok to krok 8. W tym momencie algorytm zadecydował, że nie potrzebuje już dodawać dalszych predyktorów do modelu. Mimo że dokładność dla klientów, którzy nie odeszli, zmniejszyła się trochę do 91,2%, dokładność predykcji dla klientów odchodzących wzrosła z oryginalnej wartości 0% do 47,1%. Jest to istotna poprawa względem oryginalnego modelu zerowego, który nie używał predyktorów.



Rysunek 169. Tabela klasyfikacji — Blok 1

Dla klientów, którzy chcą ograniczyć poziom odejścia, możliwość ograniczenia go o prawie połowę byłoby dużym krokiem w stronę ochrony źródeł przychodu.

Uwaga: Ten przykład pokazuje również, jak przyjęcie wskaźnika Wartości procentowe ogółem jako oceny dokładności modelu może być w niektórych przypadkach mylące. Oryginalny model zerowy miał ogólną dokładność wynoszącą 72,6%, podczas gdy końcowy przewidywany model ma ogólną dokładność wynoszącą 79,1%. Jednak, tak jak pokazano, dokładność predykcji indywidualnych kategorii jest znacząco różna.

Aby ocenić, jak dobrze model jest dopasowany do danych, podczas budowania modelu w oknie dialogowym Zaawansowane dane wyjściowe dostępne są narzędzia diagnostyczne. Wyjaśnienia podstaw matematycznych metod modelowania używanych w programie IBM SPSS Modeler przedstawiono w publikacji *IBM SPSS Modeler Algorithms Guide* dostępnej w katalogu *\Documentation* na dysku instalacyjnym.

Należy zauważyć, że wyniki opierają się tylko na danych uczących. Aby ocenić, jak dobrze model uogólnia inne dane w rzeczywistych przypadkach, należałoby użyć węzła podziału na podzbiory, aby zatrzymać podzbiór rekordów do celów testowania i walidacji.

Rozdział 14. Prognozowanie wykorzystania pasma (Szereg czasowy)

Prognozowanie za pomocą węzła Szereg czasowy

Analityk krajowego dostawcy usług szerokopasmowych ma wygenerować prognozy abonentów w celu predykcji stopnia wykorzystania przepustowości. Prognozy są potrzebne indywidualnie dla każdego z rynków lokalnych tworzących łącznie bazę abonentów krajowych. Użytkownik zastosuje modelowanie szeregów czasowych do wygenerowania prognoz dla wybranych rynków lokalnych na następne trzy miesiące. W drugim przykładzie pokazano, jak można przekształcić dane źródłowe, jeśli nie mają poprawnego formatu dla danych wejściowych węzła Szereg czasowy.

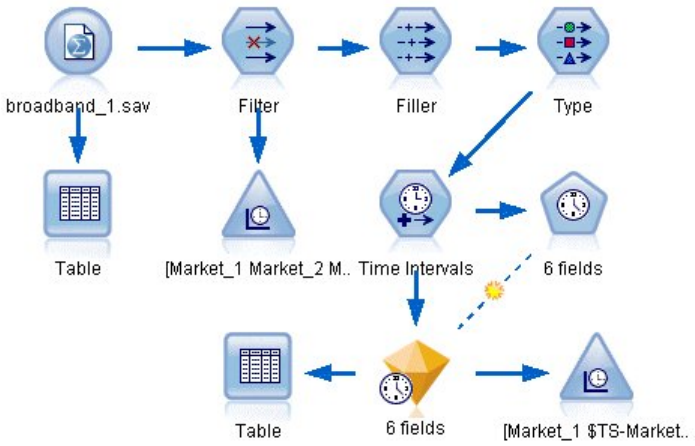
W tych przykładach zastosowano strumień o nazwie *broadband_create_models.str*, który odwołuje się do pliku danych o nazwie *broadband_1.sav*. Te pliki są dostępne w folderze *Demos* instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *broadband_create_models.str* znajduje się w folderze *streams*.

Ostatni przykład pokazuje, jak zastosować zapisane modele na zaktualizowanym zbiorze danych, aby rozszerzyć prognozy o kolejne trzy miesiące.

W programie IBM SPSS Modeler można przygotować wiele modeli szeregów czasowych w jednej operacji. Plik źródłowy, który będzie używany posiada dane szeregu czasowego dla 85 różnych rynków, mimo że w celu uproszczenia model będzie zawierał tylko pięć z tych rynków oraz sumę dla wszystkich rynków.

Plik danych *broadband_1.sav* zawiera miesięczne dane użycia dla każdego z 85 lokalnych rynków. Dla celów tego przykładu użytych zostanie tylko pierwszych pięć szeregów. Zostanie utworzony osobny model dla każdego z tych pięciu szeregów oraz dla sumy.

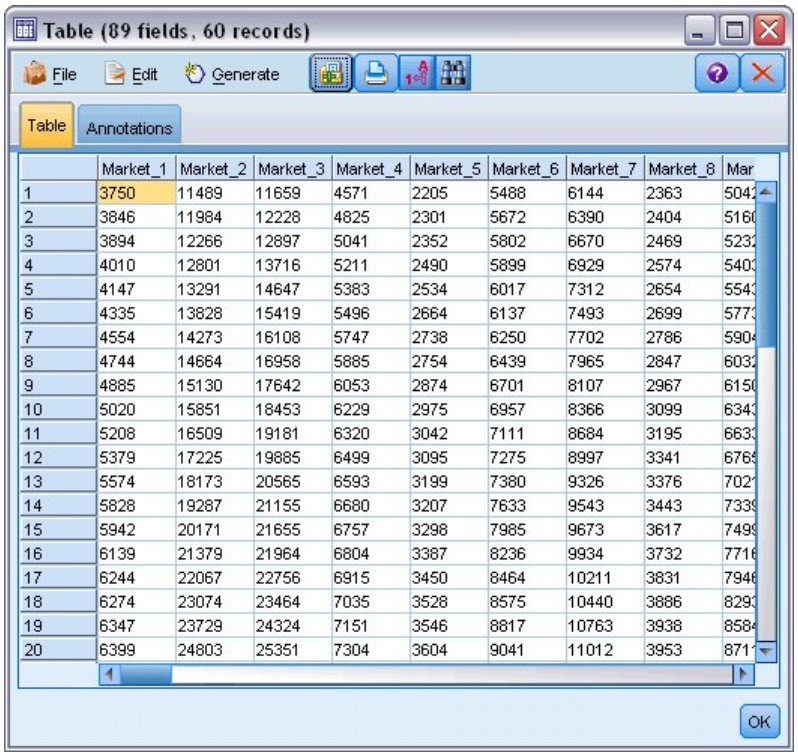
Plik zawiera również zmienną daty, która wskazuje miesiąc i rok dla każdego rekordu. Ta zmienna będzie używana w węźle Przedziały czasowe, aby określać etykiety rekordów. Zmienna danych jest wczytywana do programu IBM SPSS Modeler jako łańcuch, ale aby mogła być użyta w programie IBM SPSS Modeler, należy przekształcić typ składowania na numeryczny format daty, używając węzła wypełniania.



Rysunek 170. Przykładowy strumień przedstawiający modelowanie szeregów czasowych

Węzeł Szereg czasowy wymaga, aby każdy szereg znajdował się w osobnej kolumnie z wierszem dla każdego przedziału. Program IBM SPSS Modeler zapewnia metody przekształcania danych w celu dopasowania formatu, jeśli

jest to wymagane.



	Market_1	Market_2	Market_3	Market_4	Market_5	Market_6	Market_7	Market_8	Mar
1	3750	11489	11659	4571	2205	5488	6144	2363	5042
2	3846	11984	12228	4825	2301	5672	6390	2404	5160
3	3894	12266	12897	5041	2352	5802	6670	2469	5232
4	4010	12801	13716	5211	2490	5899	6929	2574	5403
5	4147	13291	14647	5383	2534	6017	7312	2654	5543
6	4335	13828	15419	5496	2664	6137	7493	2699	5773
7	4554	14273	16108	5747	2738	6250	7702	2786	5904
8	4744	14664	16958	5885	2754	6439	7965	2847	6032
9	4885	15130	17642	6053	2874	6701	8107	2967	6150
10	5020	15851	18453	6229	2975	6957	8366	3099	6343
11	5208	16509	19181	6320	3042	7111	8684	3195	6633
12	5379	17225	19885	6499	3095	7275	8997	3341	6765
13	5574	18173	20565	6593	3199	7380	9326	3376	7027
14	5828	19287	21155	6680	3207	7633	9543	3443	7339
15	5942	20171	21655	6757	3298	7985	9673	3617	7499
16	6139	21379	21964	6804	3387	8236	9934	3732	7716
17	6244	22067	22756	6915	3450	8464	10211	3831	7946
18	6274	23074	23464	7035	3528	8575	10440	3886	8293
19	6347	23729	24324	7151	3546	8817	10763	3938	8584
20	6399	24803	25351	7304	3604	9041	11012	3953	8711

Rysunek 171. Miesięczne dane subskrypcji dla lokalnych rynków usług szerokopasmowych

Tworzenie strumienia

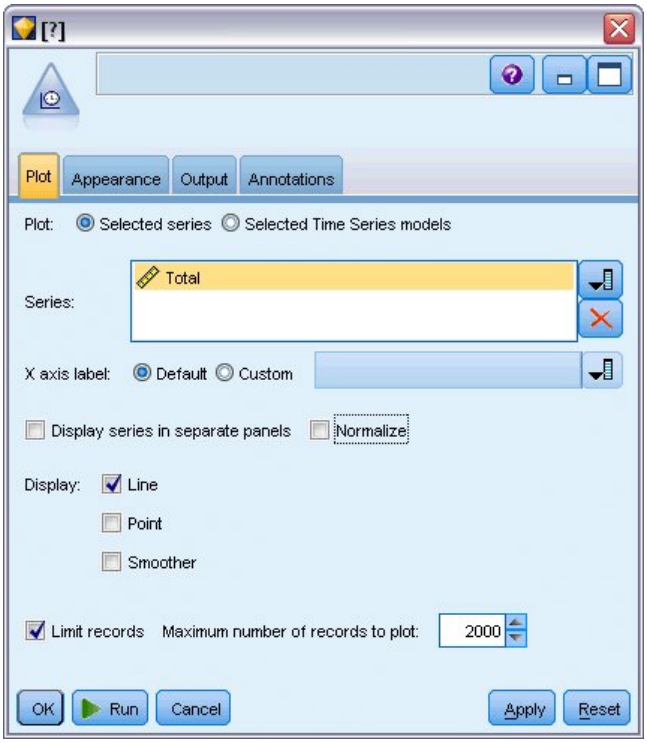
1. Utwórz nowy strumień i dodaj węzeł źródłowy Plik Statistics wskazujący na plik *broadband_1.sav*.
2. Użyj węzła filtrowania, aby odfiltrować zmienne od *Market_6* do *Market_85* oraz zmienne *MONTH_* i *YEAR_* w celu uproszczenia modelu.

Wskazówka: Aby zaznaczyć wiele sąsiadujących zmiennych w jednej operacji, kliknij zmienną *Market_6*, przytrzymaj naciśnięty lewy przycisk myszy i przeciągnij w dół do zmiennej *Market_85*. Zaznaczone pola są podświetlane na niebiesko. Aby dodać inne zmienne, przytrzymaj naciśnięty klawisz Ctrl i kliknij zmienne *MONTH_* i *YEAR_*.

Rysunek 172. Upraszczanie modelu

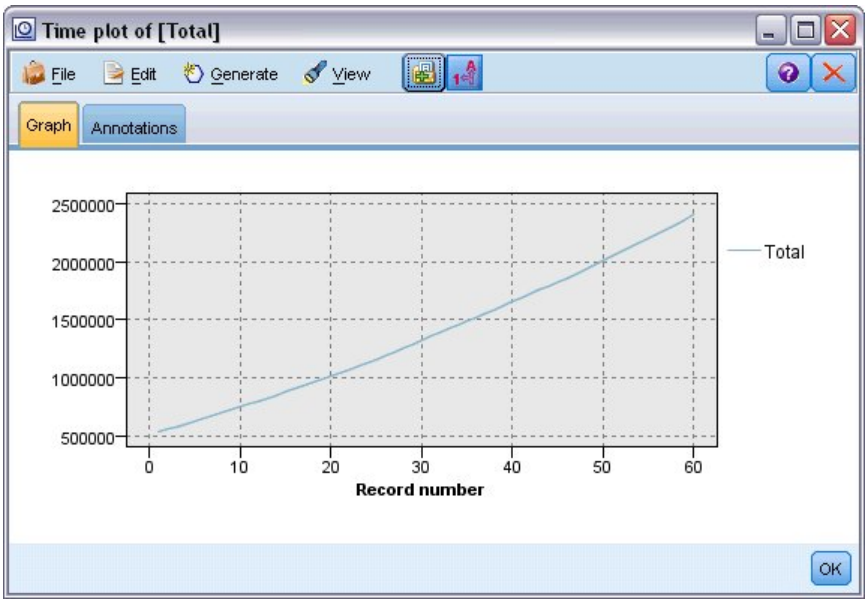
Badanie danych

Zawsze dobrze jest zrozumieć charakter danych przed rozpoczęciem budowania modelu. Czy dane wykazują zmienność sezonową? Mimo że narzędzie Automatyczny dobór modelu może automatycznie znaleźć najlepszy model sezonowy i niesezonowy dla każdego szeregu, można uzyskać szybsze wyniki, ograniczając wyszukiwanie do modeli niesezonowych, jeśli sezonowość nie występuje w danych. Bez badania danych dla każdego lokalnego rynku można uzyskać ogólny obraz obecności lub braku sezonowości, kreśląc całkowitą liczbę użytkowników usług na wszystkich pięciu rynkach.



Rysunek 173. Tworzenie wykresu całkowitej liczby użytkowników usługi

1. Na palecie Wykresy dołącz węzeł Sekwencyjny do węzła filtrowania.
2. Dodaj zmienną *Ogółem* do listy Szeregi.
3. Usuń zaznaczenie pól **Wyświetl szeregi w oddzielnych panelach** i **Normalizuj**.
4. Kliknij przycisk **Uruchom**.

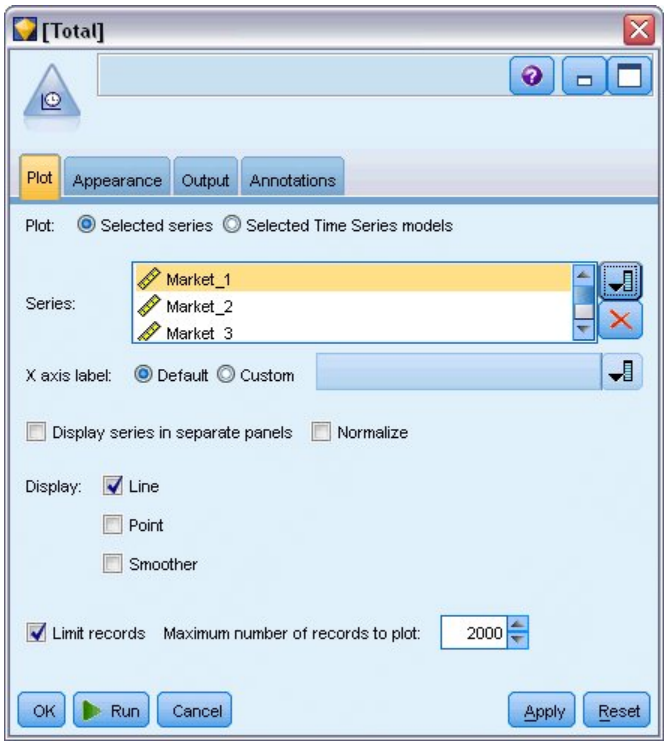


Rysunek 174. Wykres sekwencyjny zmiennej Ogółem

Ten szereg wykazuje bardzo gładki trend w górę bez oznak wariacji sezonowych. Mogą istnieć indywidualne serie bez sezonowości, ale wygląda na to, że sezonowość nie jest istotną cechą tych danych.

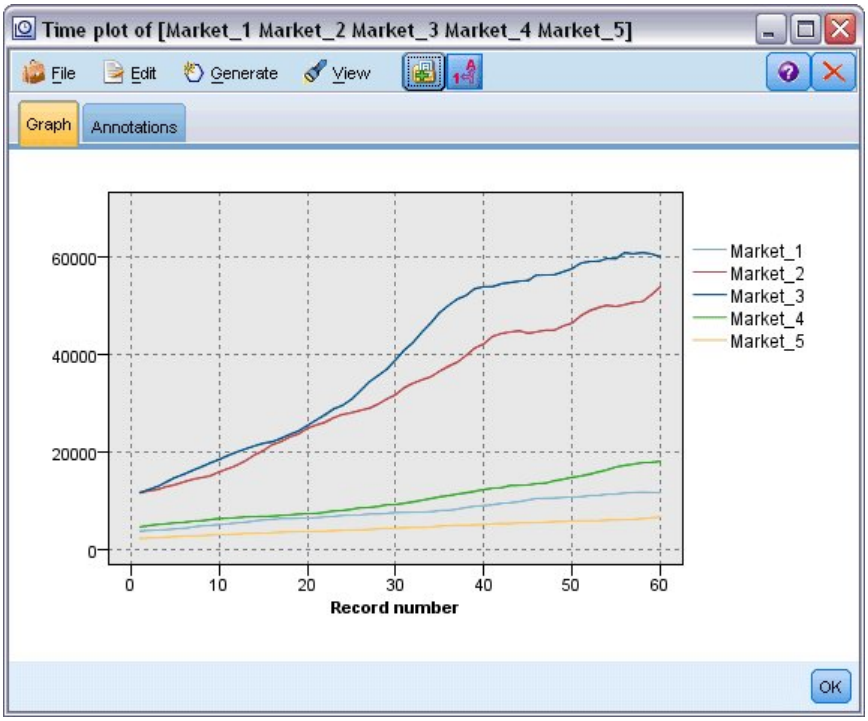
Oczywiście należy zbadać każdy szereg przed wykluczeniem modeli sezonowych. Można następnie wydzielić szeregi wykazujące sezonowość i utworzyć osobny model.

Program IBM SPSS Modeler umożliwia łatwe tworzenie wykresów dla wielu szeregów.



Rysunek 175. Tworzenie wykresów wielu szeregów czasowych

5. Otwórz ponownie węzeł wykresu sekwencyjnego.
6. Usuń zmienną *Total* z listy szeregów (zaznacz ją i kliknij czerwony przycisk X).
7. Dodaj do listy zmienne od *Market_1* do *Market_5*.
8. Kliknij przycisk **Uruchom**.



Rysunek 176. Wykres sekwencyjny wielu zmiennych

Badanie każdego rynku ujawnia stały trend rosnący w każdym z przypadków. Mimo że niektóre rynki są bardziej zmienne niż inne, nie istnieje dowód na sezonowość.

Definiowanie dat

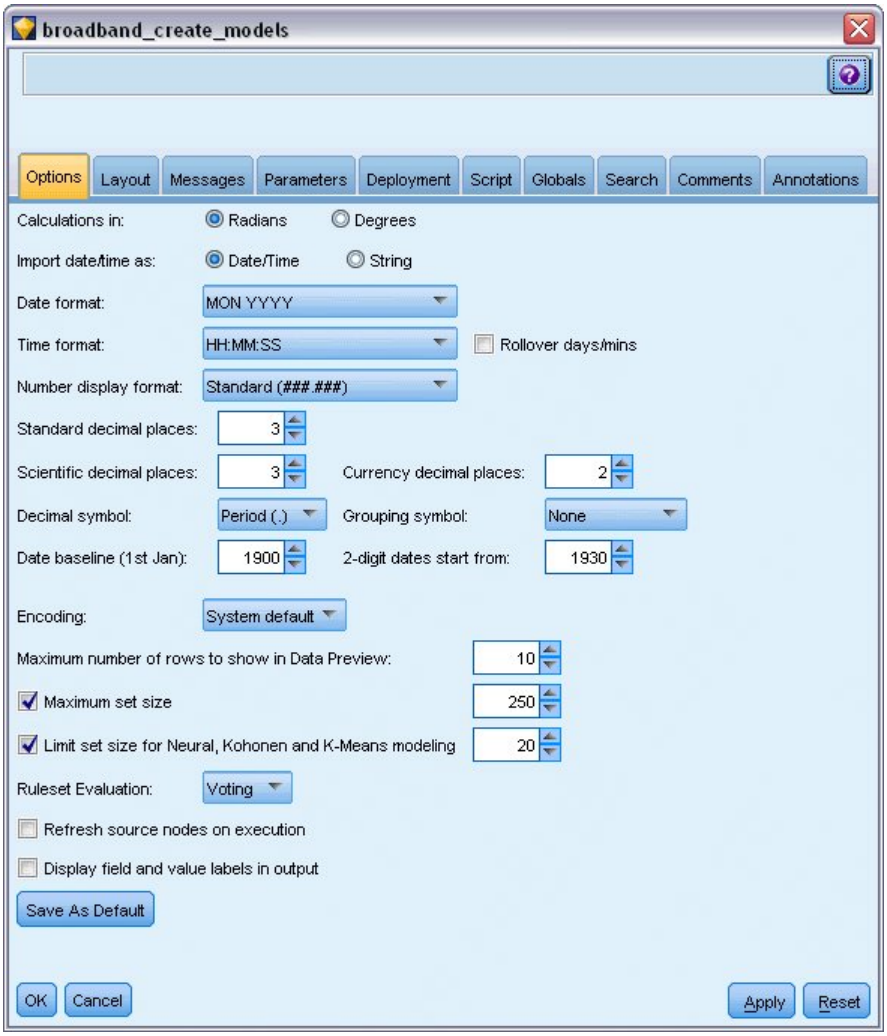
Teraz trzeba zmienić typ składowania zmiennej *DATE_* na format daty.

1. Załącz węzeł wypełniania do węzła filtrowania.
2. Otwórz węzeł wypełniania i kliknij przycisk wyboru zmiennych.
3. Wybierz zmienną **DATE_**, aby dodać ją do obszaru **Wypełnij zmienne**.
4. Ustaw wartość warunku **Zamień** na **Zawsze**.
5. Ustaw wartość obszaru **Zamień na** na **to_date(DATE_)**.

Rysunek 177. Ustawianie typu składowania danych

Zmień domyślny format daty, aby odpowiadał formatowi zmiennej daty. Jest to niezbędne, aby przekształcenie zmiennej daty działało w oczekiwany sposób.

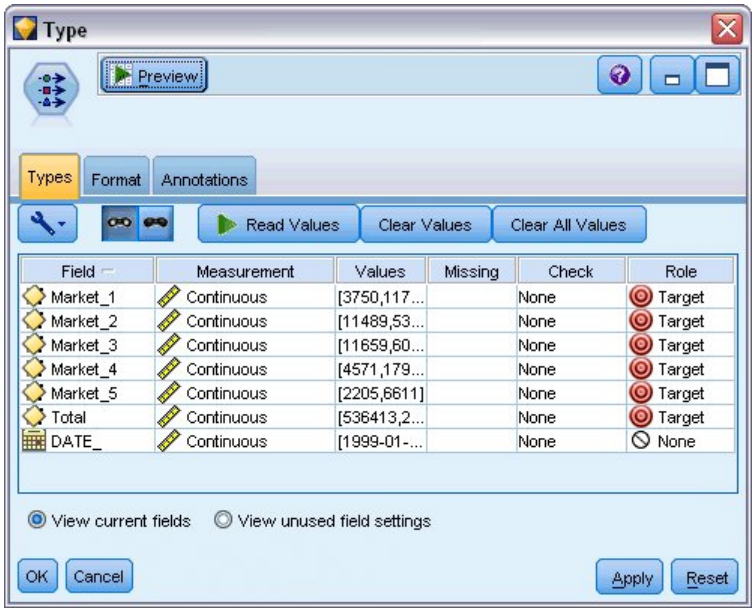
6. W menu wybierz kolejno opcje **Narzędzia > Właściwości strumienia > Opcje**, aby wyświetlić okno dialogowe opcji strumienia.
7. Ustaw domyślną wartość parametru **Format daty** na **MMM RRRR** .



Rysunek 178. Ustawianie formatu daty

Definiowanie zmiennych przewidywanych

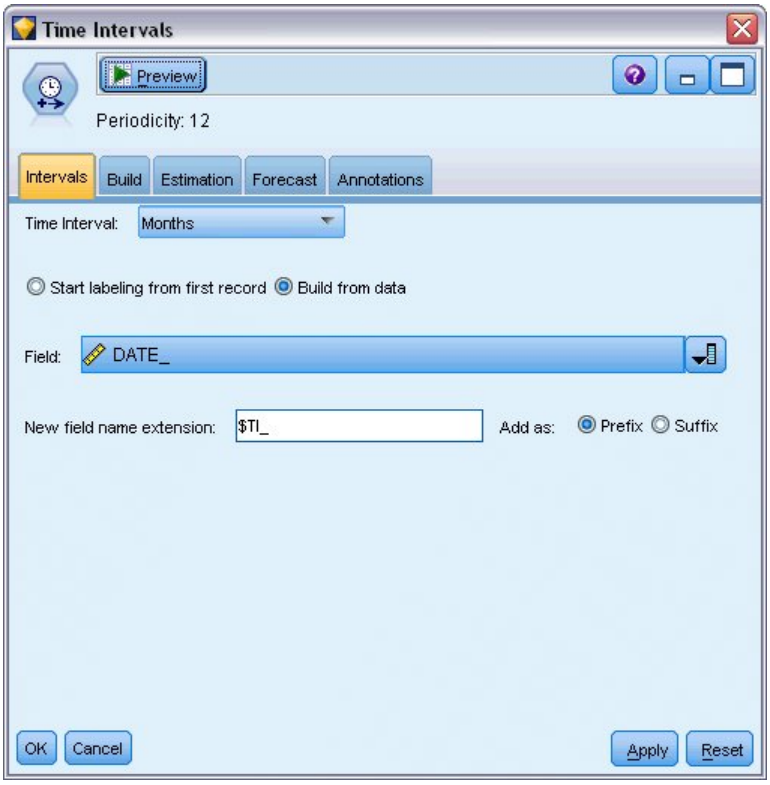
1. Dodaj węzeł typu i ustaw rolę na **Brak** dla zmiennej *DATE_* . Ustaw rolę na **Przewidywana** dla wszystkich pozostałych zmiennych (zmiennych *Market_n* i zmiennej *Total*).
2. Kliknij przycisk **Odczytaj wartości**, aby wypełnić kolumnę Wartości.



Rysunek 179. Ustawianie roli dla wielu zmiennych

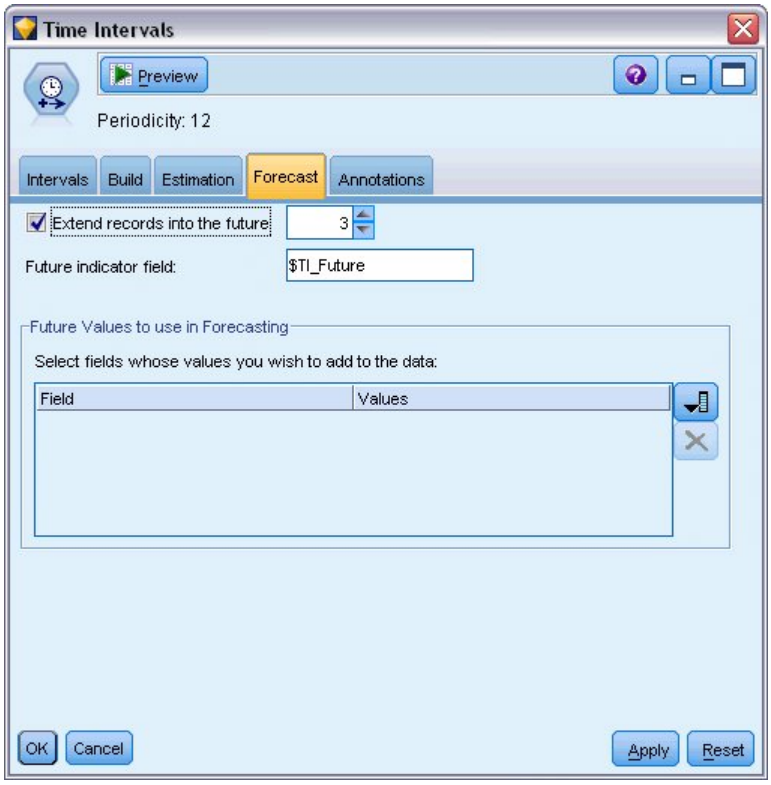
Ustawianie przedziałów czasowych

1. Dodaj węzeł Przedziały czasowe (z palety węzłów operacji na zmiennych).
2. Na karcie Przedziały wybierz **Miesiące** jako przedział czasowy.
3. Zaznacz opcję **Utwórz na podstawie danych**.
4. Wybierz **DATE_** jako zmienną budowy.



Rysunek 180. Ustawianie przedziałów czasowych

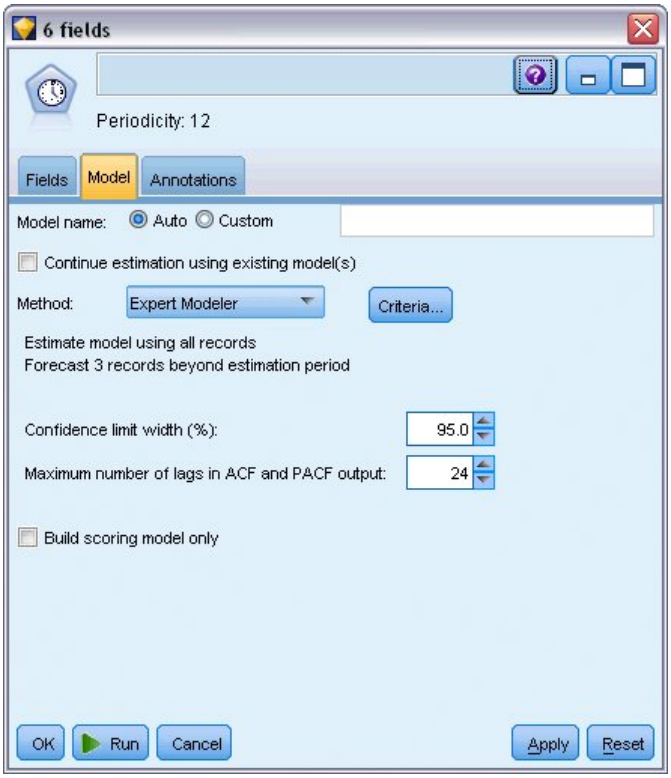
5. Na karcie Prognoza zaznacz pole wyboru **Rozszerz rekordy na przedziały z przyszłości**.
6. Ustaw wartość na **3**.
7. Kliknij przycisk **OK**.



Rysunek 181. Ustawianie okresu prognozy

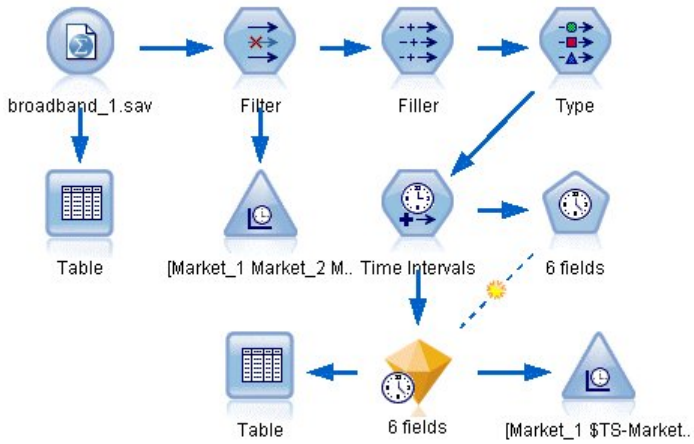
Tworzenie modelu

1. Z palety Modelowanie dodaj węzeł Szereg czasowy do strumienia i dołącz go do węzła Przedziały czasowe.
2. Kliknij przycisk **Uruchom** na węźle Szereg czasowy, używając ustawień domyślnych. W ten sposób narzędzie Automatyczny dobór modelu decyduje, jaki jest najbardziej odpowiedni model do użycia dla każdego szeregu czasowego.



Rysunek 182. Wybieranie narzędzia Automatyczny dobór modelu dla szeregów czasowych

3. Załącz model użytkowy Szereg czasowy do węzła Przedziały czasowe.
4. Załącz węzeł tabeli do modelu Szereg czasowy i kliknij przycisk **Uruchom**.



Rysunek 183. Przykładowy strumień przedstawiający modelowanie szeregów czasowych

Istnieją teraz trzy nowe wiersze (61 do 63) dodane do oryginalnych danych. Są to wiersze dla okresu prognozy, w tym przypadku od stycznia do marca 2004 r.

Istnieje teraz kilka nowych kolumn: kolumny *$TI_* dodane przez węzeł Przedziały czasowe oraz kolumny *$TS-* dodane przez węzeł Szereg czasowy. Kolumny wskazują następujące informacje dla każdego wiersza (tzn. każdego przedziału w danych szeregów czasowych):

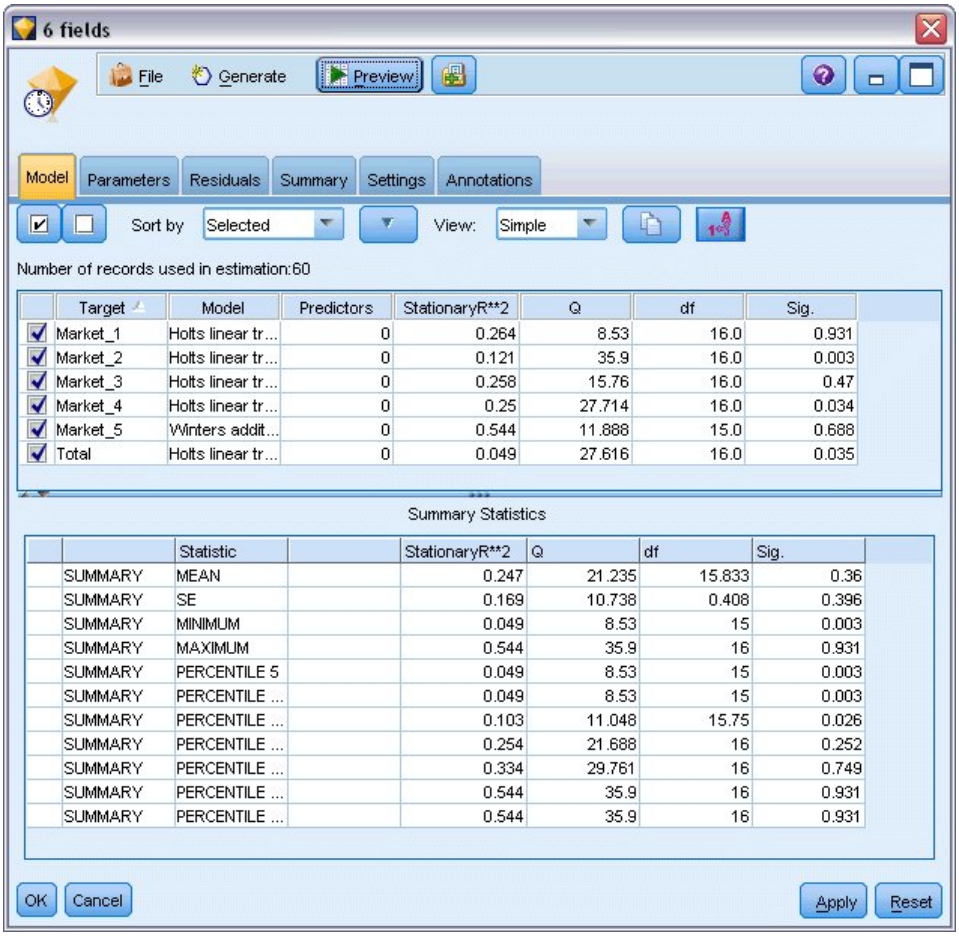
Kolumna	Opis
$TI_TimeIndex	Wartość indeksu przedziału czasowego dla tego wiersza.
$TI_TimeLabel	Etykieta przedziału czasowego dla tego wiersza.
$TI_Year	Wskaźniki roku i miesiąca dla danych wygenerowanych w tym wierszu.
$TI_Month	
$TI_Count	Liczba rekordów zaangażowanych w określenie nowych danych dla tego wiersza.
$TI_Future	Wskazuje, czy ten wiersz zawiera dane prognozy.
$TS-*colname*	Dane wygenerowanego modelu dla każdej kolumny oryginalnych danych.
$TSLCI-*colname*	Niższa wartość przedziału ufności dla każdej kolumny wygenerowanych danych modelu.
$TSUCI-*colname*	Wyższa wartość przedziału ufności dla każdej kolumny wygenerowanych danych modelu.
$TS-Total	Suma wartości $TS-*colname* dla tego wiersza.
$TSLCI-Total	Suma wartości $TSLCI-*colname* dla tego wiersza.
$TSUCI-Total	Suma wartości $TSUCI-*colname* dla tego wiersza.

Najbardziej znaczące kolumny dla operacji prognozy to *$TS-Market_n*, *$TSLCI-Market_n* i *$TSUCI-Market_n*. W szczególności te kolumny w wierszach od 61 do 63 zawierają dane prognozy subskrypcji użytkowników oraz przedziały ufności dla każdego lokalnego rynku.

Badanie modelu

1. Dwukrotnie kliknij model użytkowy Szereg czasowy, aby wyświetlić dane o modelach wygenerowanych dla każdego z rynków.

   Warto zwrócić uwagę na to, jak Automatyczny dobór modelu zadecydował, aby wygenerować inny typ modelu dla zmiennej Market_5 niż dla pozostałych rynków.



Rysunek 184. Modele szeregów czasowych wygenerowane dla rynków

Kolumna Predyktory pokazuje, ile zmiennych zostało użytych jako predyktory dla każdej zmiennej przewidywanej — w tym przypadku żadna.

Pozostałe kolumny w tym widoku przedstawiają różne miary dobroci dopasowania dla każdego modelu. Kolumna **StacjonarnyR**2** przedstawia wartość Stacjonarny *R*-kwadrat. Ta statystyka zapewnia ocenę proporcji wariancji ogółem w szeregu wyjaśnionej przez model. Im wyższa wartość (przy maksimum wynoszącym 1,0), tym lepsze dopasowanie modelu.

Kolumny **Q**, **df** i **Istotność** dotyczą statystyki Ljunga-Boxa — testu losowości błędów resztowych w modelu — im bardziej losowe błędy, tym lepszy będzie model. **Q** to sama statystyka Ljunga-Boxa, podczas gdy wartość **df** (stopnie swobody) wskazuje liczbę parametrów modelu, które mogą się zmieniać podczas szacowania konkretnej zmiennej przewidywanej.

Wartości **Istotność** przedstawia wartość istotności statystyki Ljunga-Boxa, zapewniając kolejny wskaźnik tego, czy model został określony poprawnie. Wartość istotności niższa niż 0,05 wskazuje, że błędy resztowe nie są losowe, sugerując, że w obserwowanym szeregu istnieje struktura, która nie została uwzględniona w modelu.

Biorąc pod uwagę wartości Stacjonarny *R*-kwadratowy i Istotność, modele wybrane przez narzędzie Automatyczny dobór modelu dla zmiennych *Market_1*, *Market_3* i *Market_5* są dopuszczalne. Wartości **Istotność** dla zmiennych *Market_2* i *Market_4* są mniejsze niż 0,05, co wskazuje, że niezbędne może być eksperymentowanie z lepiej dopasowanymi modelami.

Wartości podsumowania w dolnej części zapewniają informacje o dystrybucji statystyk we wszystkich modelach. Na przykład średnia wartość Stacjonarny *R*-kwadrat we wszystkich modelach to 0,247, podczas gdy wartość minimalna to 0,049 (modelu *Total*), wartość maksymalna 0,544 (wartość dla zmiennej *Market_5*).

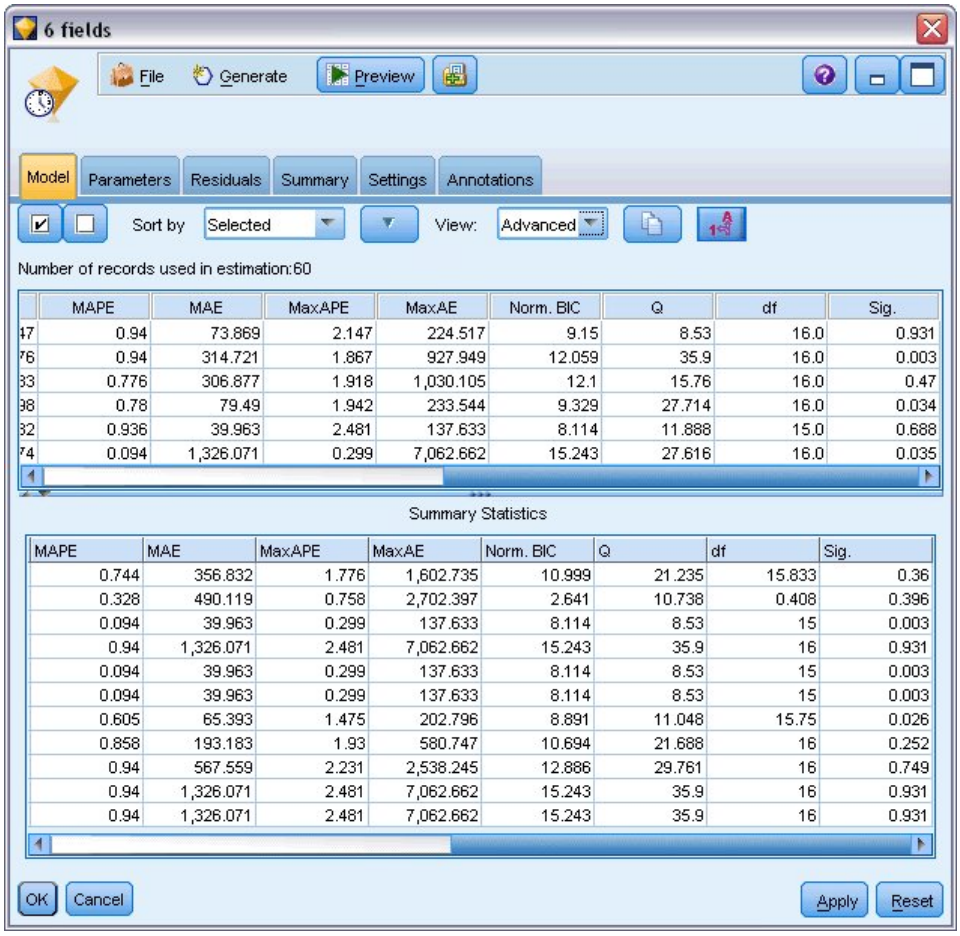
Błąd standardowy opisuje błąd standardowy we wszystkich modelach dla każdej statystyki. Na przykład błąd standardowy dla statystyki Stacjonarny *R*-kwadrat we wszystkich modelach to 0,169.

Sekcja podsumowania obejmuje również wartości percentyla, które zapewniają informacje o rozkładzie statystyk w modelach. Dla każdego percentyla taki procent modeli ma wartość dopasowanej statystyki poniżej określonej wartości.

Na przykład: tylko 25% modeli ma wartość Stacjonarny *R*-kwadrat niższą niż 0,121.

2. Kliknij listę rozwijaną Widok i wybierz pozycję **Zaawansowane**.

   W oknie wyświetlane są dodatkowe miary dobroci dopasowania. **R**2** to wartość *R*-kwadrat, estymacja wariancji ogółem w szeregu czasowym, która może być wyjaśniona modelem. Ponieważ maksymalna wartość tej statystyki to 1,0, nasze modele mają dobre wyniki pod tym względem.



Rysunek 185. Zaawansowany widok modeli szeregów czasowych

RMSE to pierwiastek błędu średniokwadratowego, miara tego, jak bardzo rzeczywiste wartości szeregu różnią się od wartości przewidywanych przez model i jest wyrażany w takich samych jednostkach, jak używane w szeregu. Ponieważ jest to miara błędu, chcemy aby wartości były tak niskie, jak to możliwe. Na pierwszy rzut oka wygląda, że modele dla zmiennych *Market_2* i *Market_3*, mimo że są wciąż dopuszczalne według statystyk, które wcześniej widzieliśmy, są mniej skuteczne niż modele dla pozostałych trzech rynków.

Te dodatkowe miary dobroci dopasowania obejmują średnie bezwzględne błędy procentowe (**MAPE**) oraz ich maksymalną wartość (**MaxAPE**). Bezwzględny błąd procentowy to miara tego, jak bardzo szereg przewidywany różni się od poziomu przewidywanego modelem i wyrażany jest jako wartość procentowa. Badając wartości średnie i maksymalne we wszystkich modelach, można uzyskać wgląd w niepewność predykcji.

Wartość MAPE pokazuje, że wszystkie modele wykazują średnią niepewność niższą niż 1%, co jest bardzo niską wartością. Wartość MaxAPE przedstawia maksymalny procentowy błąd bezwzględny i jest przydatna do obrazowania najgorszego przypadku dla prognoz. Statystyka pokazuje, że największy błąd procentowy dla każdego modelu przypada w zakresie od około 1,8% do 2,5%, co ponownie jest bardzo niską wartością.

Wartość **MAE** (średni błąd bezwzględny) przedstawia średnią wartości bezwzględnych błędów prognoz. Tak jak w przypadku wartości RMSE, wartość MAE jest wyrażana w takich samych jednostkach, jak te używane w szeregu. **MaxAE** przedstawia największy błąd prognozy w tych samych jednostkach i wskazuje najgorszy przypadek dla prognoz.

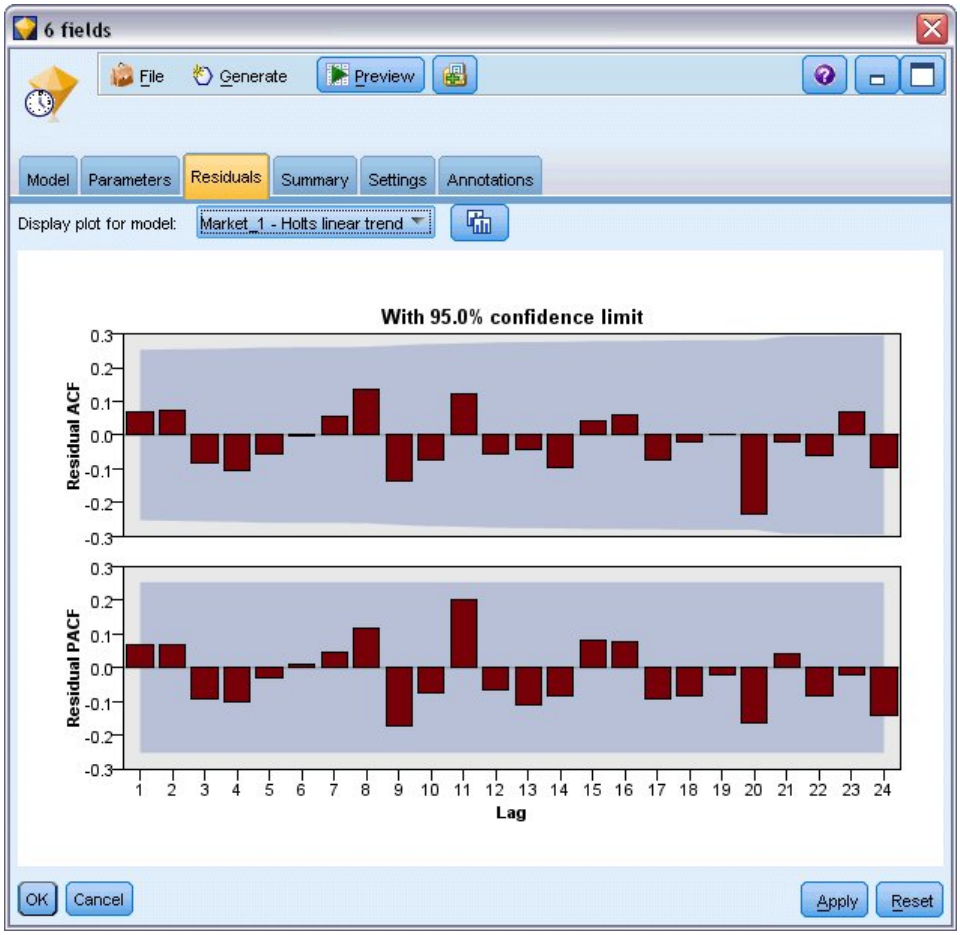
Mimo że te wartości bezwzględne są interesujące, w tym przypadku to procentowe wartości błędów (MAPE i MaxAPE) są bardziej przydatne, ponieważ szeregi przewidywane reprezentują liczbę subskrybentów na rynkach różnych wielkości.

Czy wartości MAPE i MaxAPE reprezentują dopuszczalną niepewność w tych modelach? Na pewno są bardzo niskie. W takiej sytuacji liczy się wiedza biznesowa, ponieważ dopuszczalne ryzyko zmienia się między problemami. Załóżymy, że statystyki dobroci dopasowania przypadają w dopuszczalnych zakresach i przejdziemy do poszukiwania błędów resztowych.

Badanie wartości funkcji autokorelacji (ACF) i funkcji autokorelacji cząstkowych (PACF) dla reszt modelu zapewnia bardziej ilościowy wgląd w modele niż po prostu przeglądanie statystyk dobroci dopasowania.

Dobrze określony szereg czasowy przechwyci wszystkie zmienności nielosowe, łącznie z sezonowością, trendem i zmianami cyklicznymi oraz innymi współczynnikami, które są istotne. Jeśli tak jest, błąd nie powinien być skorelowany ze sobą (nie występuje autokorelacja) w czasie. Istotna struktura w funkcjach autokorelacji wskazywałaby, że model jest niepełny.

3. Kliknij kartę Reszty, aby wyświetlić wartości funkcji autokorelacji (ACF) oraz funkcji autokorelacji cząstkowych (PACF) dla błędów resztowych w modelu dla pierwszego z lokalnych rynków.



Rysunek 186. Wartości ACF i PACF dla rynków

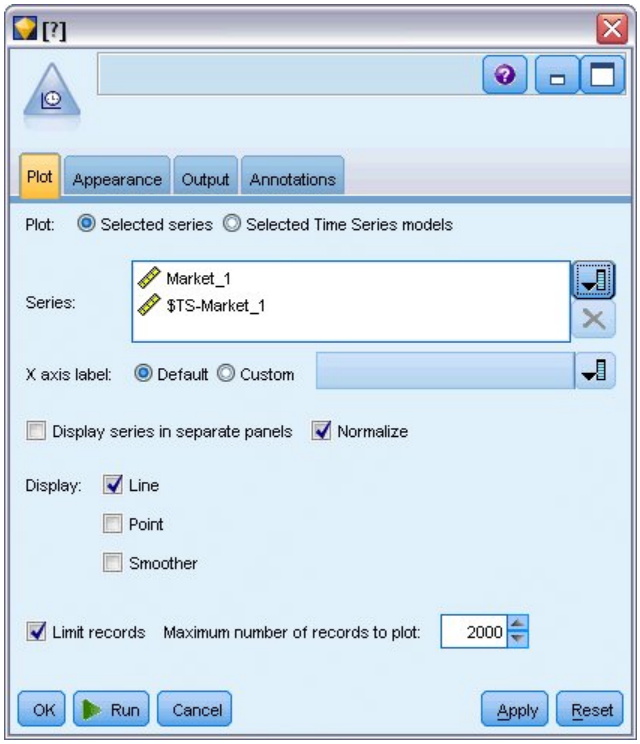
Na tych wykresach oryginalne wartości zmiennej błędu zostały opóźnione do 24 okresów czasowych i porównane z wartością oryginalną, aby zobaczyć, czy występuje korelacja w czasie. Aby model był dopuszczalny, żaden ze słupków na górnym wykresie (ACF) nie powinien wykraczać poza zacieniowany obszar w kierunku dodatnim (w górę) lub ujemnym (w dół).

Jeśli tak by się stało, należy sprawdzić dolny wykres (PACF), aby zobaczyć, czy struktura jest tam potwierdzona. Wykres PACF poszukuje korelacji po skontrolowaniu wartości szeregu pomiędzy punktami w czasie.

Wszystkie wartości zmiennej *Market_1* przypadają w zacieniowanym obszarze, więc możemy kontynuować i sprawdzić wartości dla innych rynków.

4. Kliknij listę rozwijaną **Wyświetl wykres dla modelu**, aby wyświetlić te wartości dla innych rynków i sum.

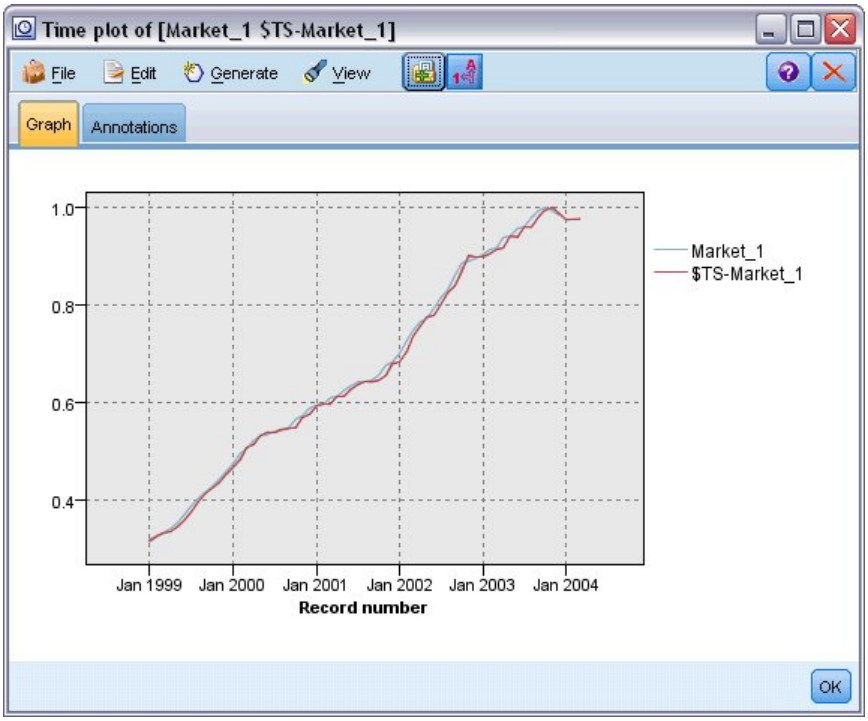
   Wartości dla zmiennych *Market_2* i *Market_4* nie są alarmujące, potwierdzając to, co było widać z ich wartości **Istotność** . Na pewnym etapie konieczne będzie wypróbowanie różnych modeli dla tych rynków, aby zobaczyć, czy można uzyskać lepsze dopasowanie, ale w pozostałej części tego przykładu skoncentrujemy się na tym, czego jeszcze możemy dowiedzieć się z modelu *Market_1*.
5. Na palecie Wykresy dołącz węzeł Sekwencyjny do modelu użytkowego Szereg czasowy.
6. Na karcie Wykres usuń zaznaczenie pola wyboru **Wyświetl szeregi w oddzielnych panelach**.
7. Na liście **Szeregi** kliknij przycisk wyboru zmiennych i zaznacz zmienne *Market_1* i *$TS-Market_1*, a następnie kliknij przycisk **OK**, aby dodać je do listy.
8. Kliknij przycisk **Uruchom**, aby wyświetlić wykres liniowy danych rzeczywistych i prognozowanych dla pierwszego z lokalnych rynków.



Rysunek 187. Wybieranie zmiennych do umieszczenia na wykresie

Zwróćmy uwagę, jak linia prognozy (*$TS-Market_1*) rozciąga się poza koniec rzeczywistych danych. Jest to prognoza oczekiwanego popytu dla następnych trzech miesięcy na tym rynku.

Linie danych rzeczywistych i prognozowanych w całym szeregu czasowym są bardzo blisko siebie na wykresie, co wskazuje, że jest to rzetelny model dla tego szeregu czasowego.

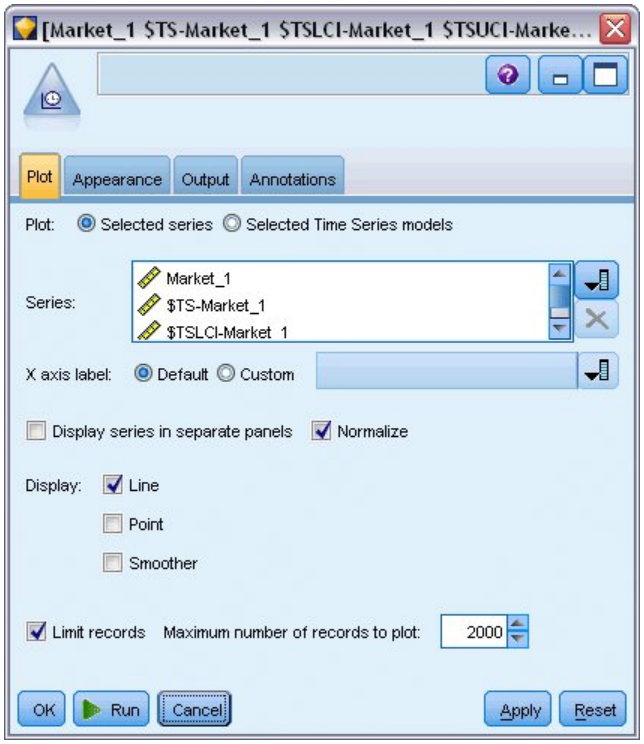


Rysunek 188. Wykres sekwencyjny danych rzeczywistych i prognozowanych dla zmiennej Market_1

Zapisz model w pliku do użycia w przyszłym przykładzie:

9. Kliknij przycisk **OK**, aby zamknąć bieżący wykres.
10. Otwórz model użytkowy Szereg czasowy.
11. Wybierz opcje **Plik > Zapisz węzeł** i określ lokalizację pliku.
12. Kliknij przycisk **Zapisz**.

    Mamy rzetelny model dla tego konkretnego rynku, ale jaką granicę błędu ma prognoza? Można uzyskać pewne informacje o tym, badając przedział ufności.
13. Dwukrotnie kliknij ostatni węzeł wykresu sekwencyjnego w strumieniu (z etykietą **Market_1 $TS-Market_1**), aby ponownie otworzyć jego okno dialogowe.
14. Kliknij przycisk wyboru zmiennych i dodaj zmienne *$TSLCI-Market_1* i *$TSUCI-Market_1* do listy **Szeregi**.
15. Kliknij przycisk **Uruchom**.

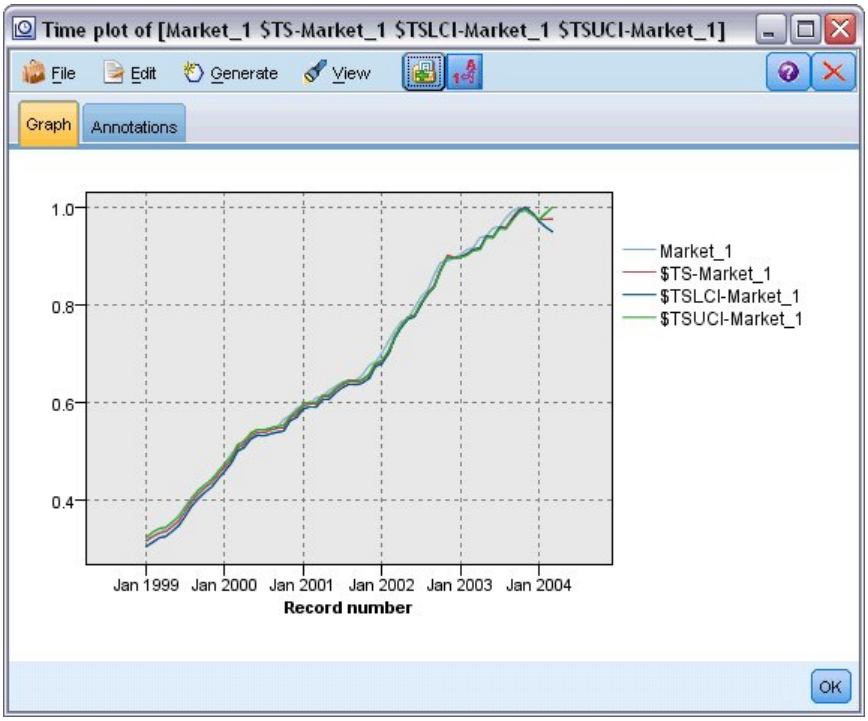


Rysunek 189. Dodawanie kolejnych zmiennych do wykresu

Jest to ten sam wykres, co wcześniej, ale ma dodaną górną (*$TSUCI*) i dolną (*$TSLCI*) granicę przedziału ufności.

Zwróćmy uwagę, jak granice przedziału ufności rozbiegają się w okresie prognozy, wskazując rosnącą niepewność prognozy w przyszłości.

Po upływie każdego okresu użytkownik otrzyma kolejną miesięczną (w tym przypadku) porcję rzeczywistych danych z użycia, na których można oprzeć prognozę. Można wczytać nowe dane do strumienia i zastosować ponownie model, o którym wiadomo, że jest rzetelny. Więcej informacji można znaleźć w temacie “Ponowne stosowanie modelu Szereg czasowy” na stronie 170.



Rysunek 190. Wykres sekwencyjny z dodanym przedziałem ufności

Podsumowanie

Dowiedzieliśmy się, jak używać narzędzia Automatyczny dobór modelu do opracowania prognoz dla wielu szeregów czasowych i użytkownik zapisał wynikowe modele w pliku zewnętrznym.

W następnym przykładzie zobaczymy, jak przekształcić niestandardowe dane szeregów czasowych do formatu dostosowanego do węzła Szereg czasowy.

Ponowne stosowanie modelu Szereg czasowy

Ten przykład stosuje modele szeregów czasowych z pierwszego przykładu, ale może być również używany niezależnie. Więcej informacji można znaleźć w temacie “Prognozowanie za pomocą węzła Szereg czasowy” na stronie 151.

Tak jak w oryginalnym scenariuszu, analityk krajowego dostawcy usług szerokopasmowych ma utworzyć prognozy subskrypcji usług dla każdego z kilku rynków lokalnych, aby przywidzieć wymogi dotyczące przepustowości. Użytkownik użył już narzędzia Automatyczny dobór modelu, aby utworzyć modele i prognozę na trzy miesiące.

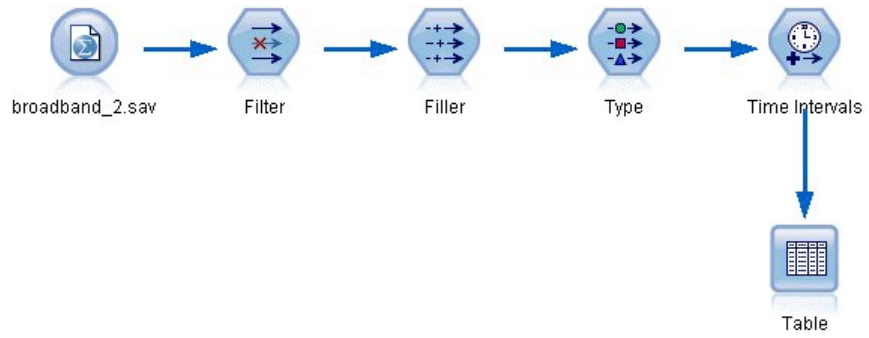
Hurtownia danych została zaktualizowana za pomocą rzeczywistych danych oryginalnego okresu prognozy, więc chcesz użyć tych danych, aby rozszerzyć horyzont prognozy o kolejne trzy miesiące.

W tym przykładzie zastosowano strumień o nazwie *broadband_apply_models.str*, który odwołuje się do pliku danych o nazwie *broadband_2.sav*. Te pliki są dostępne w folderze *Demos* instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *broadband_apply_models.str* znajduje się w folderze *streams*.

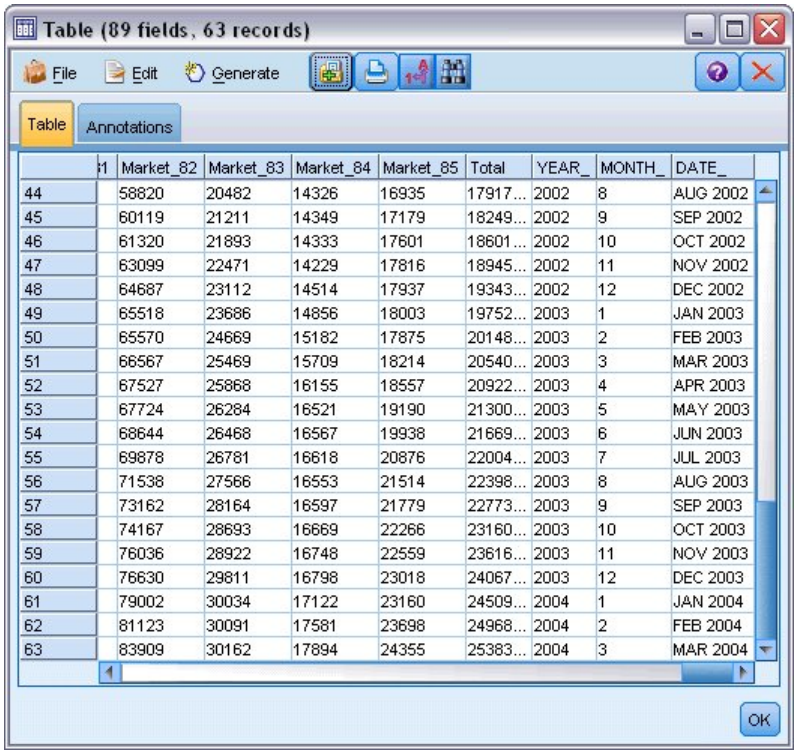
Pobieranie strumienia

W tym przykładzie użytkownik odtworzy węzeł Szereg czasowy z modelu Szereg czasowy zapisanego w pierwszym przykładzie. Jeśli model nie został zapisany, kopię można znaleźć w folderze *Demos*.

1. Otwórz strumień *broadband_apply_models.str* z folderu *streams* katalogu *Demos*.



Rysunek 191. Otwieranie strumienia



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

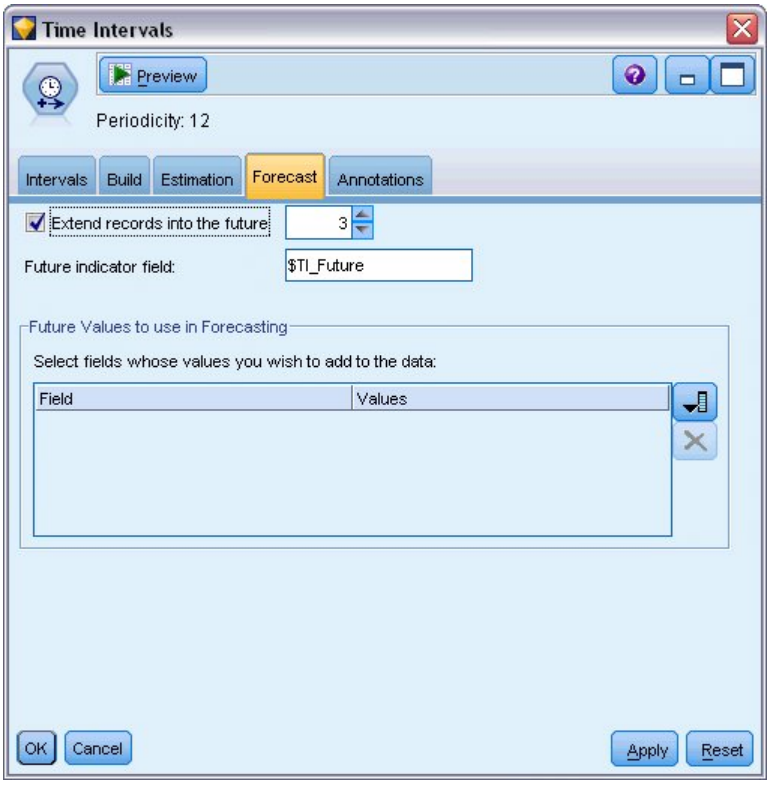
Rysunek 192. Aktualizowane dane sprzedaży

Zaktualizowane dane miesięczne zgromadzono w pliku *broadband_2.sav*.

2. Załącz węzeł tabeli do węzła źródłowego Plik IBM SPSS Statistics, otwórz węzeł tabeli i kliknij przycisk **Uruchom**.

   Uwaga: Plik danych został zaktualizowany rzeczywistymi danymi sprzedaży od stycznia do marca 2004 w wierszach od 61 do 63.

3. Otwórz węzeł Przedziały czasowe w strumieniu.
4. Kliknij kartę **Prognoza**.
5. Upewnij się, że wartość opcji **Rozszerz rekordy na przedziały z przyszłości** to **3**.

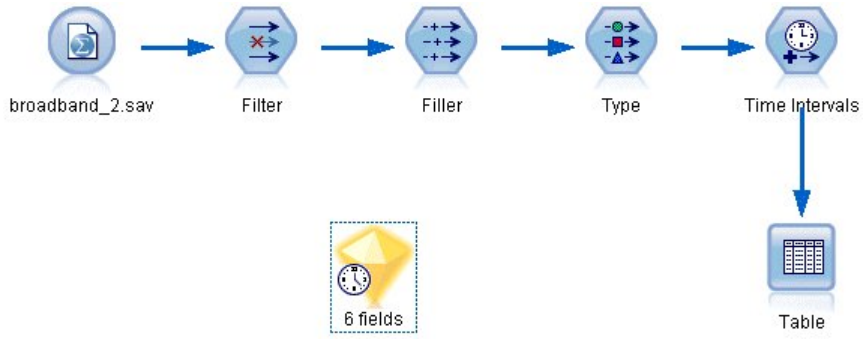


Rysunek 193. Sprawdzanie ustawienia okresu prognozy

Pobieranie zapisanego modelu

1. W menu programu IBM SPSS Modeler wybierz opcje **Wstaw > Węzeł z pliku** i wybierz plik *TSmodel.nod* z folderu *Demos* (lub użyj modelu Szereg czasowy zapisanego w pierwszym przykładzie szeregów czasowych).

Ten plik zawiera modele szeregów czasowych z poprzedniego przykładu. Operacja wstawiania umieszcza odpowiedni model użytkowy Szereg czasowy w obszarze roboczym.

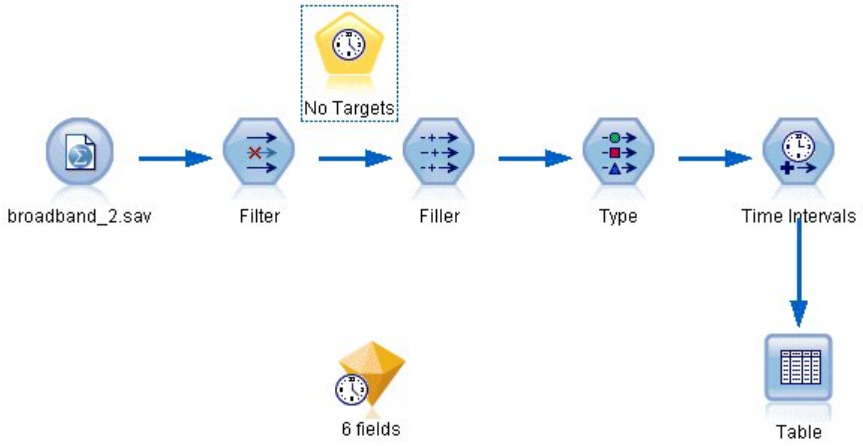


Rysunek 194. Dodawanie modelu użytkowego

Generowanie węzła modelowania

1. Otwórz model użytkowy Szereg czasowy i wybierz opcje **Utwórz > Utwórz węzeł modelowania**.

Powoduje to umieszczenie węzła modelowania szeregów czasowych w obszarze roboczym.

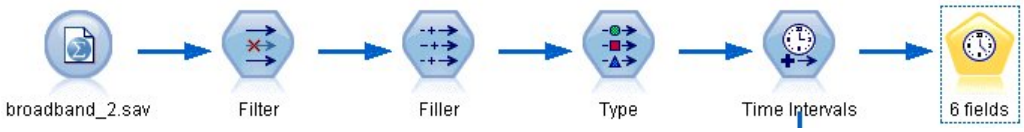


Rysunek 195. Generowanie węzła modelowania z modelu użytkowego

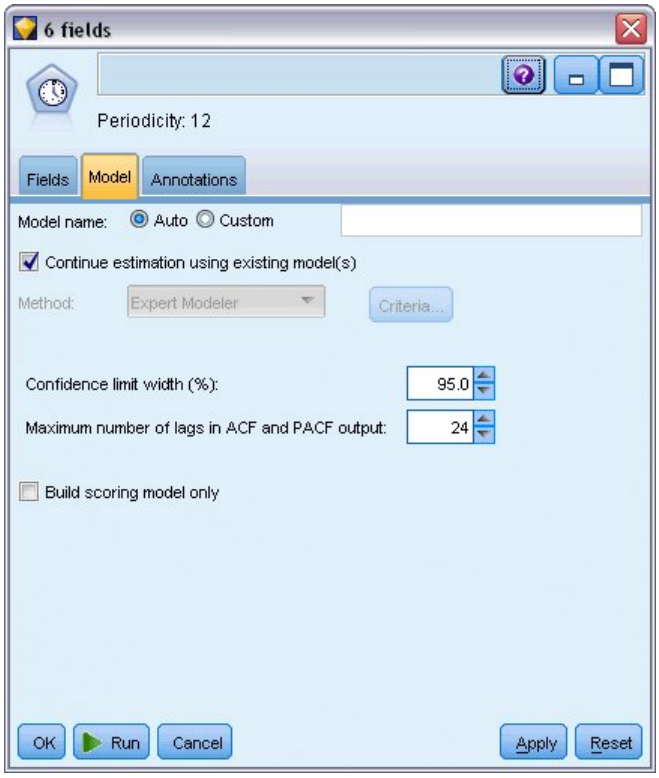
Generowanie nowego modelu

1. Zamknij model użytkowy Szereg czasowy i usuń go z obszaru roboczego.

   Stary model zbudowano na 60 wierszach danych. Teraz należy wygenerować nowy model na podstawie zaktualizowanych danych sprzedaży (63 wiersze).
2. Dołącz nowo wygenerowany węzeł budowania szeregu czasowego do strumienia.



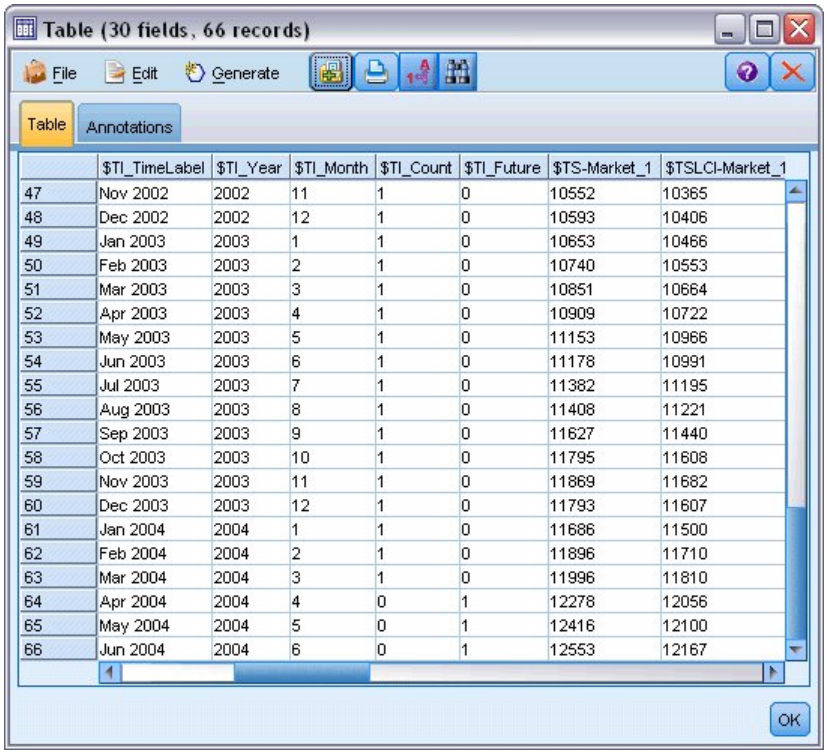
Rysunek 196. Załączanie węzła modelowania do strumienia



Rysunek 197. Ponowne używanie zapisanych ustawień dla modelu szeregów czasowych

3. Otwórz węzeł Szereg czasowy.
4. Na karcie **Model** sprawdź, czy zaznaczona jest opcja **Kontynuuj oszacowanie za pomocą istniejących modeli**.
5. Kliknij przycisk **Uruchom**, aby umieścić nowy model użytkowy w obszarze roboczym i na palecie modeli.

Badanie nowego modelu

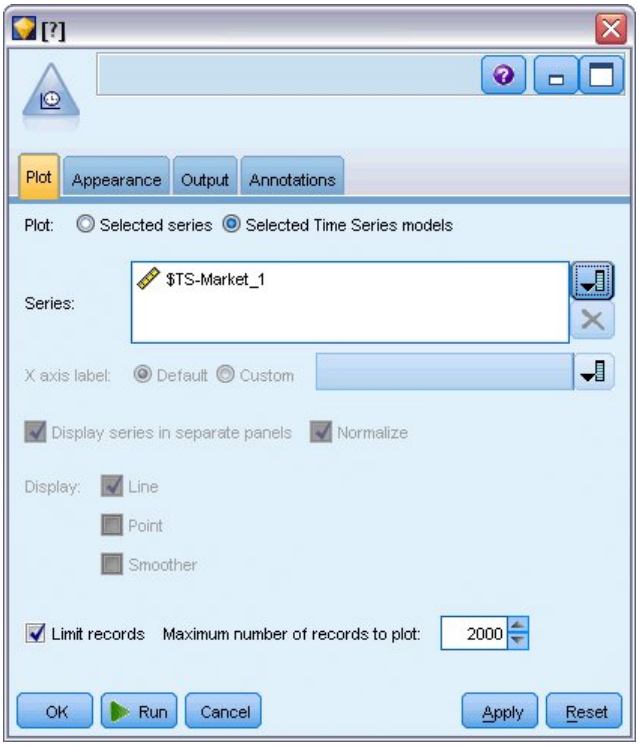


	$TI_TimeLabel	$TI_Year	$TI_Month	$TI_Count	$TI_Future	$TS-Market_1	$TSLCI-Market_1
47	Nov 2002	2002	11	1	0	10552	10365
48	Dec 2002	2002	12	1	0	10593	10406
49	Jan 2003	2003	1	1	0	10653	10466
50	Feb 2003	2003	2	1	0	10740	10553
51	Mar 2003	2003	3	1	0	10851	10664
52	Apr 2003	2003	4	1	0	10909	10722
53	May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

Rysunek 198. Tabela przedstawiająca nową prognozę

1. Załącz węzeł tabeli do nowego modelu użytkowego Szereg czasowy w obszarze roboczym.
2. Otwórz węzeł Tabela i kliknij przycisk **Uruchom**.

   Nowy model wciąż przedstawia prognozę na trzy miesiące, ponieważ używane są zapisane ustawienia. Tym razem prognoza obejmuje jednak okres od kwietnia do czerwca, ponieważ okres estymacji (określony w węźle Przedziały czasowe) kończy się teraz w marcu zamiast w styczniu.



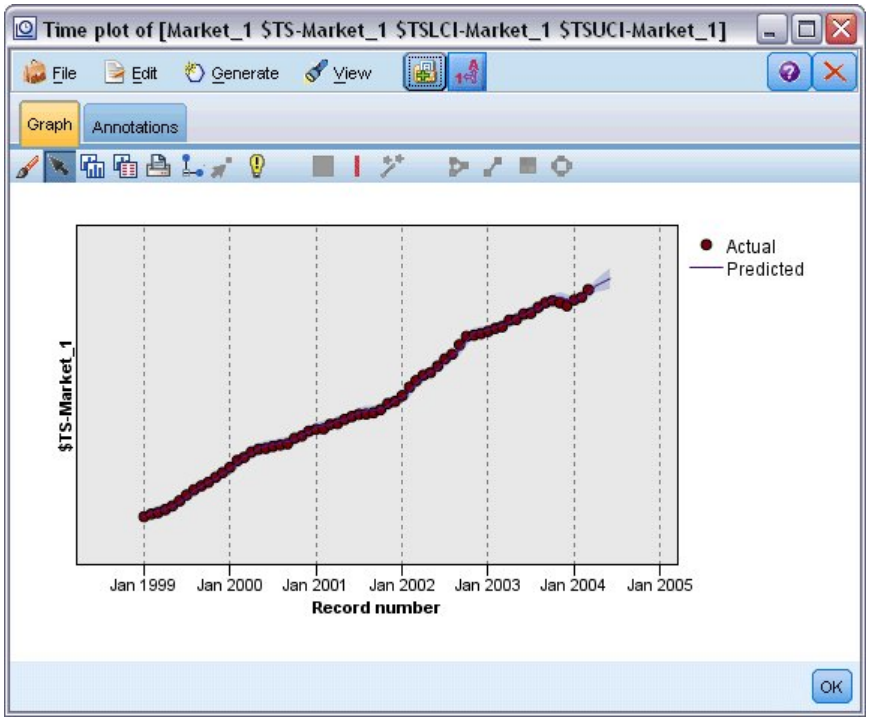
Rysunek 199. Określanie zmiennych do umieszczenia na wykresie

3. Dołącz węzeł wykresu sekwencyjnego do modelu użytkowego Szereg czasowy.

   Tym razem użyjemy okna wykresu sekwencyjnego przeznaczonego specjalnie dla modeli szeregów czasowych.
4. Na karcie Wykres zaznacz opcję **Wybrane modele szeregów czasowych**.
5. Na liście **Szeregi** kliknij przycisk wyboru zmiennych i zaznacz zmienną *$TS-Market_1*, a następnie kliknij przycisk **OK**, aby dodać je do listy.
6. Kliknij przycisk **Uruchom**.

Gotowy jest wykres przedstawiający rzeczywistą sprzedaż dla rynku *Market_1* do marca 2004 r. razem z prognozą sprzedaży oraz przedziałem ufności (określonym niebieskim zacieniowanym obszarem) do czerwca 2004 r.

Tak jak w pierwszym przykładzie wartości prognozy znajdują się blisko rzeczywistych danych w całym okresie, ponownie potwierdzając, że model jest dobry.



Rysunek 200. Prognoza rozszerzona do czerwca

Podsumowanie

Dowiedzieliśmy się, jak zastosować zapisane modele, aby rozszerzyć poprzednie prognozy, gdy dostępne są bardziej aktualne dane, i zrobiliśmy to bez przebudowywania modeli. Oczywiście, jeśli istnieje powód sugerujący, że model zmienił się, należy go przebudować.

Rozdział 15. Prognozowanie sprzedaży katalogowej (Szereg czasowy)

Firma sprzedaży katalogowej jest zainteresowana prognozowaniem miesięcznej sprzedaży męskiej linii odzieżowej w oparciu o dane sprzedaży z ostatnich 10 lat.

W tym przykładzie zastosowano strumień o nazwie *catalog_forecast.str*, który odwołuje się do pliku danych o nazwie *catalog_seasfac.sav*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *catalog_forecast.str* znajduje się w katalogu *streams*.

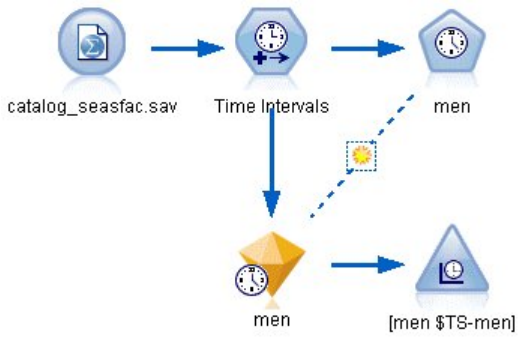
W poprzednim przykładzie użyto narzędzia Automatyczny dobór modelu, aby zadecydować, jaki jest najbardziej odpowiedni model dla szeregów czasowych. Teraz przyjrzymy się bliżej dwóm metodom, które są dostępne podczas samodzielnego wyboru modeli: wygładzanie wykładnicze i ARIMA.

Aby pomóc w podjęciu decyzji o wyborze odpowiedniego modelu, dobrze jest najpierw wykonać wykres szeregu czasowego. Wizualna inspekcja szeregu czasowego może być skuteczną pomocą w dokonaniu wyboru. Należy zadać sobie w szczególności następujące pytania:

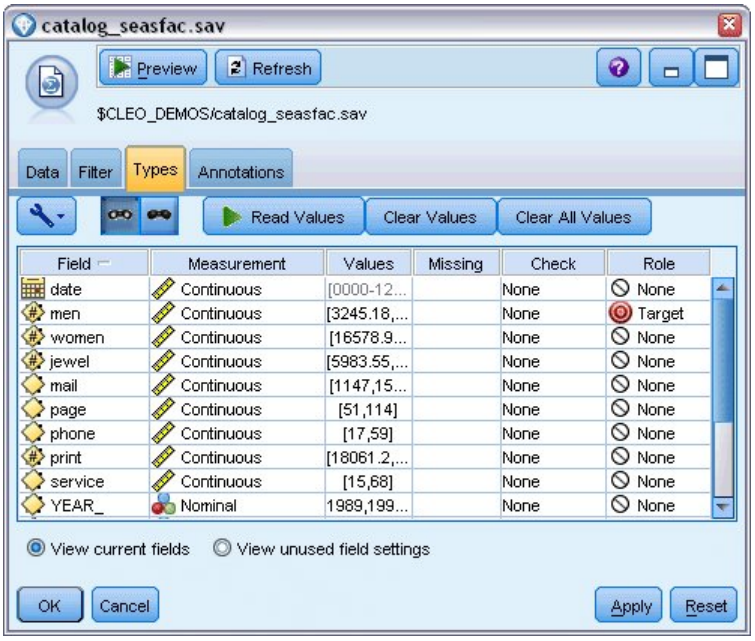
- Czy ten szereg ma ogólny trend? Jeśli tak, to czy trend jest stały, czy maleje z czasem?
- Czy szereg wykazuje sezonowość? Jeśli tak, czy fluktuacje sezonowe rosną z czasem, czy wydają się stałe w kolejnych okresach?

Tworzenie strumienia

1. Utwórz nowy strumień i dodaj węzeł źródłowy Plik Statistics wskazujący na plik *catalog_seasfac.sav*.

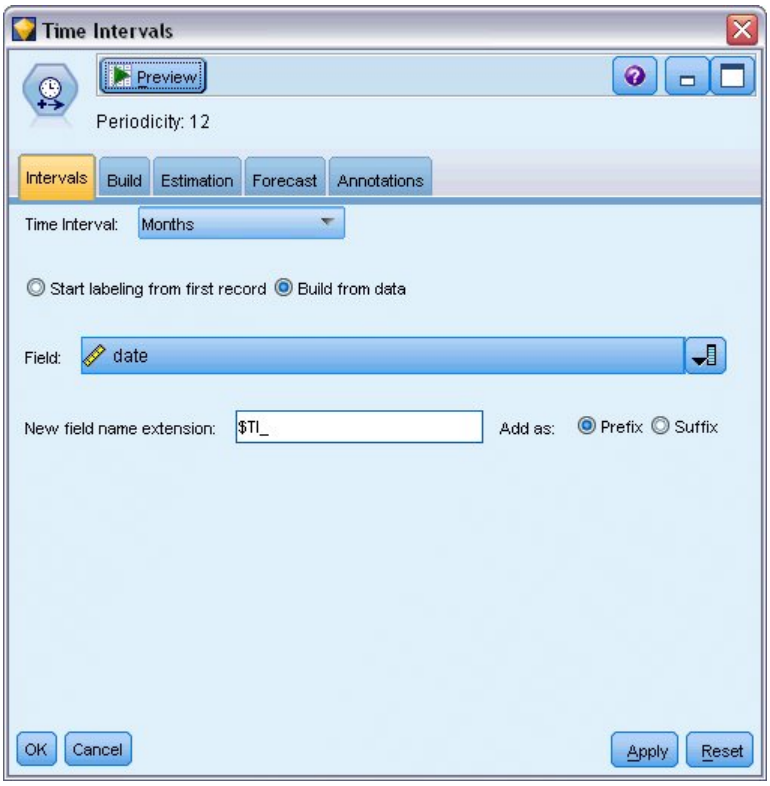


Rysunek 201. Prognozowanie sprzedaży katalogowej



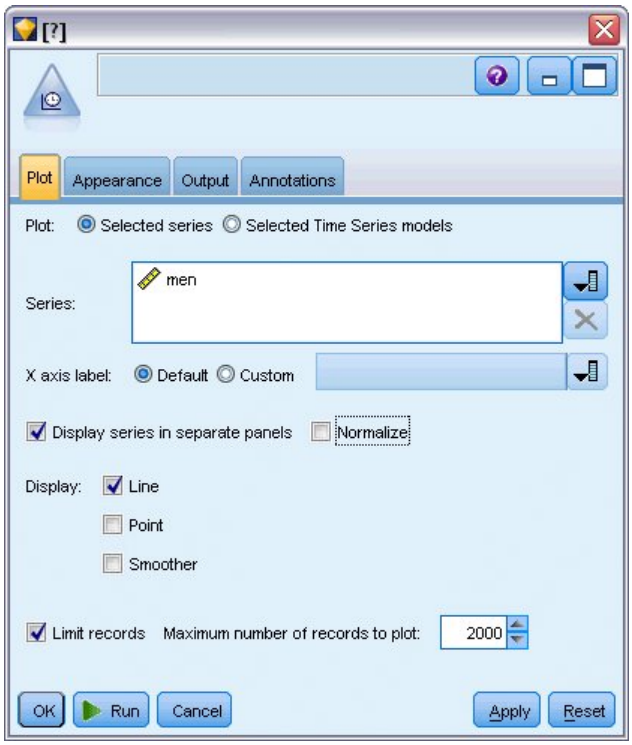
Rysunek 202. Określanie zmiennej przewidywanej

2. Otwórz węzeł źródłowy Plik IBM SPSS Statistics i wybierz kartę Typy.
3. Kliknij przycisk **Odczytaj wartości**, a następnie **OK**.
4. Kliknij kolumnę *Rola* dla zmiennej *men* i ustaw rolę na **Przewidywana**.
5. Ustaw rolę dla wszystkich pozostałych zmiennych na **Brak** i kliknij przycisk **OK**.



Rysunek 203. Ustawianie przedziałów czasowych

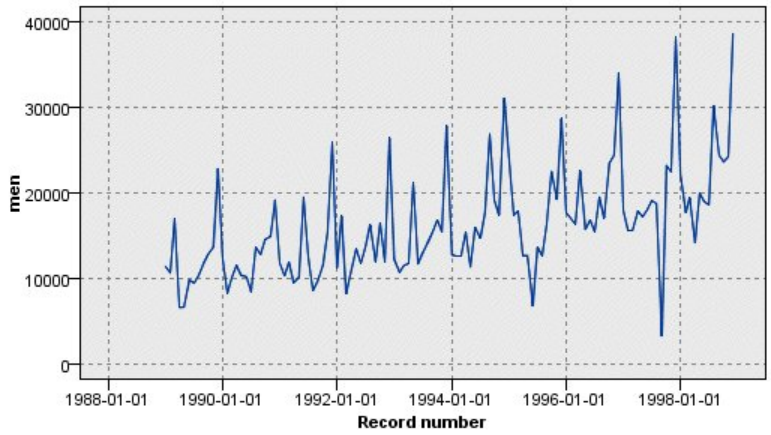
6. Dołącz węzeł Przedziały czasowe do węzła źródłowego Plik IBM SPSS Statistics.
7. Otwórz węzeł Przedziały czasowe i ustaw opcję **Przedział czasowy** na **Miesiące**.
8. Zaznacz opcję **Utwórz na podstawie danych**.
9. Ustaw opcję **Zmienna** na **date** i kliknij przycisk **OK**.



Rysunek 204. Tworzenie wykresów szeregów czasowych

10. Załącz węzeł wykresu sekwencyjnego do węzła Przedziały czasowe.
11. Na karcie Wykres dodaj zmienną **men** do listy Szeregi.
12. Usuń zaznaczenie pola wyboru **Normalizuj**.
13. Kliknij przycisk **Uruchom**.

Badanie danych



Rysunek 205. Rzeczywista sprzedaży odzieży męskiej

Szereg wykazuje ogólny trend w górę. Wartości szeregu ogólnie rosną w czasie. Trend rosnący wydaje się stały, co wskazywałoby na trend liniowy.

Szereg ma również wyraźny wzorzec sezonowy z corocznym maksimum w grudniu, na co wskazują pionowe linie na wykresie. Zmienność sezonowa wydaje się rosnąć razem z trendem rosnącym, co wskazywałoby na sezonowość multiplikatywną, a nie addytywną.

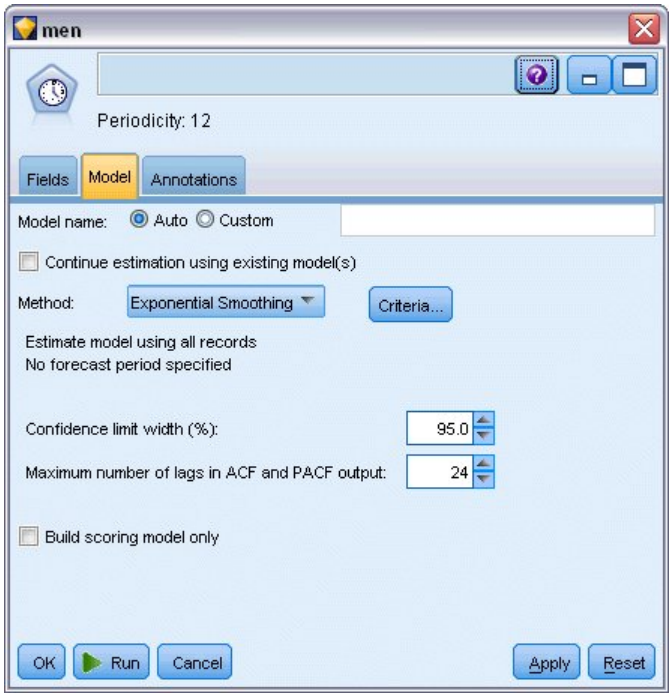
1. Kliknij przycisk **OK**, aby zamknąć wykres.

Teraz, gdy zidentyfikowano cechy szeregu, można go modelować. Metoda wygładzania wykładniczego jest szczególnie przydatna w przypadku prognozowania szeregów wykazujących trend, sezonowość lub obie te cechy. Jak stwierdziliśmy, dane wykazują obie cechy.

Wygładzanie wykładnicze

Budowanie najlepiej dopasowanego modelu wygładzania wykładniczego obejmuje określenie typu modelu — czy model musi obejmować trend, sezonowość, czy obie te cechy — a następnie uzyskanie parametrów najlepszego dopasowania dla wybranego modelu.

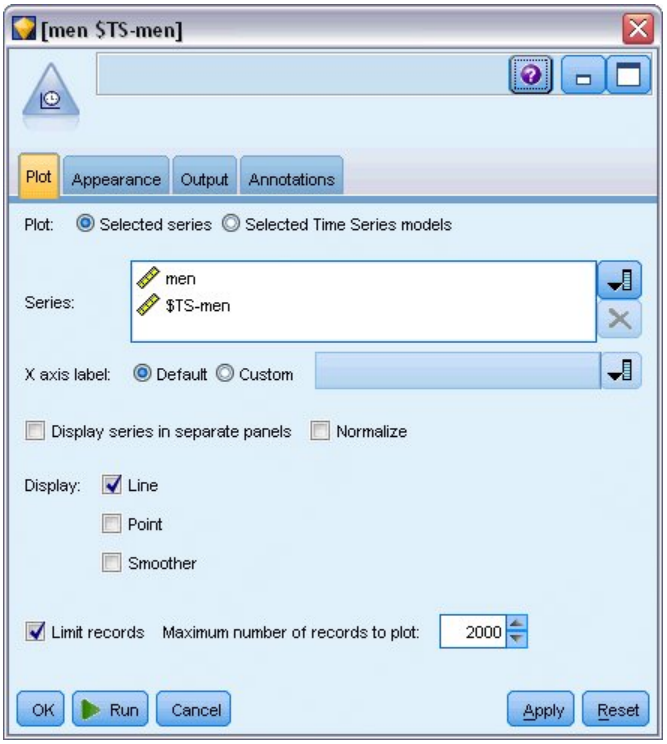
Wykres sprzedaży odzieży męskiej sugeruje model z liniowym składnikiem trendu i multiplikatywnym składnikiem sezonowości. Sugeruje to model Wintersa. Najpierw zbadamy jednak prosty model (bez trendu i sezonowości), a następnie model Holta (zawiera trend liniowy, ale nie uwzględnia sezonowości). Zapewni to użytkownikowi praktykę w identyfikowaniu, kiedy model nie jest dobrze dopasowany do danych, co jest niezbędną umiejętnością w skutecznym budowaniu modelu.



Rysunek 206. Określanie wygładzania wykładniczego

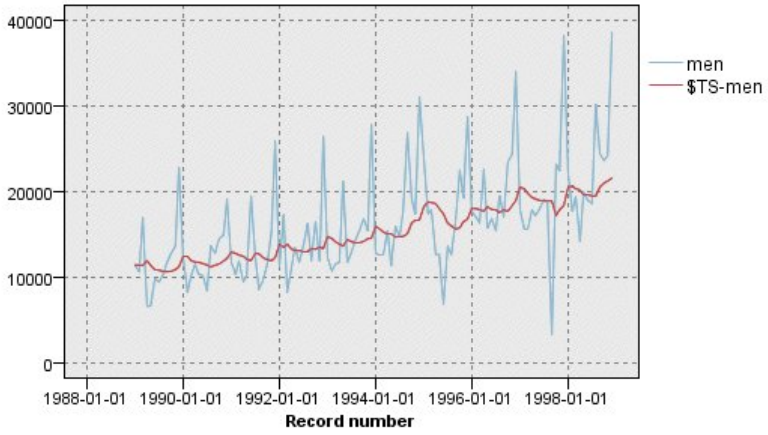
Rozpoczniemy od prostego modelu wygładzania wykładniczego.

1. Załącz węzeł Szereg czasowy do węzła Przedziały czasowe.
2. Na karcie **Model** ustaw opcję **Metoda** na **Wygładzanie wykładnicze**.
3. Kliknij przycisk **Uruchom**, aby utworzyć model użytkowy.



Rysunek 207. Tworzenie wykresu modelu Szereg czasowy

4. Dołącz węzeł wykresu sekwencyjnego do modelu użytkowego.
5. Na karcie **Wykres** dodaj zmienne *men* i *$TS-men* do listy **Szeregi**.
6. Usuń zaznaczenie pól **Wyświetl szeregi w oddzielnych panelach** i **Normalizuj**.
7. Kliknij przycisk **Uruchom**.

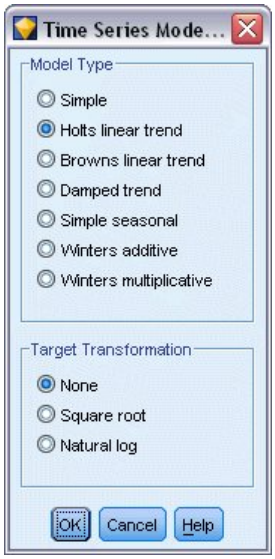


Rysunek 208. Prosty model wygładzania wykładniczego

Wykres **men** reprezentuje rzeczywiste dane, podczas gdy **$TS-men** to model szeregów czasowych.

Mimo że prosty model wykazuje stopniowy (i raczej powolny) trend rosnący, nie uwzględnia sezonowości. Bez obaw można odrzucić ten model.

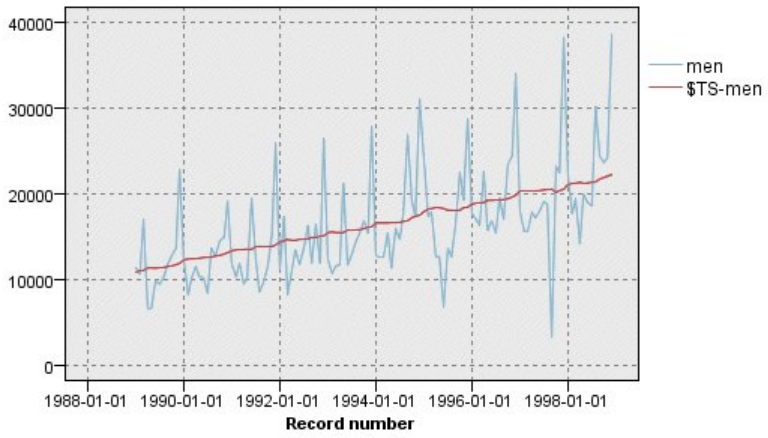
8. Kliknij przycisk **OK**, aby zamknąć okno wykresu sekwencyjnego.



Rysunek 209. Wybieranie modelu Holta

Wypróbujmy model liniowy Holta. Ten model powinien lepiej uchwycić trend niż prosty model, ale mało prawdopodobne jest, że odzwierciedli sezonowość.

9. Otwórz ponownie węzeł wykresu sekwencyjnego.
10. Na karcie **Model** przy wybranej metodzie **Wygładzanie wykładnicze** kliknij opcję **Kryteria**.
11. W oknie dialogowym Kryteria wygładzania wykładniczego wybierz opcję **Trend liniowy Holta**.
12. Kliknij przycisk **OK**, aby zamknąć okno dialogowe.
13. Kliknij przycisk **Uruchom**, aby utworzyć ponownie model użytkowy.
14. Otwórz ponownie węzeł wykresu sekwencyjnego i kliknij przycisk **Uruchom**.

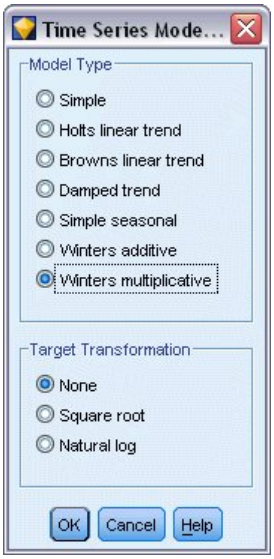


Rysunek 210. Model trendu liniowego Holta

Model Holta wykazuje gładszy trend rosnący niż prosty model, ale wciąż nie uwzględnia sezonowości, więc można odrzucić również ten model.

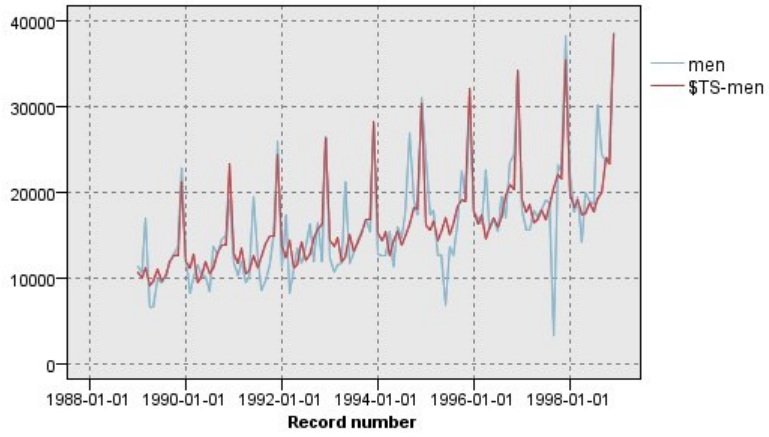
15. Zamknij okno wykresu sekwencyjnego.

Wstępny wykres sprzedaży odzieży męskiej w czasie sugerował model obejmujący trend liniowy i sezonowość multiplikatywną.Dlatego też bardziej odpowiednim kandydatem jest model Wintersa.



Rysunek 211. Wybieranie modelu Wintersa

16. Otwórz ponownie węzeł wykresu sekwencyjnego.
17. Na karcie **Model** przy wybranej metodzie **Wygładzanie wykładnicze** kliknij opcję **Kryteria**.
18. W oknie dialogowym Kryteria wygładzania wykładniczego wybierz opcję **Multiplikatywny Wintersa**.
19. Kliknij przycisk **OK**, aby zamknąć okno dialogowe.
20. Kliknij przycisk **Uruchom**, aby utworzyć ponownie model użytkowy.
21. Otwórz węzeł wykresu sekwencyjnego i kliknij przycisk **Uruchom**.



Rysunek 212. Model multiplikatywny Wintersa

Ten wykres wygląda lepiej — model odzwierciedla zarówno trend, jak i sezonowość danych.

Zbiór danych obejmuje okres 10 lat i 10 sezonowych szczytów występujących w grudniu każdego roku. 10 szczytów obecnych w przewidywanych wynikach dobrze pasuje do 10 corocznych szczytów danych rzeczywistych.

Wyniki uwydatniają jednak ograniczenia procedury wygładzania wykładniczego. Nagłe skoki w górę i w dół pokazują, że istnieje istotna struktura, która nie została uwzględniona.

Jeśli podstawowym celem jest modelowanie długoterminowego trendu ze zmiennością sezonową, to wygładzanie wykładnicze może być dobrym wyborem. Aby utworzyć model bardziej złożonej struktury, takiej jak ta, należy rozważyć użycie procedury ARIMA.

ARIMA

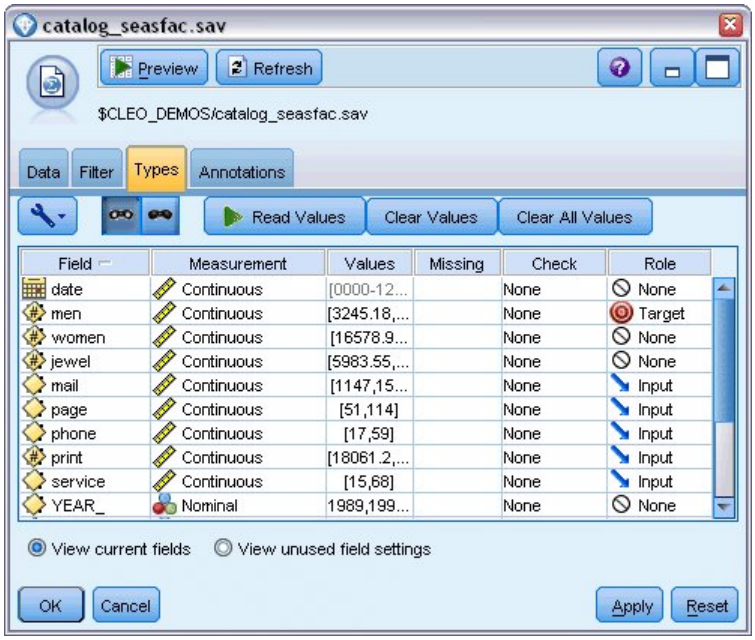
Procedura ARIMA pozwala na utworzenie modelu autoregresyjnej zintegrowanej średniej ruchomej (ARIMA) umożliwiającego dokładne modelowanie szeregów czasowych. Modele ARIMA oferują bardziej wyrafinowane metody modelowania składników trendu i sezonowości niż modele wygładzania wykładniczego oraz mają dodatkową zaletę polegającą na uwzględnianiu w modelu zmiennych predyktorów.

Kontynuując przykład firmy sprzedaży katalogowej, która chce opracować model prognozowania, wiedzieliśmy, że firma zgromadziła dane dotyczące miesięcznej sprzedaży odzieży męskiej razem z kilkoma szeregami, których można użyć do wyjaśnienia niektórych zmienności w sprzedaży. Potencjalne predyktory obejmują liczbę wysłanych katalogów, liczbę stron w katalogu, liczbę otwartych linii telefonicznych, kwotę wydaną na reklamy w prasie oraz liczbę pracowników działu obsługi klienta.

Czy niektóre z tych predyktorów są przydatne do prognozy? Czy model z predyktorami jest naprawdę lepszy niż model bez predyktorów? Używając procedury ARIMA, możemy utworzyć model prognozowania z predyktorami i zobaczyć, czy wystąpi istotna różnica zdolności predykcyjnej w porównaniu do modelu wygładzania wykładniczego bez predyktorów.

Metoda ARIMA pozwala na dostosowanie modelu, określając kolejność autoregresji, różnicowania i średniej kroczącej, jak również sezonowych odpowiedników tych składników. Ręczne określanie najlepszych wartości dla tych składników może być czasochłonnym procesem obejmującym pracę metodą prób i błędów, więc w tym przykładzie pozwolimy, aby narzędzie Automatyczny dobór modelu wybrało za nas model ARIMA.

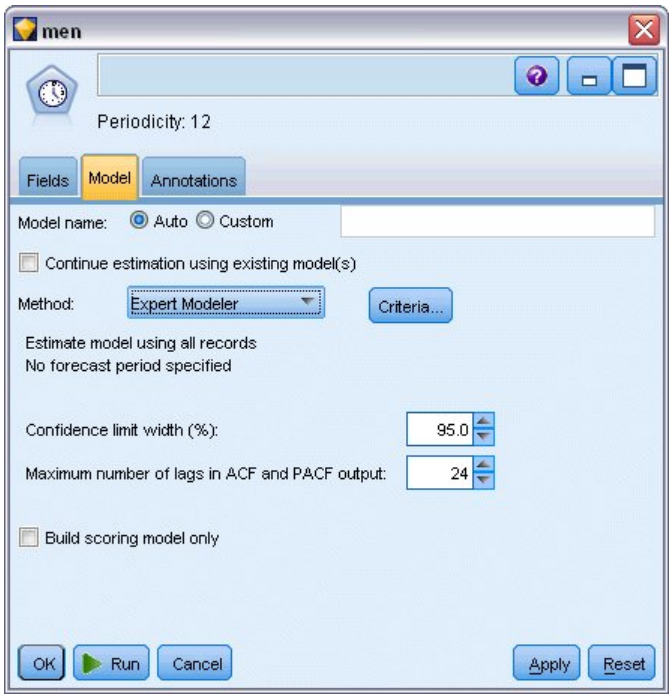
Spróbujemy utworzyć lepszy model, traktując niektóre zmienne w zbiorze danych jako zmienne predyktorów. Zmienne, które wydają się najbardziej przydatne, to liczba wysłanych katalogów (*mail*), liczba stron w katalogu (*page*), liczba linii telefonicznych otwartych do zamówień (*phone*), kwota wydana na reklamy w prasie *(print)* oraz liczba pracowników działu obsługi klienta *(service)*.



Rysunek 213. Ustawianie zmiennych predykcyjnych

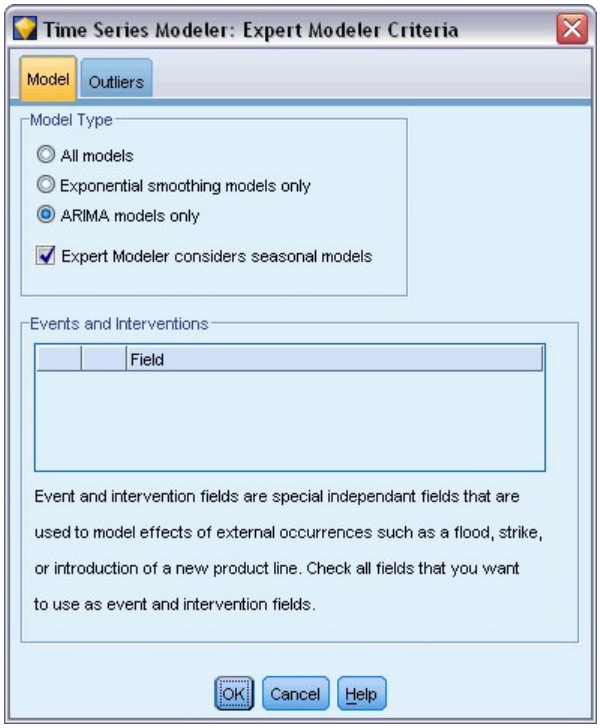
1. Otwórz węzeł źródłowy Plik IBM SPSS Statistics.

2. Na karcie Typy ustaw wartość pola *Rola* dla zmiennych *mail*, *page*, *phone*, *print* i *service* na **Dane wejściowe**.
3. Upewnij się, że rola zmiennej **men** jest ustawiona jako **Przewidywana** i wszystkie pozostałe zmienne są ustawione na **Brak**.
4. Kliknij przycisk **OK**.



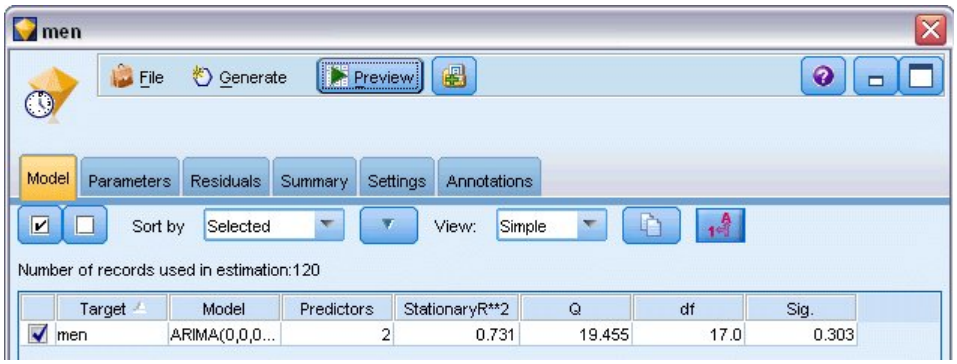
Rysunek 214. Wybieranie opcji Automatyczny dobór modelu

5. Otwórz węzeł Szereg czasowy.
6. Na karcie Model ustaw opcje **Metoda** na **Automatyczny dobór modelu** i kliknij przycisk **Kryteria**.



Rysunek 215. Wybieranie tylko modeli ARIMA

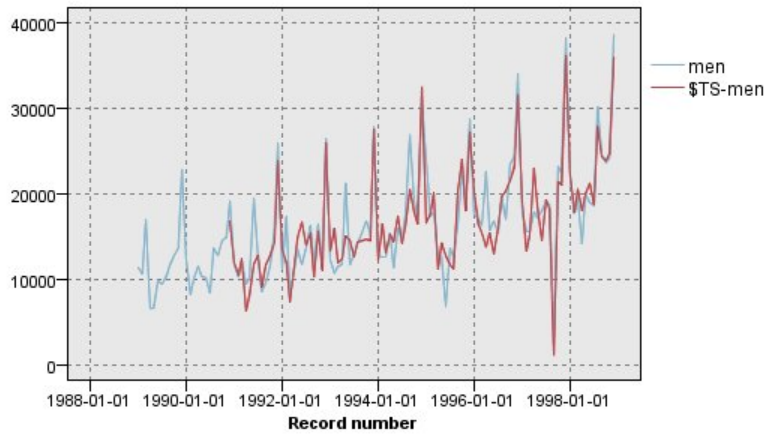
7. W oknie dialogowym Kryteria doboru automatycznego wybierz opcję **Tylko modele ARIMA** i upewnij się, że zaznaczona jest opcja **Automatyczny dobór modelu uwzględnia modele sezonowe**.
8. Kliknij przycisk **OK**, aby zamknąć okno dialogowe.
9. Kliknij przycisk **Uruchom** na karcie Model, aby ponownie utworzyć model użytkowy.



Rysunek 216. Automatyczny dobór modelu wybiera dwa predyktory

10. Otwórz model użytkowy.

    Zauważ, że opcja Automatyczny dobór modelu wybrała tylko dwa z pięciu określonych predyktorów jako istotne dla modelu.
11. Kliknij przycisk **OK**, aby zamknąć model użytkowy.
12. Otwórz węzeł wykresu sekwencyjnego i kliknij przycisk **Uruchom**.



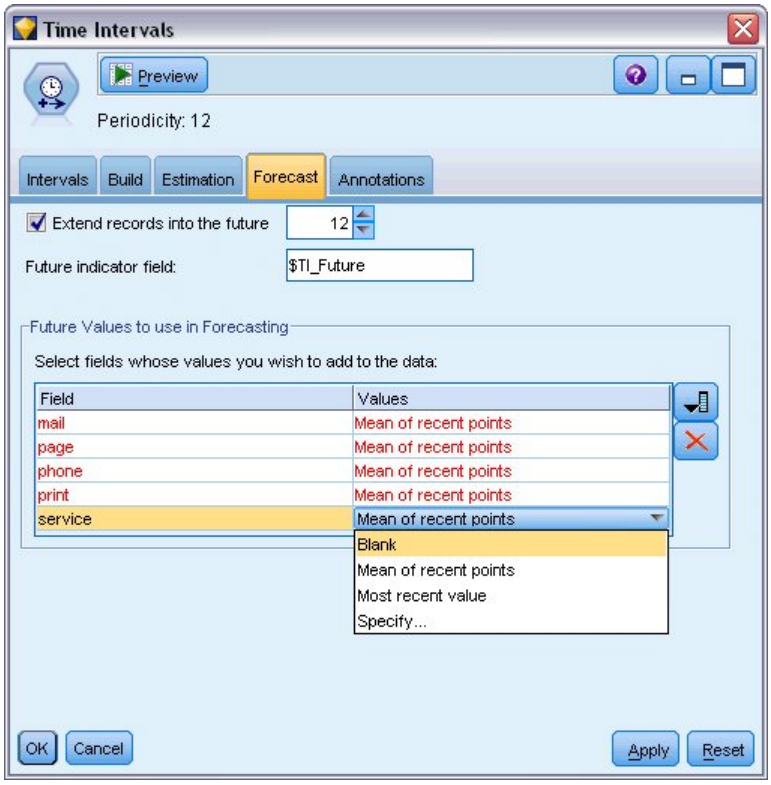
Rysunek 217. Model ARIMA z określonymi predyktorami

Model jest usprawnieniem poprzedniego, ponieważ odwzorowuje duży spadek, co sprawia, że jest najlepiej dopasowanym modelem do tej pory.

Można by dostosować model jeszcze bardziej, ale wszelkie usprawnienia od tego miejsca będą prawdopodobnie minimalne. Ustaliliśmy, że preferowany jest model ARIMA z predyktorami, użyjmy więc zbudowanego właśnie modelu. Do celów tego przykładu przygotujemy prognozę dla nadchodzącego roku.

13. Kliknij przycisk **OK**, aby zamknąć okno wykresu sekwencyjnego.
14. Otwórz węzeł Przedziały czasowe i wybierz kartę *Prognoza*.
15. Zaznacz pole wyboru *Rozszerz rekordy na przedziały z przyszłości* i ustaw jego wartość na 12.

    Użycie predyktorów podczas przewidywania wymaga określenia szacowanych wartości dla tych zmiennych w okresie prognozy, aby program mógł dokładniej przewidzieć zmienną przewidywaną.

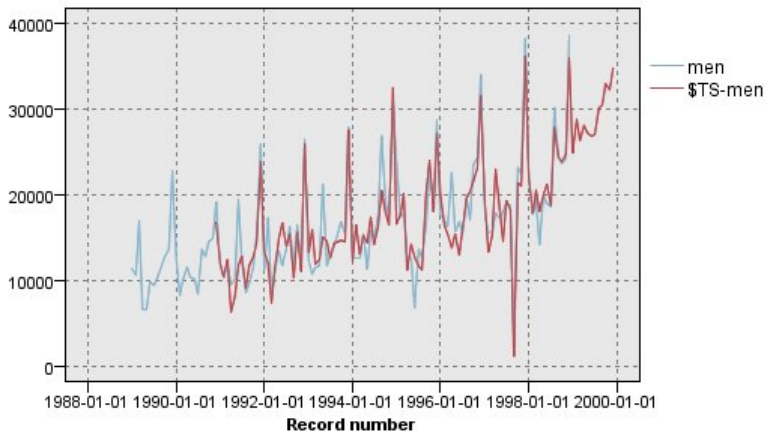


Rysunek 218. Określanie przyszłych wartości zmiennych predykcyjnych

16. W grupie **Wartości przyszłe używane w prognozowaniu** kliknij przycisk wyboru zmiennych po prawej stronie kolumny Wartości.
17. W oknie dialogowym Wybierz zmienne zaznacz zmienne od **mail** do **service** i kliknij przycisk **OK**.

    W rzeczywistej sytuacji użytkownik określiłby wartości ręcznie, ponieważ te pięć predyktorów jest powiązanych z elementami kontrolowanymi przez użytkownika. Dla celów tego przykładu użyjemy jednej z predefiniowanych funkcji, aby nie musieć określać 12 wartości dla każdego predyktora. (Po zaznajomieniu się z tym przykładem można spróbować eksperymentować z różnymi przyszłymi wartościami, aby zobaczyć, jaki mają wpływ na model).
18. Dla każdej zmiennej po kolei kliknij pole **Wartości**, aby wyświetlić listę możliwych wartości, i wybierz pozycję **Średnia z ostatnich punktów**. Ta opcja oblicza średnią z trzech ostatnich punktów danych dla tej zmiennej i używa jej jako szacowanej wartości w każdym przypadku.
19. Kliknij przycisk **OK**.
20. Otwórz węzeł Szereg czasowy i kliknij przycisk **Uruchom**, aby ponownie utworzyć model użytkowy.
21. Otwórz węzeł wykresu sekwencyjnego i kliknij przycisk **Uruchom**.

Prognoza dla roku 1999 wygląda dobrze — zgodnie z oczekiwaniami po grudniu następuje powrót do normalnej sprzedaży, w drugiej połowie roku występuje trend rosnący z ogólną sprzedażą powyżej wartości z poprzedniego roku.



Rysunek 219. Prognoza sprzedaży z określonymi predyktorami

Podsumowanie

Pomyślnie utworzyliśmy model skomplikowanego szeregu czasowego, uwzględniającego nie tylko trend rosnący, ale również zmienności sezonowe i inne. Zobaczyliśmy, jak metodą prób i błędów można dochodzić do dokładnego modelu, którego można następnie użyć do przewidywania przyszłej sprzedaży.

W praktyce należałoby ponownie zastosować model razem z aktualizacją rzeczywistych danych sprzedaży (na przykład co miesiąc lub co kwartał) i przygotować aktualizowaną prognozę. Więcej informacji można znaleźć w temacie “Ponowne stosowanie modelu Szereg czasowy” na stronie 170.

Rozdział 16. Składanie ofert klientom (Samonauczanie)

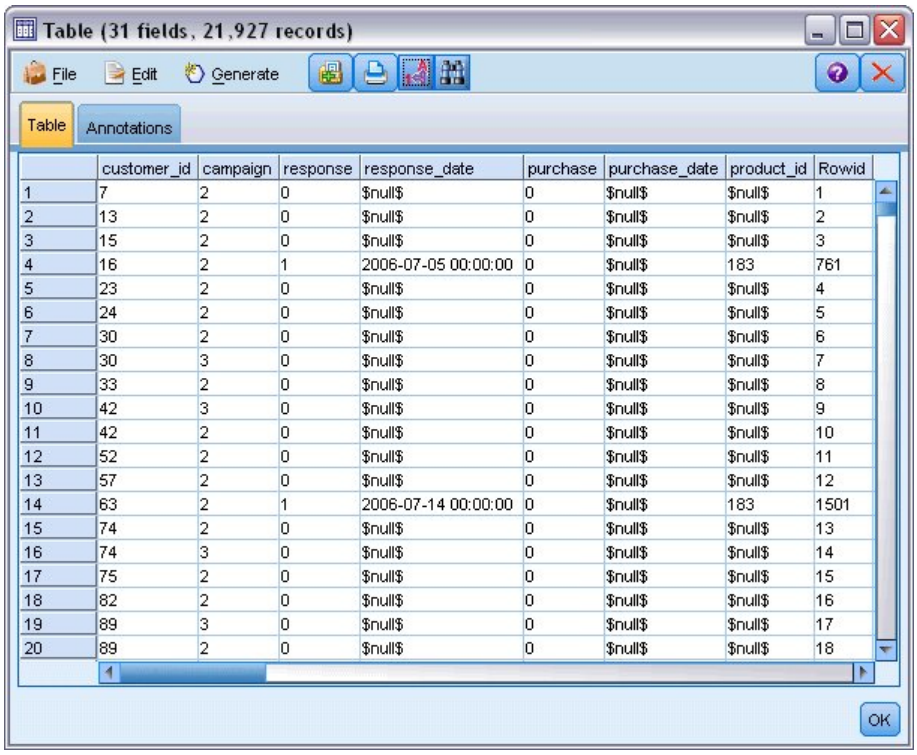
Węzeł samouczącego się modelu SLRM generuje model i umożliwiaj jego aktualizowanie, co pozwala przewidzieć, które oferty są najbardziej odpowiednie dla klientów, i określić, jakie jest prawdopodobieństwo przyjęcia oferty. Takie modele są najbardziej użyteczne w zarządzaniu relacjami z klientami, takimi jak zastosowania marketingowe lub centra obsługi.

Ten przykład wykorzystuje fikcyjny bank. Dział marketingu chce przyczynić się do zwiększenia zysków, organizując kampanię, w której oferty finansowe będą precyzyjnie dopasowane do charakterystyki poszczególnych klientów. Przykład używa samouczącego się modelu SLRM, aby określić cechy klientów, którzy z największym prawdopodobieństwem pozytywnie zareagują na ofertę, na podstawie wcześniejszych ofert i odpowiedzi w celu promowania najlepszej obecnej oferty na podstawie wyników.

Ten przykład używa strumienia *pm_selflearn.str*, który odwołuje się do plików danych *pm_customer_train1.sav*, *pm_customer_train2.sav* i *pm_customer_train3.sav*. Te pliki są dostępne w folderze *Demos* instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *pm_selflearn.str* znajduje się w katalogu *streams*.

Istniejące dane

Firma posiada dane historyczne śledzące oferty przedstawione klientom w poprzednich kampaniach, razem z odpowiedziami na te oferty. Te dane obejmują również informacje demograficzne i finansowe, których można użyć do przewidywania wskaźników odpowiedzi dla różnych klientów.

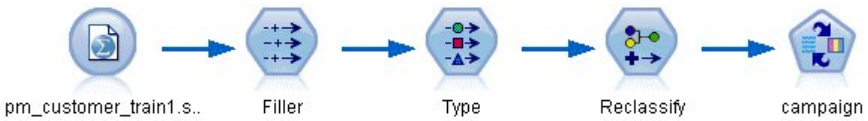


	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7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

Rysunek 220. Odpowiedzi na poprzednie oferty

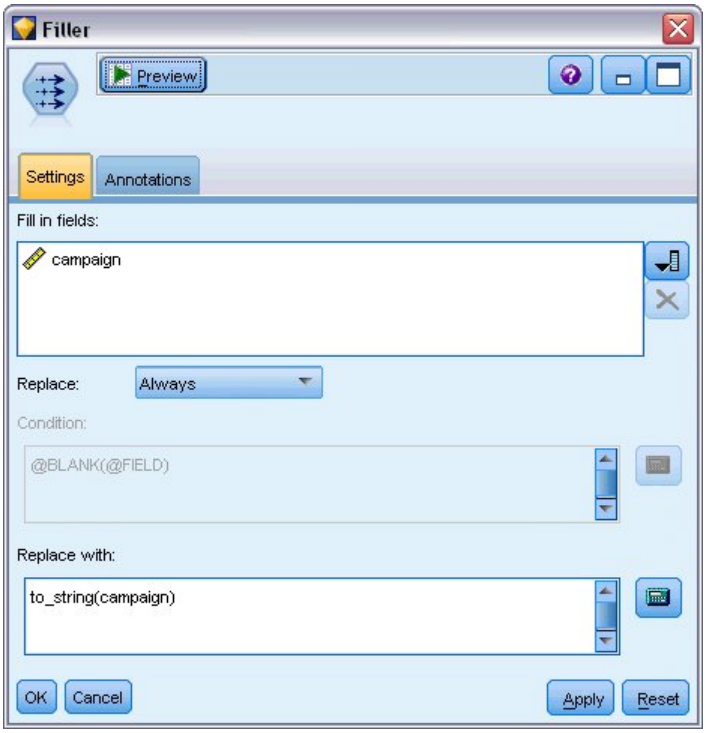
Tworzenie strumienia

1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *pm_customer_train1.sav* znajdujący się w folderze *Demos* w folderze instalacji IBM SPSS Modeler.



Rysunek 221. Przykładowy strumień SLRM

2. Dodaj węzeł wypełniania i wybierz campaign jako wypełnianą zmienną.
3. Dla opcji Zamień wybierz typ **Zawsze**.
4. W polu tekstowym Zamień na wprowadź łańcuch to_string(campaign) i kliknij przycisk **OK**.



Rysunek 222. Wyliczanie zmiennej campaign

5. Dodaj węzeł typu i w kolumnie *Rola* ustaw wartość na **Brak** dla zmiennych *customer_id*, *response_date*, *purchase_date*, *product_id*, *Rowid* i *X_random*.

Rysunek 223. Zmienianie ustawień węzła typu

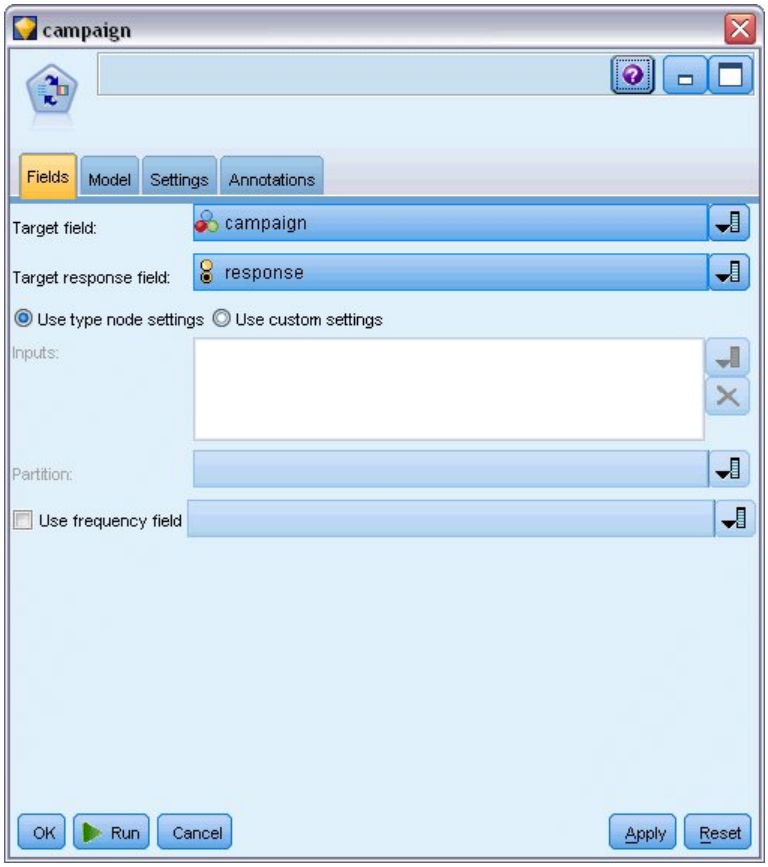
6. Ustaw pole *Rola* na **Przewidywana** dla zmiennych *campaign* i *response*. To są zmienne, na których mają opierać się predykcje.

   Ustaw pole **Poziom pomiaru** na **Flaga** dla zmiennej *response*.
7. Kliknij przycisk **Odczytaj wartości**, a następnie **OK**.

   Ponieważ dane zmiennej campaign są widoczne jako lista numerów (1, 2, 3 i 4), można rekodować zmienne, aby miały bardziej znaczące tytuły.
8. Dołącz węzeł rekodowania do węzła typu.
9. W polu **Rekoduj na** wybierz opcję **Istniejąca zmienna**.
10. Na liście **Rekoduj zmienną** wybierz pozycję **campaign**.
11. Kliknij przycisk **Uzyskaj**. Wartości zmiennej campaign zostaną dodane do kolumny *Wartość oryginalna*.
12. W kolumnie *Nowa wartość* wprowadź następujące nazwy kampanii w pierwszych czterech wierszach:
    - **Mortgage**
    - **Car loan**
    - **Savings**
    - **Pension**
13. Kliknij przycisk **OK**.

Rysunek 224. Rekodowanie nazw kampanii

14. Załącz węzeł modelowania SLRM do węzła rekodowania. Na karcie Zmienne wybierz zmienną **campaign** w pozycji Wartość przewidywana i **response** w pozycji Przewidywana zmienna odpowiedzi.

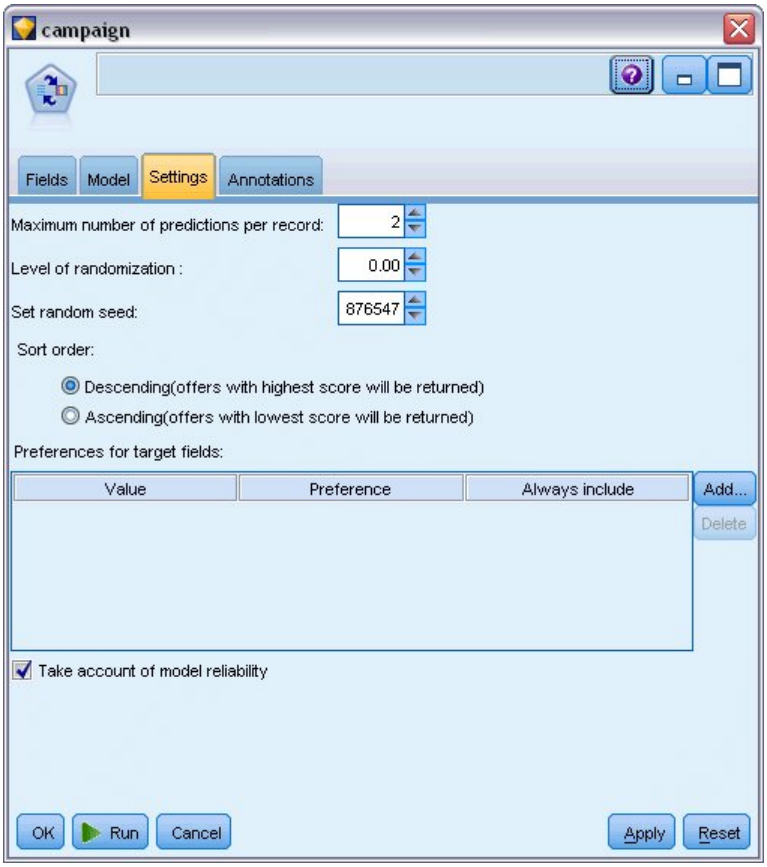


Rysunek 225. Wybieranie zmiennej przewidywanej i przewidywanej zmiennej odpowiedzi

15. Na karcie Ustawienia w polu Maksymalna liczba predykcji na rekord zmniejsz liczbę do 2.

    Oznacza to, że dla każdego klienta zidentyfikowane zostaną dwie oferty z najwyższym prawdopodobieństwem akceptacji.

16. Upewnij się, że zaznaczona jest opcja **Weź pod uwagę niezawodność modelu**, i kliknij przycisk **Uruchom**.

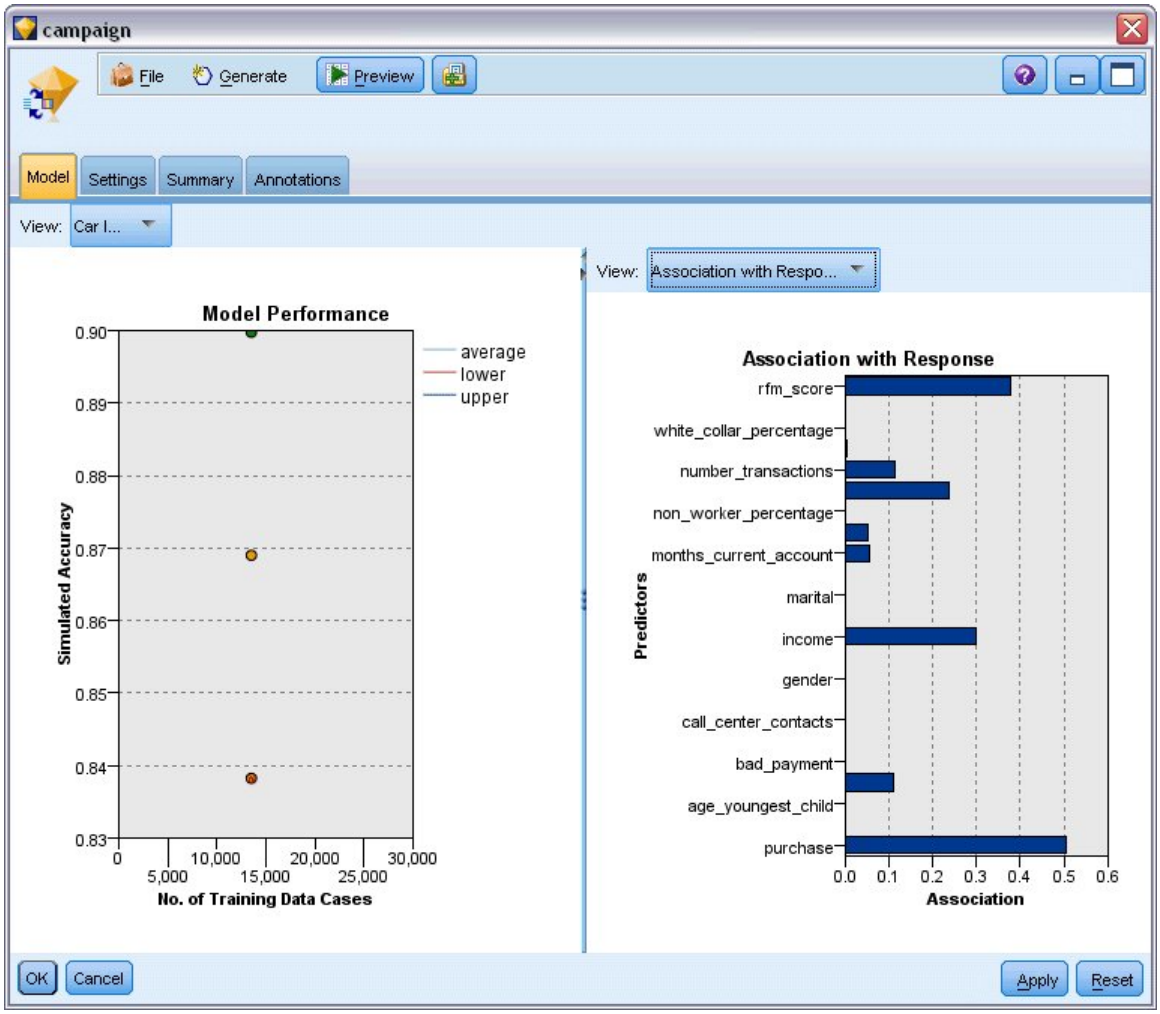


Rysunek 226. Ustawienia węzła SLRM

Przeglądanie modelu

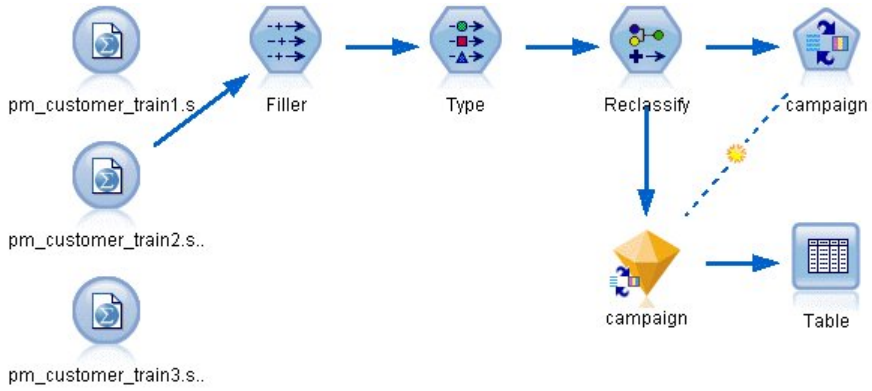
1. Otwórz model użytkowy. Karta Model wstępnie przedstawia szacowaną dokładność predykcji dla każdej oferty oraz powiązanej ważności każdego predyktora w oszacowaniu modelu.

   Aby wyświetlić korelację każdego predyktora ze zmienną przewidywaną, wybierz opcję **Związek z odpowiedzią** z listy **Widok** w oknie po prawej stronie.
2. Aby przełączać pomiędzy czterema ofertami, dla których istnieją predykcje, wybierz ofertę z listy **Widok** w oknie po lewej stronie.



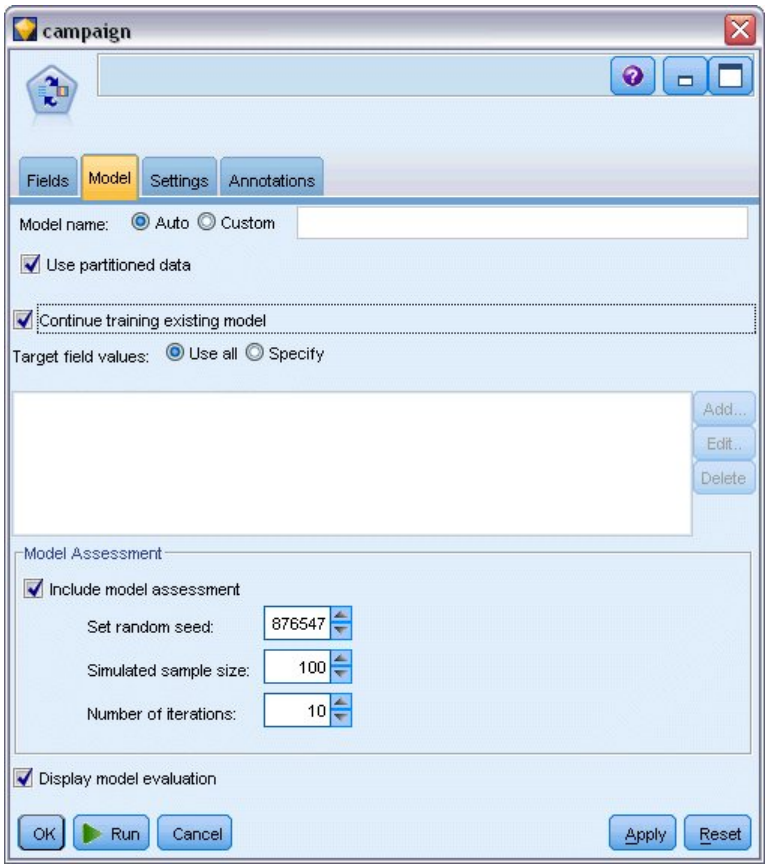
Rysunek 227. Model użytkowy SLRM

3. Zamknij okno modelu użytkowego.
4. W obszarze roboczym strumienia odłącz węzeł źródłowy Plik IBM SPSS Statistics wskazujący na plik *pm_customer_train1.sav*.
5. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *pm_customer_train2.sav* znajdujący się w folderze *Demos* w folderze instalacji IBM SPSS Modeler i połącz go z węzłem wypełniania.



Rysunek 228. Dołączanie drugiego źródła danych do strumienia SLRM

6. Na karcie Model węzła SLRM zaznacz opcję **Kontynuuj uczenie istniejącego modelu**.

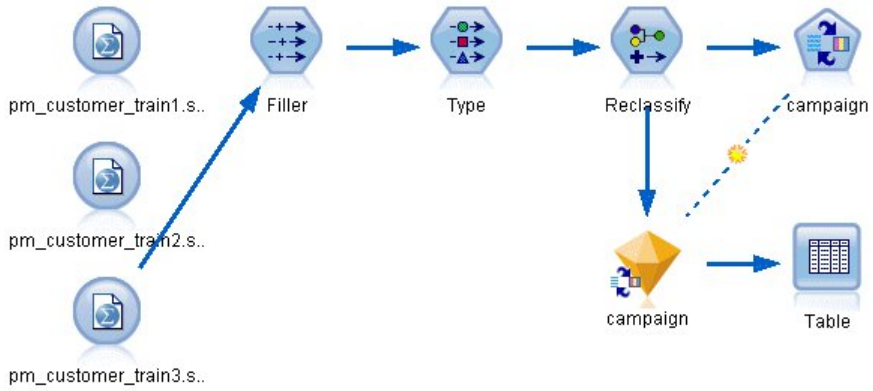


Rysunek 229. Kontynuowanie uczenia modelu

7. Kliknij przycisk **Uruchom**, aby utworzyć ponownie model użytkowy. Aby wyświetlić jego szczegóły, dwukrotnie kliknij model użytkowy w obszarze roboczym.

   Karta Model przedstawia teraz poprawione oszacowania dokładności predykcji dla każdej oferty.

8. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *pm_customer_train3.sav* znajdujący się w folderze *Demos* instalacji programu IBM SPSS Modeler i połącz go z węzłem wypełniania.

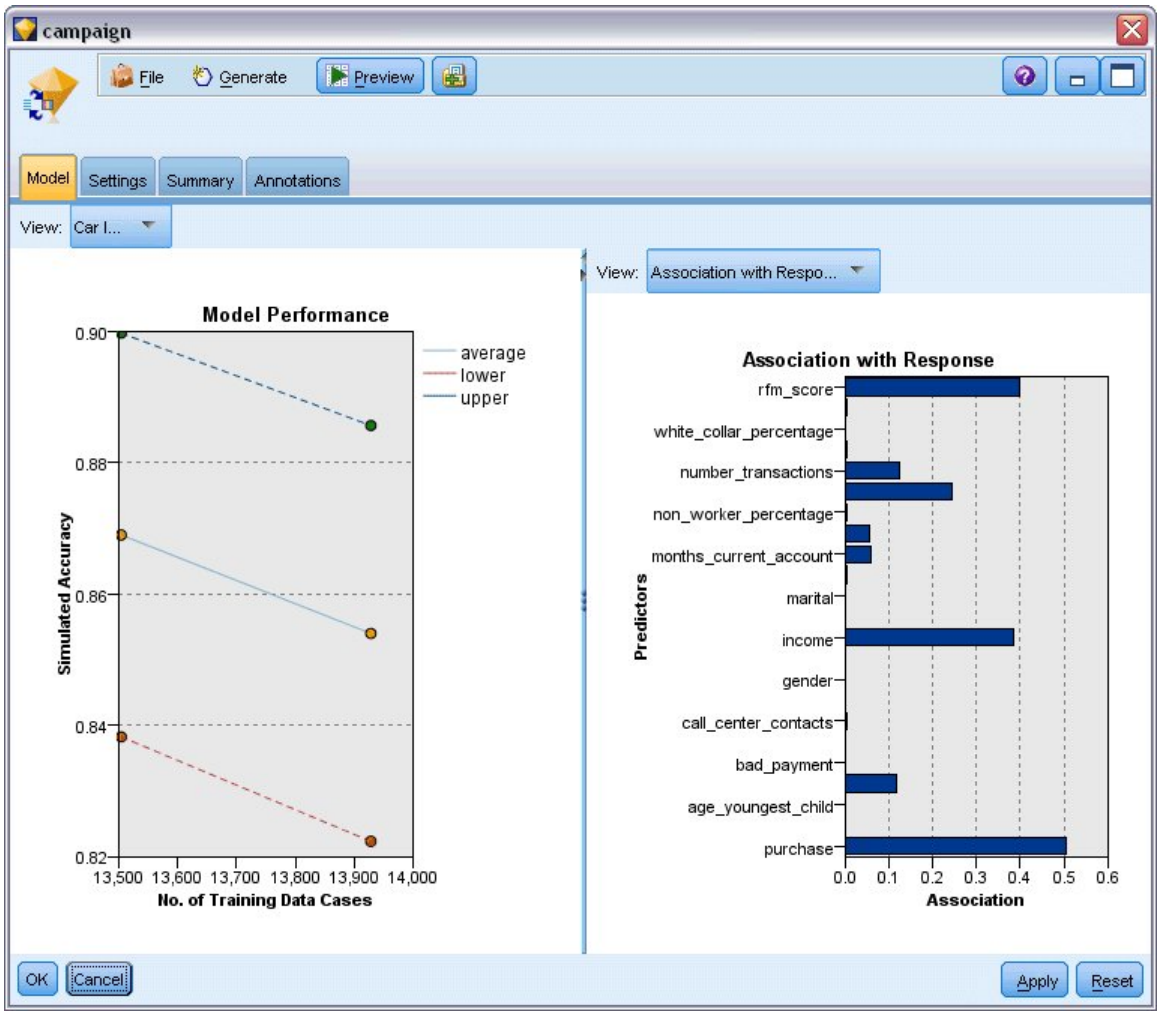


Rysunek 230. Dołączanie trzeciego źródła danych do strumienia SLRM

9. Kliknij przycisk **Uruchom**, aby jeszcze raz utworzyć model użytkowy. Aby wyświetlić jego szczegóły, dwukrotnie kliknij model użytkowy w obszarze roboczym.

10. Karta Model przedstawia teraz końcową szacowaną dokładność predykcji dla każdej oferty.

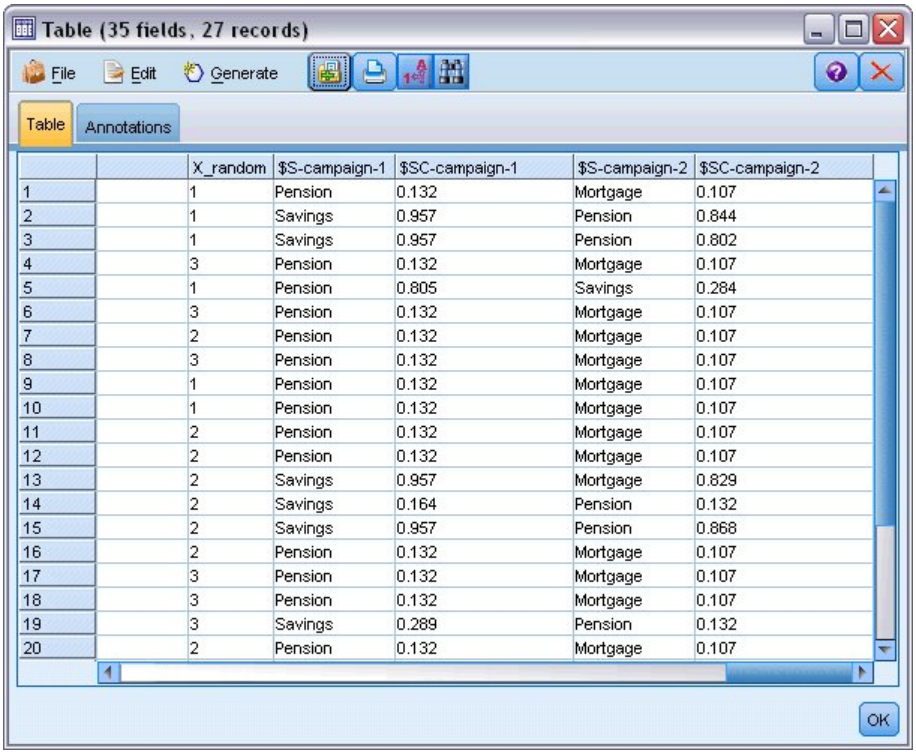
    Tak jak widać, średnia dokładność nieznacznie spadła (z 86,9% do 85,4%), gdy dodano dodatkowe źródła danych. Taka fluktuacja ma minimalną wartość i można ją przypisać lekkim anomaliom w dostępnych danych.



Rysunek 231. Aktualizowany model użytkowy SLRM

11. Załącz węzeł tabeli do ostatniego (trzeciego) wygenerowanego modelu i wykonaj węzeł tabeli.
12. Przewiń tabelę w prawo. Predykcje pokazują oferty, które klient zaakceptuje z największym prawdopodobieństwem, oraz ufność tego, że zaakceptuje ofertę, w zależności od szczegółowych informacji o kliencie.

Na przykład: w pierwszym wierszu przedstawionej tabeli istnieje tylko 13,2% ufności (zapisanej jako 0,132 w kolumnie *$SC-campaign-1*) , że klient, który wcześniej wziął kredyt na samochód, zaakceptuje fundusz emerytalny, jeśli taka oferta zostanie mu zaoferowana. W drugim i trzecim wierszu również przedstawiono klientów, którzy skorzystali z kredytu na samochód. W ich przypadku istnieje 95,7% ufności, że oni oraz inni klienci z podobną historią otworzyliby konto oszczędnościowe, jeśli przedstawiono by im taką ofertę, i ponad 80% ufności, że zaakceptowaliby ofertę funduszu emerytalnego.



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

Rysunek 232. Wyniki modelu — przewidywane oferty i wartości ufności

Wyjaśnienia podstaw matematycznych metod modelowania używanych w programie IBM SPSS Modeler przedstawiono w publikacji *IBM SPSS Modeler Algorithms Guide* dostępnej jako plik PDF w pobranym produkcie.

Należy zauważyć, że wyniki opierają się tylko na danych uczących. Aby ocenić, jak dobrze model uogólnia inne dane w rzeczywistych przypadkach, należałoby użyć węzła podziału na podzbiory, aby zatrzymać podzbiór rekordów do celów testowania i walidacji.

Rozdział 17. Przewidywanie osób niespłacających kredytu (Sieć Bayesa)

Sieci bayesowskie umożliwiają utworzenie modelu prawdopodobieństwa przez połączenie zaobserwowanych i zarejestrowanych dowodów ze „zdroworozsądkową” wiedzą rzeczywistą w celu ustanowienia prawdopodobieństwa występowania zdarzeń na podstawie pozornie niepowiązanych ze sobą atrybutów.

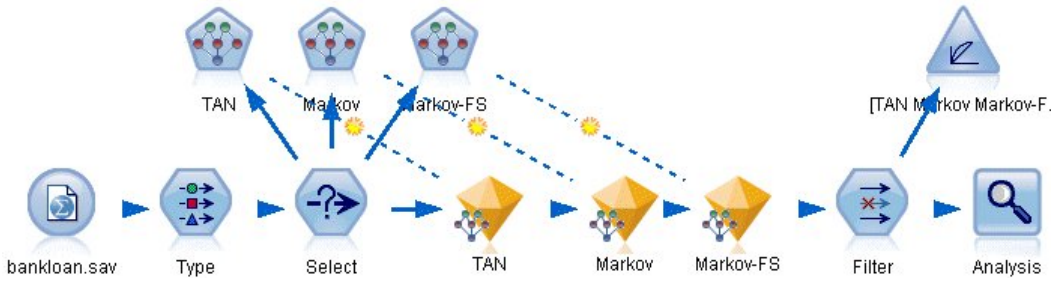
W tym przykładzie zastosowano strumień o nazwie *bayes_bankloan.str*, który odwołuje się do pliku danych o nazwie *bankloan.sav*. Pliki są dostępne w katalogu *Demos* instalacji programu IBM SPSS Modeler i można do nich uzyskać dostęp z grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *bayes_bankloan.str* znajduje się w katalogu *streams*.

W tym przykładzie załóżmy, że bank interesuje się potencjalnym ryzykiem tego, że niektóre kredyty nie zostaną spłacone. Jeśli można będzie użyć danych o wcześniejszych niespłaconych kredytach do przewidzenia tego, którzy klienci mogą mieć problemy ze spłatą kredytów, to takim „ryzykownym” osobom można odmówić pożyczki lub zaoferować alternatywne produkty.

Ten przykład koncentruje się na użyciu istniejących danych o niespłaconych kredytach, aby przewidzieć potencjalne osoby, które nie spłacą kredytu w przyszłości. Przykład analizuje trzy różne typy modeli sieci bayesowskiej, aby określić, która jest lepsza przy przewidywaniu takiej sytuacji.

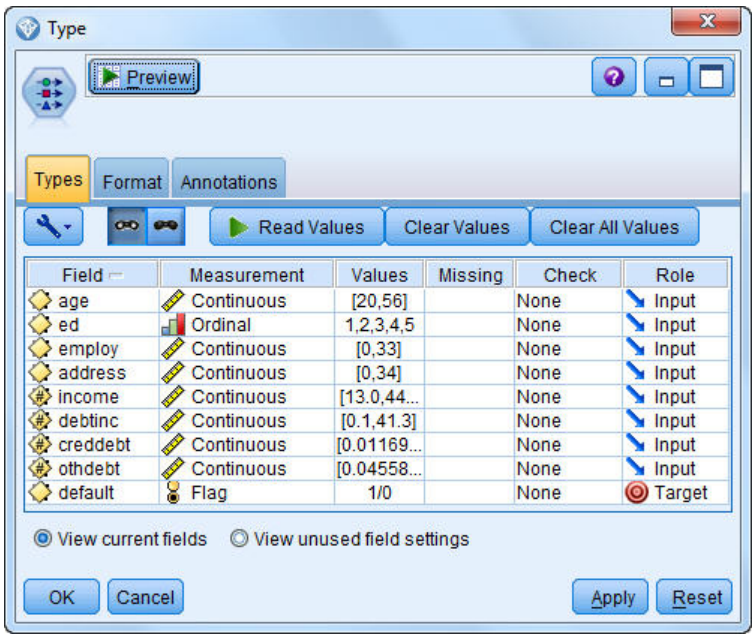
Tworzenie strumienia

1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *bankloan.sav* znajdujący się w folderze *Demos*.



Rysunek 233. Przykładowy strumień Sieć Bayesa

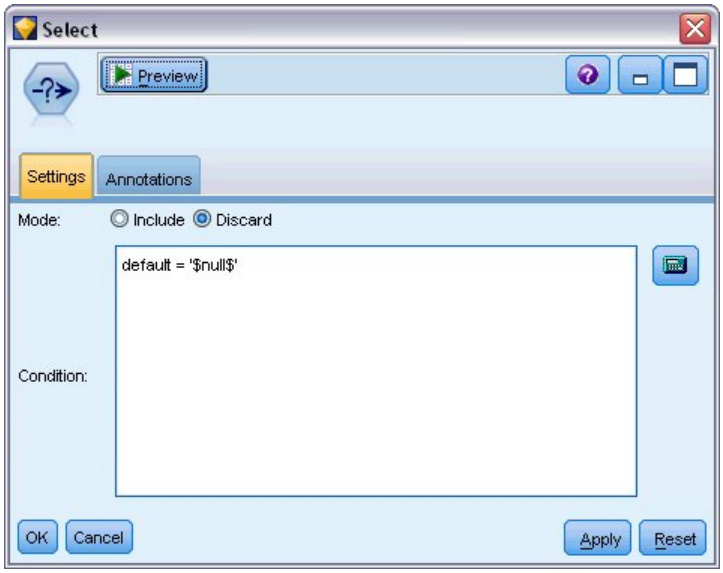
2. Dodaj węzeł typu do węzła źródłowego i ustaw rolę dla zmiennej **default** na **Przewidywana**. Wszystkie pozostałe zmienne powinny mieć ustawioną rolę na **Dane wejściowe**.
3. Kliknij przycisk **Odczytaj wartości**, aby wypełnić kolumnę *Wartości*.



Rysunek 234. Wybieranie zmiennej przewidywanej

Przypadki, gdzie zmienna przewidywana ma wartość null, nie są przydatne przy budowaniu modelu. Można wyłączyć te przypadki, aby uniemożliwić ich użycie w ocenie modelu.

4. Dołącz węzeł selekcji do węzła typu.
5. Dla opcji Tryb wybierz **Odrzuć**.
6. W polu Warunek wpisz **default = '$null$'**.

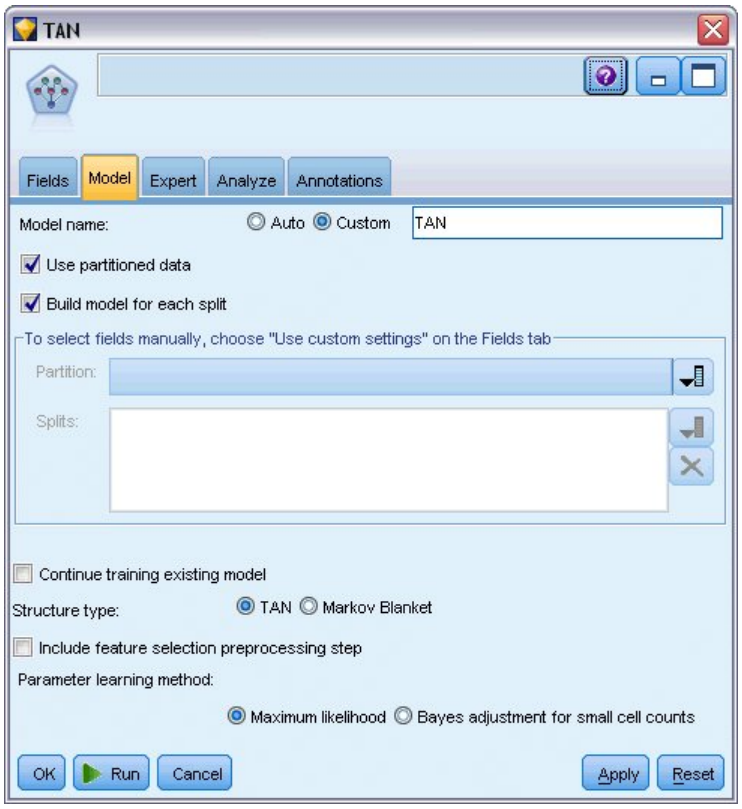


Rysunek 235. Odrzucanie wartości null zmiennej przewidywanej

Ponieważ można zbudować kilka różnych typów sieci bayesowskiej, warto porównać kilka, aby zobaczyć, który model zapewnia najlepsze predykcje. Pierwszym modelem do utworzenia jest Tree Augmented Naïve Bayes (TAN).

7. Załącz węzeł Sieć Bayesa do węzła selekcji.
8. Na karcie Model w pozycji Nazwa modelu wybierz **Użytkownika** i wprowadź TAN w polu tekstowym.

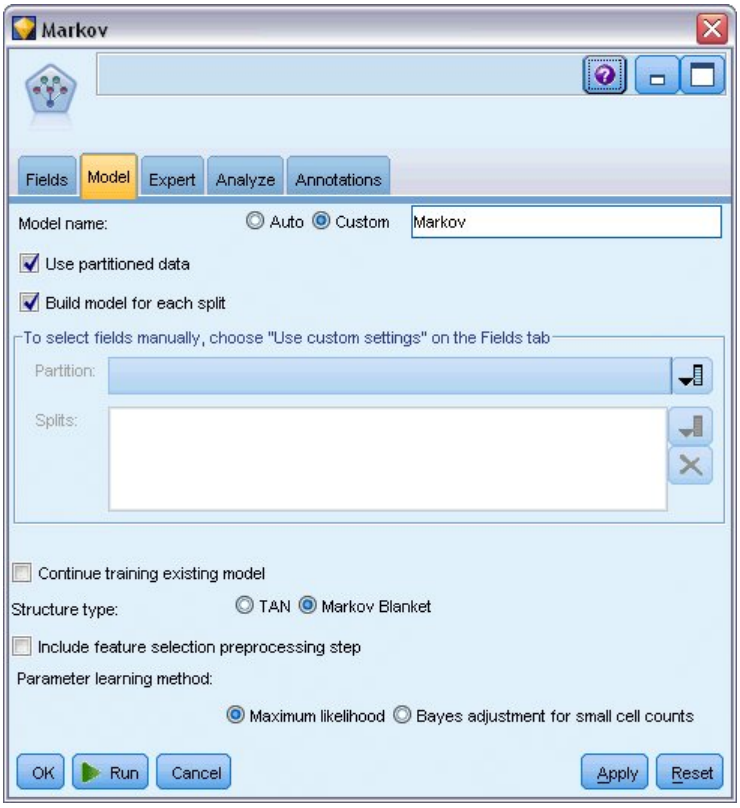
9. W opcji Typ struktury wybierz **TAN** i kliknij przycisk **OK**.



Rysunek 236. Tworzenie modelu Tree Augmented Naïve Bayes (TAN)

Drugi typ modelu do zbudowania ma strukturę modelu Marcov Blanket.

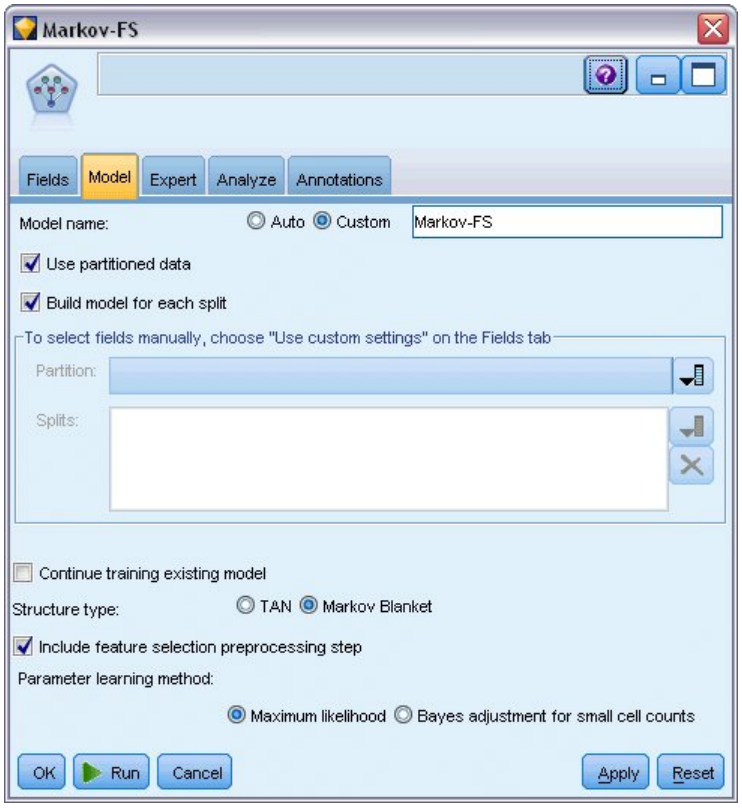
10. Załącz drugi węzeł Sieć Bayesa do węzła selekcji.
11. Na karcie Model w pozycji Nazwa modelu wybierz **Użytkownika** i wprowadź Markov w polu tekstowym.
12. W opcji Typ struktury wybierz **Koc Markowa** i kliknij przycisk **OK**.



Rysunek 237. Tworzenie modelu Koc Markowa

Trzeci typ modelu do zbudowania ma strukturę koca Markowa i używa również wstępnego przetwarzania przy wyborze predyktorów, aby wybrać dane wejściowe, które są istotnie powiązane ze zmienną przewidywaną.

13. Załącz trzeci węzeł Sieć Bayesa do węzła selekcji.
14. Na karcie Model w pozycji Nazwa modelu wybierz **Użytkownika** i wprowadź Markov-FS w polu tekstowym.
15. W opcji Typ struktury wybierz **Koc Markowa**.
16. Zaznacz opcję **Uwzględnij krok wstępnego przetwarzania przy wyborze predyktorów** i kliknij przycisk **OK**.



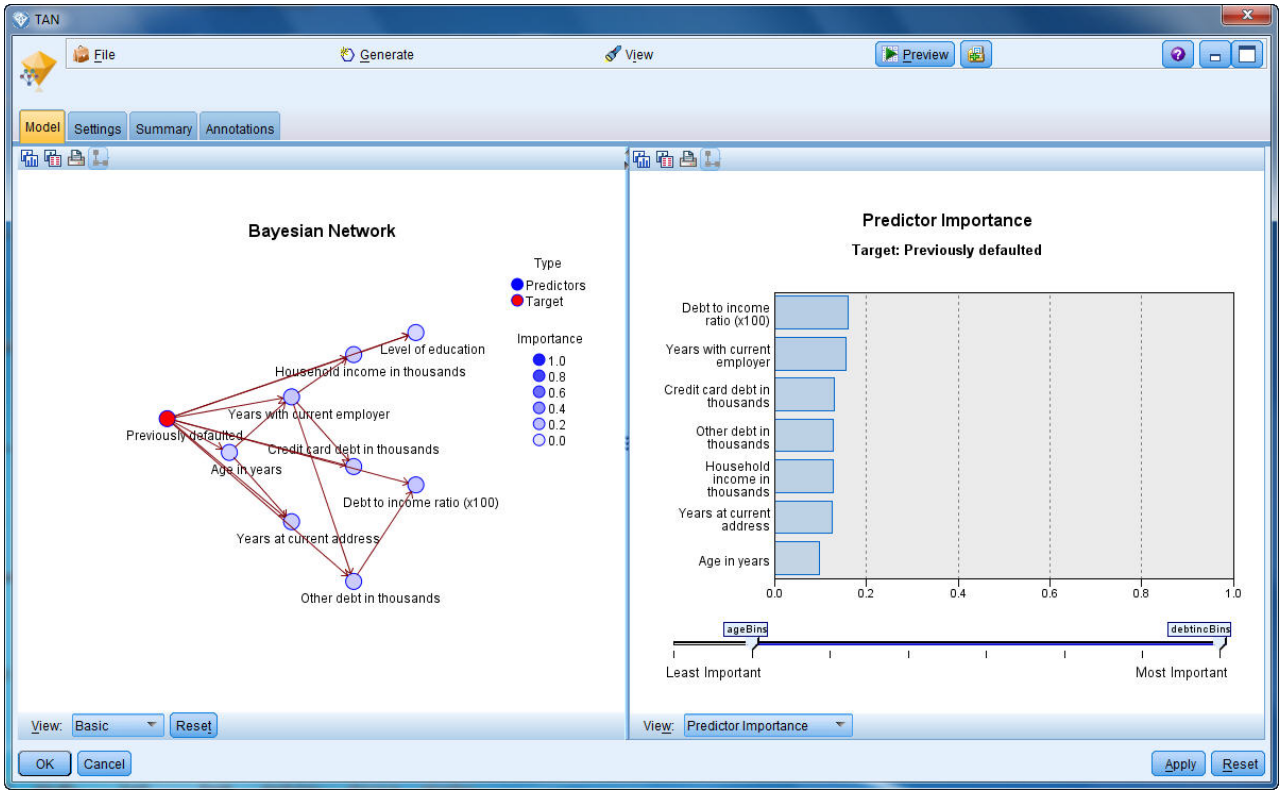
Rysunek 238. Tworzenie modelu Koc Markowa ze wstępnym przetwarzaniem wyboru predyktorów

Przeglądanie modelu

1. Uruchom strumień, aby utworzyć modele użytkowe, które są dodawane do strumienia i palety modeli w prawym górnym rogu. Aby wyświetlić ich szczegóły, dwukrotnie kliknij dowolny model użytkowy w strumieniu.

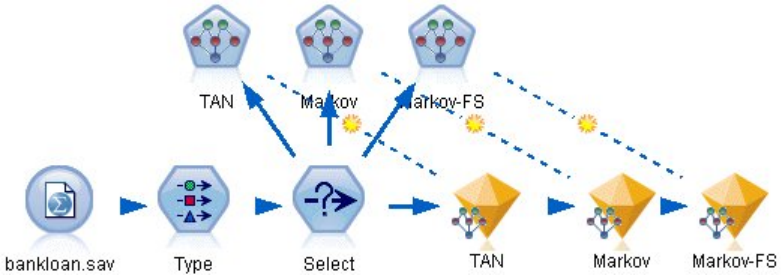
   Karta Model modelu użytkowego jest podzielona na dwa panele: lewy panel zawiera wykres sieci węzłów obrazujący relacje między zmienną przewidywaną a jej najważniejszymi predyktorami oraz relacje między predyktorami.

   Prawy panel przedstawia obszar *Ważność predyktorów*, który wskazuje relatywną ważność każdego predyktora w oszacowaniu modelu, lub obszar *Prawdopodobieństwa warunkowe*, który zawiera wartość prawdopodobieństwa warunkowego dla każdej wartości węzłów i każdej kombinacji wartości w ich węzłach nadrzędnych.



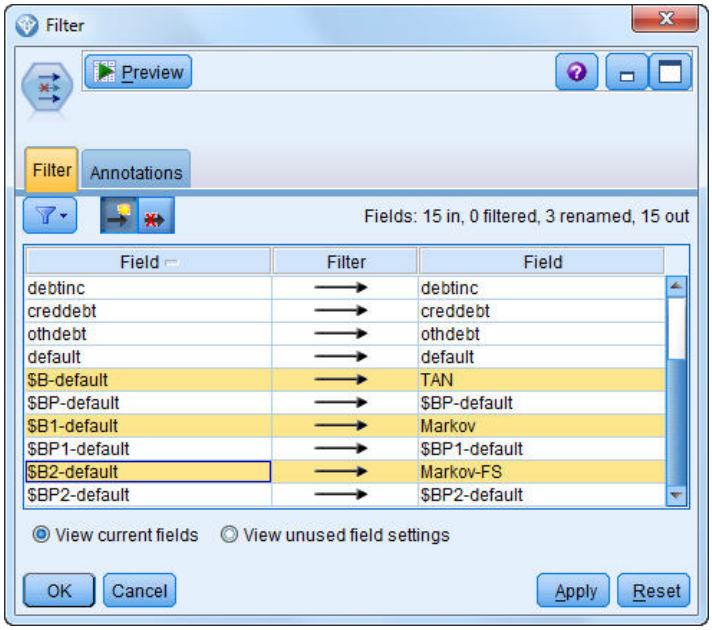
Rysunek 239. Wyświetlanie modelu Tree Augmented Naïve Bayes (TAN)

2. Podłącz model użytkowy TAN do modelu użytkowego Markov (wybierz opcję **Zamień** w oknie ostrzeżenia).
3. Podłącz model użytkowy Markov do modelu użytkowego Markov-FS (wybierz opcję **Zamień** w oknie ostrzeżenia).
4. Wyrównaj trzy modele użytkowe względem węzła selekcji, aby zapewnić łatwe przeglądanie.



Rysunek 240. Wyrównanie modeli użytkowych w strumieniu

5. Aby zmienić nazwę wyników modelu w celu zapewnienia przejrzystości na wykresie ewaluacyjnym, który będzie tworzony, załącz węzeł filtrowania do modelu użytkowego Markov-FS.
6. W prawej kolumnie *Zmienna* zmień wartość $B-default na TAN, $B1-default na Markov i $B2-default na Markov-FS.

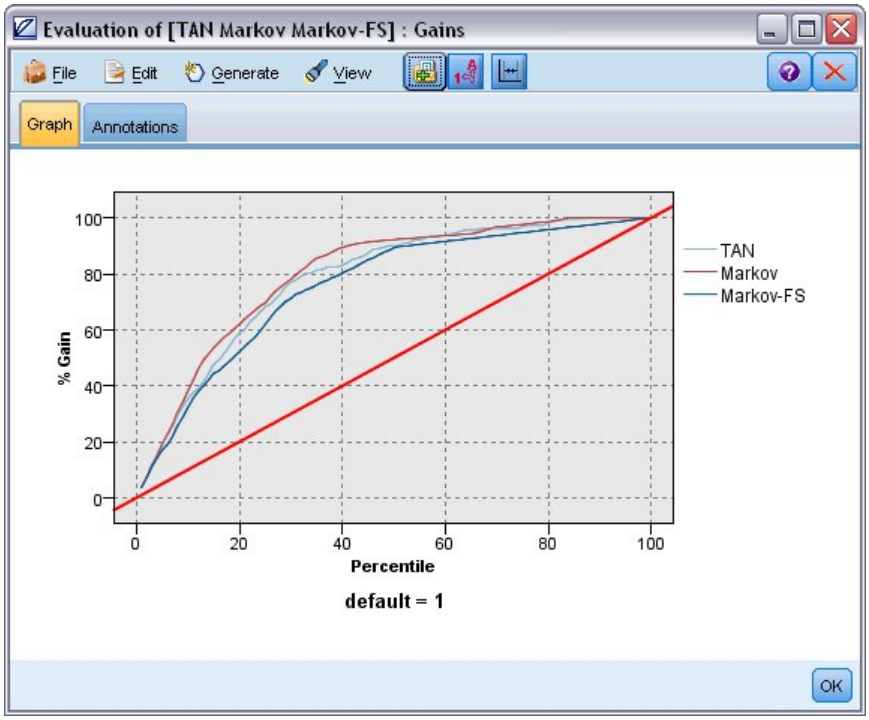


Rysunek 241. Zmiana nazw zmiennych modelu

Aby porównać przewidywaną dokładność modeli, można zbudować wykres korzyści.

7. Załącz węzeł Wykres ewaluacyjny do węzła filtrowania i wykonaj węzeł wykresu, używając jego domyślnych ustawień.

   Wykres pokazuje, że każdy typ modelu daje podobne wyniki, jednak model Markov jest nieznacznie lepszy.

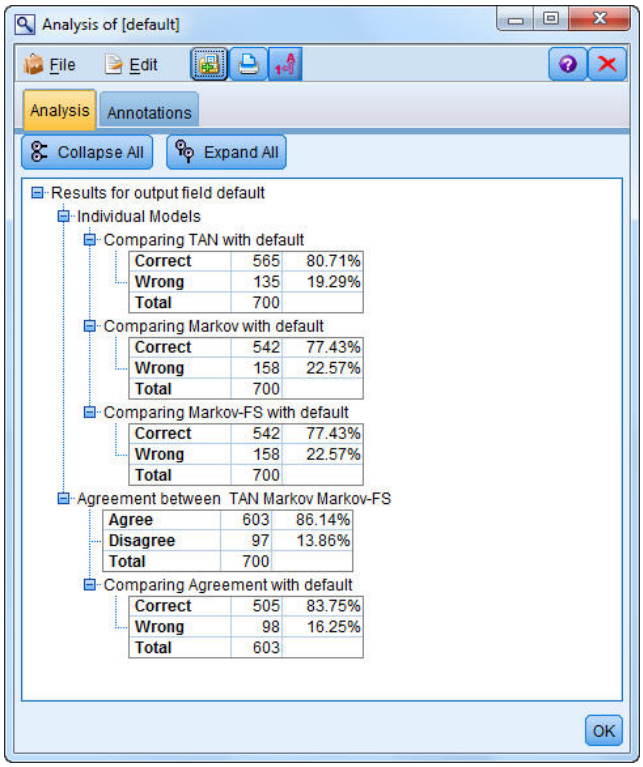


Rysunek 242. Ocena dokładności modeli

Aby sprawdzić, jak dokładne są prognozy każdego modelu, można użyć węzła analizy zamiast wykresu ewaluacyjnego. Pokazuje to dokładność jako wartość procentową dla poprawnych i niepoprawnych predykcji.

8. Załącz węzeł analizy do węzła filtrowania i wykonaj węzeł analizy, używając jego domyślnych ustawień.

Tak jak w przypadku wykresu ewaluacyjnego, analiza pokazuje, że model Markov jest nieznacznie lepszy w prawidłowych predykcjach, jednak model Markov-FS znajduje się tylko kilka punktów procentowych za modelem Markov. Oznacza to, że lepiej użyć modelu Markov-FS, ponieważ używa mniejszej liczby danych wejściowych do obliczania wyników, co pozwala oszczędzić czas gromadzenia i wprowadzania danych oraz przetwarzania.



Rysunek 243. Analizowanie dokładności modeli

Wyjaśnienia podstaw matematycznych metod modelowania używanych w programie IBM SPSS Modeler przedstawiono w publikacji *IBM SPSS Modeler Algorithms Guide* dostępnej w katalogu *\Documentation* na dysku instalacyjnym.

Należy zauważyć, że wyniki opierają się tylko na danych uczących. Aby ocenić, jak dobrze model uogólnia inne dane w rzeczywistych przypadkach, należałoby użyć węzła podziału na podzbiory, aby zatrzymać podzbiór rekordów do celów testowania i walidacji.

Rozdział 18. Ponowne uczenie modelu co miesiąc (Sieć Bayesa)

Sieci bayesowskie umożliwiają utworzenie modelu prawdopodobieństwa przez połączenie zaobserwowanych i zarejestrowanych dowodów ze „zdroworozsądkową” wiedzą rzeczywistą w celu ustanowienia prawdopodobieństwa występowania zdarzeń na podstawie pozornie niepowiązanych ze sobą atrybutów.

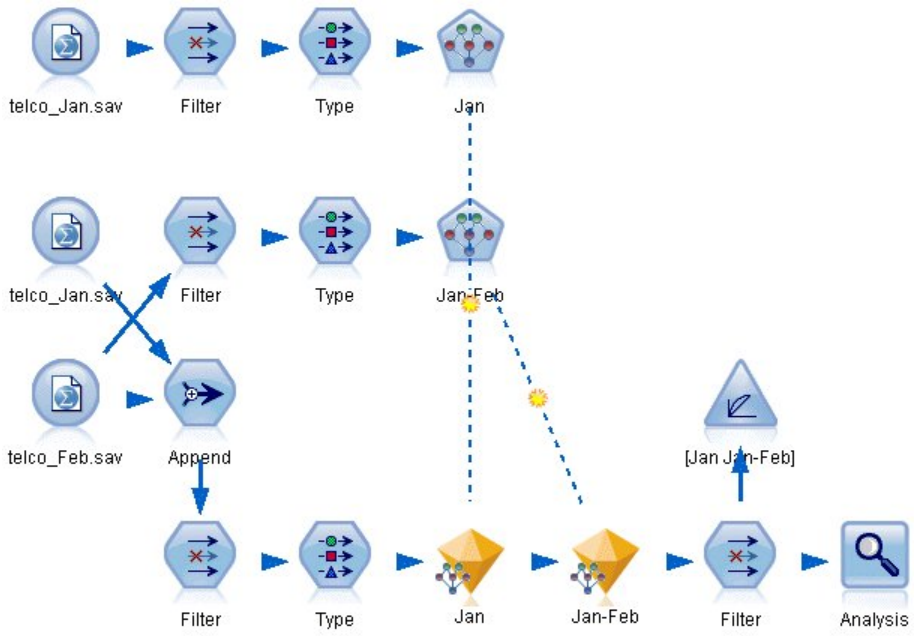
W tym przykładzie zastosowano strumień o nazwie *bayes_churn_retrain.str*, który odwołuje się do plików danych o nazwie *telco_Jan.sav* i *telco_Feb.sav*. Pliki są dostępne w katalogu *Demos* instalacji programu IBM SPSS Modeler i można do nich uzyskać dostęp z grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *bayes_churn_retrain.str* znajduje się w katalogu *streams*.

Załóżmy na przykład, że operator telekomunikacyjny jest zaniepokojony liczbą klientów odchodzących do konkurencji (odchodzenie). Jeśli można użyć historycznych danych klientów do przewidywania klientów, którzy z większym prawdopodobieństwem mogą odejść w przyszłości, można do nich skierować specjalne oferty, aby zniechęcić ich do przejścia do innego dostawcy usług.

Ten przykład koncentruje się na użyciu miesięcznych danych odejścia, aby przewidzieć klientów, którzy z większym prawdopodobieństwem mogą odejść w przyszłości, a następnie dodaniu danych z kolejnego miesiąca, aby udoskonalić i nauczyć ponownie model.

Tworzenie strumienia

1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *telco_Jan.sav* znajdujący się w folderze *Demos*.

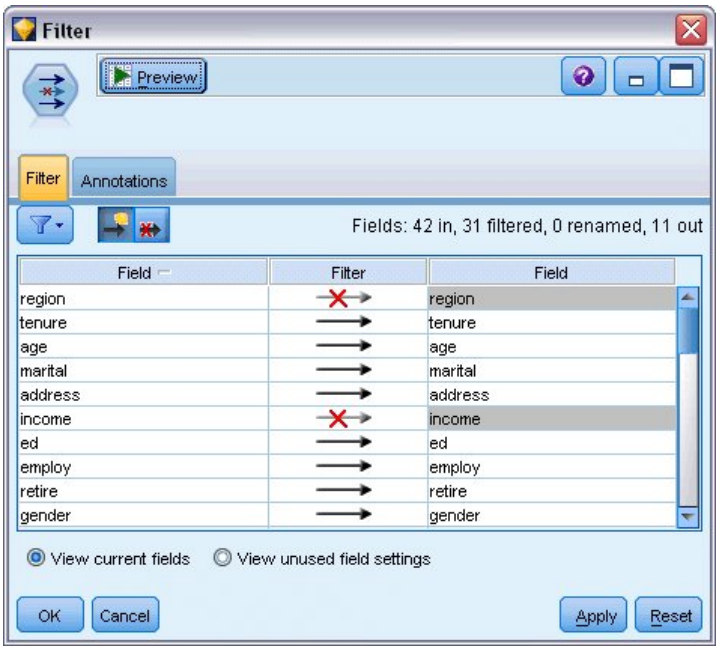


Rysunek 244. Przykładowy strumień Sieć Bayesa

   Wcześniejsza analiza pokazała, że kilka zmiennych danych ma małe znaczenie przy przewidywaniu odejścia. Te zmienne można odfiltrować ze zbioru danych, aby zwiększyć szybkość przetwarzania podczas budowy i oceny modeli.

2. Dodaj węzeł filtrowania do węzła źródłowego.

3. Wyłącz wszystkie zmienne za wyjątkiem *address*, *age*, *churn*, *custcat*, *ed*, *employ*, *gender*, *marital*, *reside*, *retire* i *tenure*.
4. Kliknij przycisk **OK**.



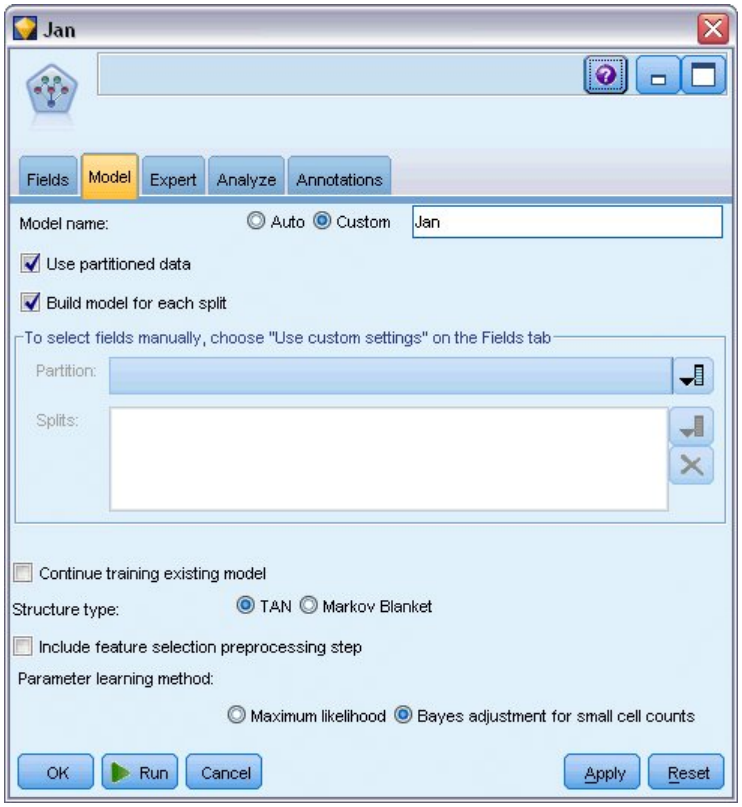
Rysunek 245. Filtrowanie zbędnych zmiennych

5. Załącz węzeł typu do węzła filtrowania.
6. Otwórz węzeł typu i kliknij przycisk **Odczytaj wartości**, aby wypełnić kolumnę *Wartości*.
7. Aby węzeł Ocena mógł ocenić, która wartość jest prawdziwa, a która fałszywa, ustaw poziom pomiaru zmiennej *churn* na **Flaga** i ustaw jej rolę na **Przewidywana**. Kliknij przycisk **OK**.

Rysunek 246. Wybieranie zmiennej przewidywanej

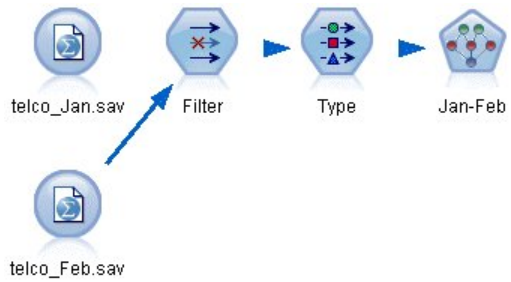
Można zbudować kilka różnych typów sieci bayesowskich, ale w tym przykładzie utworzymy model Tree Augmented Naïve Bayes (TAN). Tworzy to dużą sieć i zapewnia, że uwzględniono wszystkie możliwe łącza pomiędzy zmiennymi danych, budując w ten sposób wydajny model początkowy.

8. Załącz węzeł Sieć Bayesa do węzła typu.
9. Na karcie Model w pozycji Nazwa modelu wybierz **Użytkownika** i wprowadź Jan w polu tekstowym.
10. W opcji Metoda uczenia parametrów wybierz pozycję **Korekta Bayesa dla niewielkiej liczby komórek**.
11. Kliknij przycisk **Uruchom**. Model użytkowy zostaje dodany do obszaru roboczego oraz na palecie modeli w prawym górnym rogu.



Rysunek 247. Tworzenie modelu Tree Augmented Naïve Bayes (TAN)

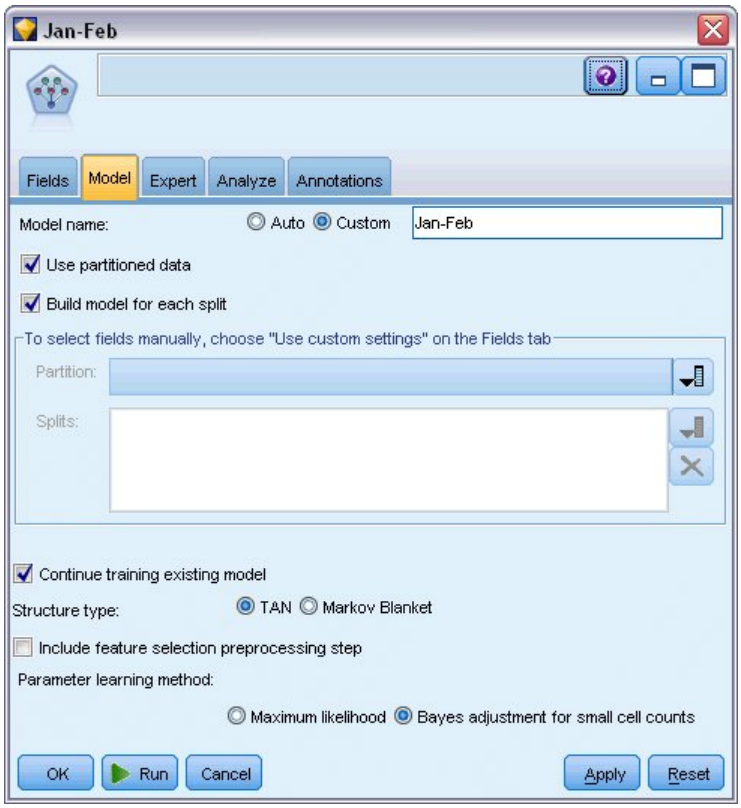
12. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *telco_Feb.sav* znajdujący się w folderze *Demos*.
13. Dołącz ten nowy węzeł źródłowy do węzła filtrowania (w oknie ostrzeżenia wybierz opcję **Zamień**, aby zastąpić połączenie z poprzednim węzłem źródłowym).



Rysunek 248. Dodawanie danych drugiego miesiąca

14. Na karcie Model węzła Sieć Bayesa w pozycji Nazwa modelu wybierz **Użytkownika** i wprowadź Jan-Feb w polu tekstowym.

15. Zaznacz opcję **Kontynuuj uczenie istniejącego modelu**.
16. Kliknij przycisk **Uruchom**. Model użytkowy zastępuje istniejący model w strumieniu, ale zostaje też dodany na palecie modeli w prawym górnym rogu.



Rysunek 249. Ponowne uczenie modelu

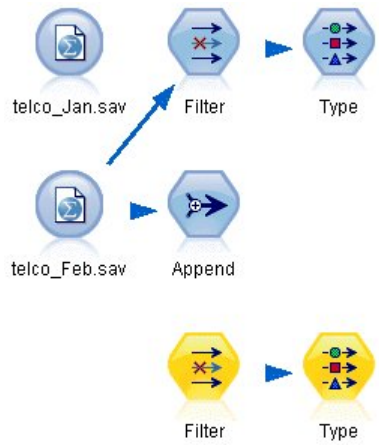
Ewaluacja modelu

Aby porównać modele, należy połączyć dwa zbiory danych.

1. Dodaj węzeł Dołączanie i przyłącz do niego oba węzły źródłowe *telco_Jan.sav* i *telco_Feb.sav*.

Rysunek 250. Dołączanie dwóch źródeł danych

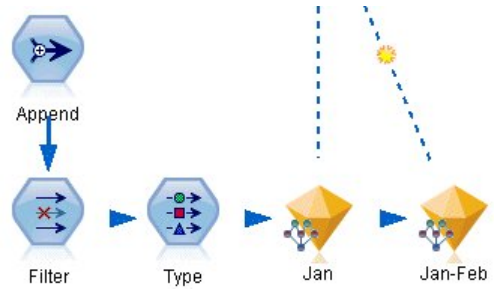
2. Skopiuj węzły filtrowania i typu z wcześniejszej części strumienia i wklej je w obszarze roboczym strumienia.
3. Przyłącz węzeł Dołączanie do nowo skopiowanego węzła filtrowania.



Rysunek 251. Wklejanie skopiowanych węzłów do strumienia

Modele użytkowe dwóch modeli Sieć Bayesa znajdują się na palecie modeli w prawym górnym rogu.

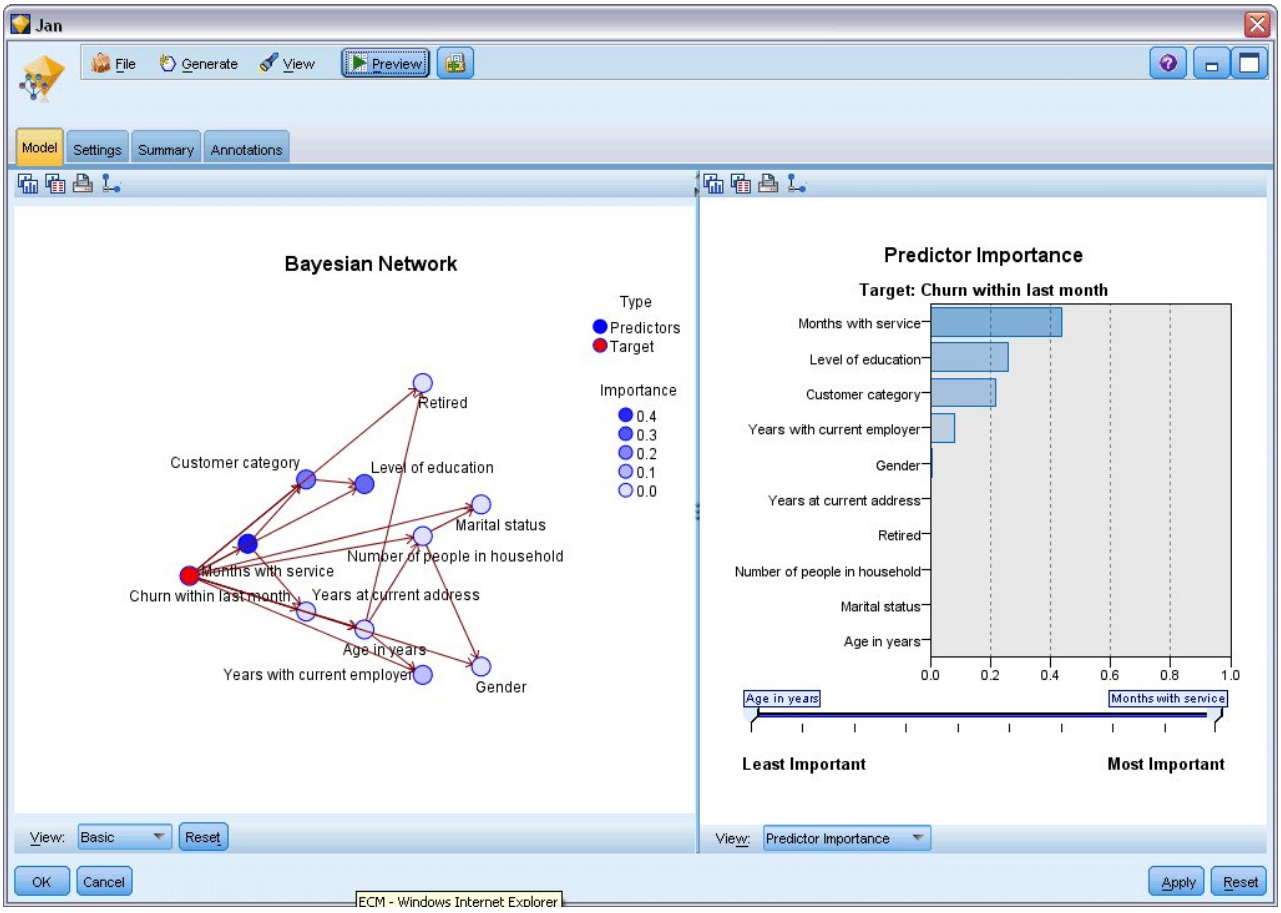
4. Dwukrotnie kliknij model użytkowy Jan, aby przenieść go do strumienia, i dołącz go do nowo skopiowanego węzła typu.
5. Dołącz model użytkowy Jan-Feb znajdujący się już w strumieniu do modelu użytkowego Jan.
6. Otwórz model użytkowy Jan.



Rysunek 252. Dodawanie modeli użytkowych do strumienia

Karta Model modelu użytkowego Sieć Bayesa jest podzielona na dwie kolumny. Lewa kolumna zawiera wykres sieci węzłów obrazujący relacje między zmienną przewidywaną a jej najważniejszymi predyktorami oraz relacje między predyktorami.

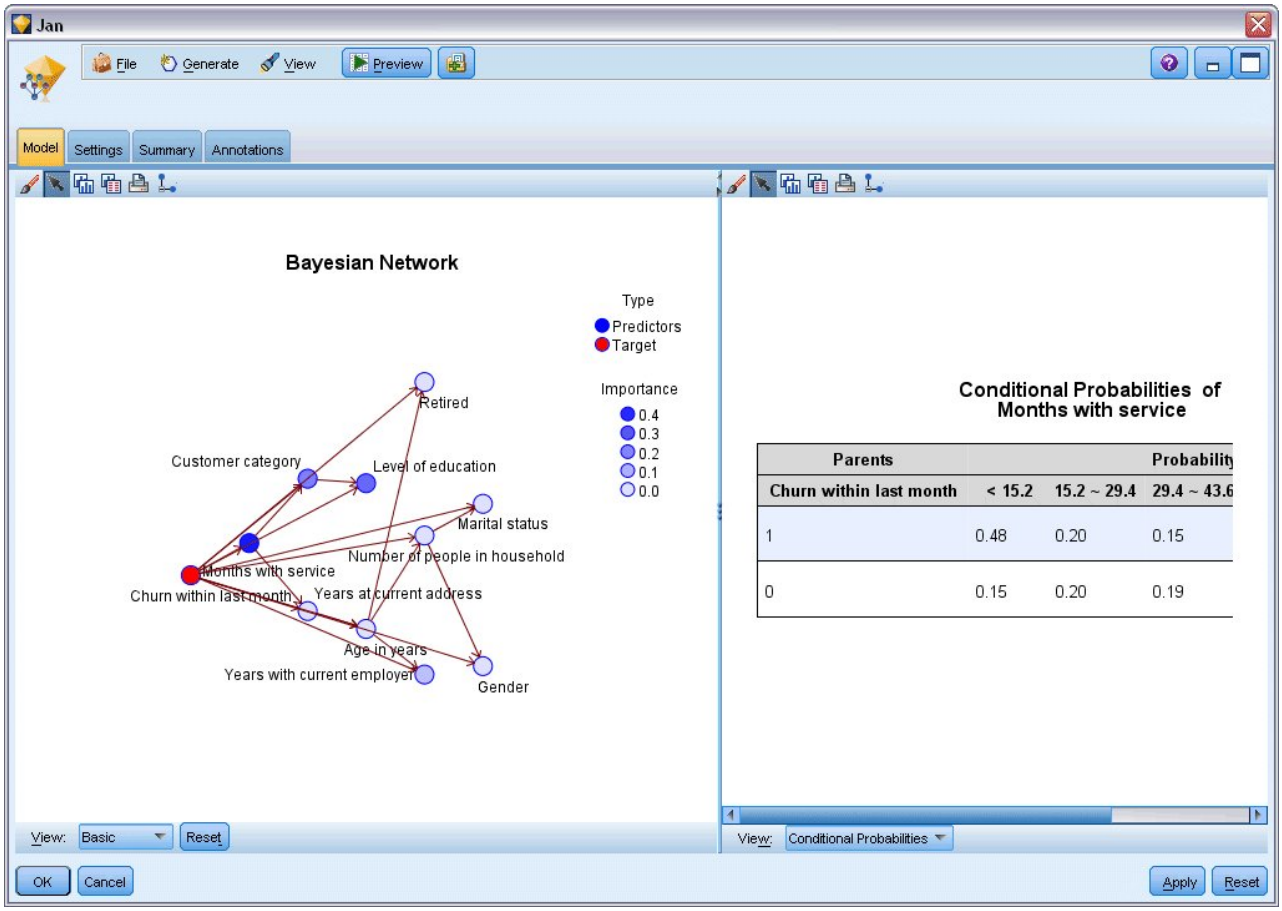
Prawa kolumna przedstawia obszar *Ważność predyktorów*, który wskazuje relatywną ważność każdego predyktora w oszacowaniu modelu, lub obszar *Prawdopodobieństwa warunkowe*, który zawiera wartość prawdopodobieństwa warunkowego dla każdej wartości węzłów i każdej kombinacji wartości w ich węzłach nadrzędnych.



Rysunek 253. Model Sieć Bayesa przedstawiający ważność predyktorów

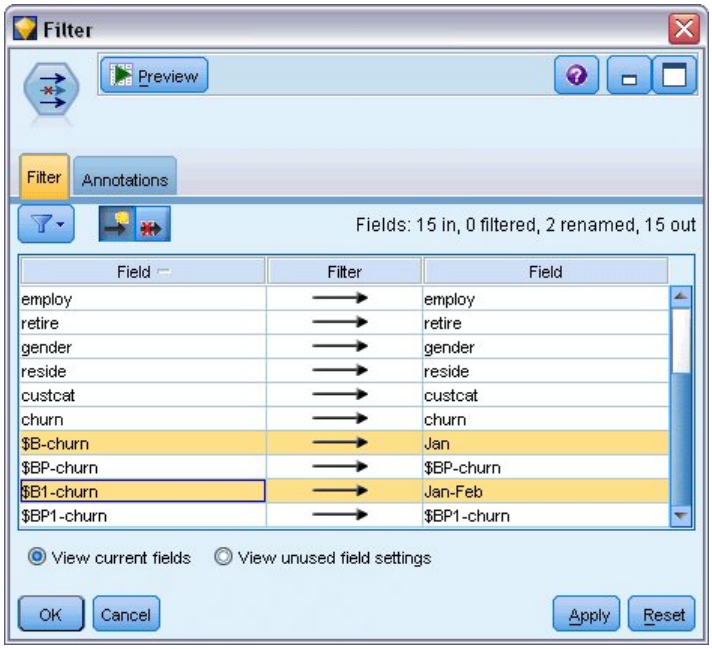
Aby wyświetlić prawdopodobieństwo warunkowe dla dowolnego węzła, kliknij węzeł w lewej kolumnie. Prawa kolumna jest aktualizowana, aby wyświetlić wymagane szczegóły.

Prawdopodobieństwo warunkowe jest pokazane dla każdego przedziału, na który zostały podzielone wartości danych względem węzłów nadrzędnych i równorzędnych węzła.



Rysunek 254. Model Sieć Bayesa przedstawiający prawdopodobieństwa warunkowe

7. Aby zmienić nazwę wyników modelu w celu zapewnienia przejrzystości, załącz węzeł filtrowania do modelu użytkowego Jan-Feb.
8. W prawej kolumnie *Zmienna* zmień nazwę $B-churn na Jan i $B1-churn na Jan-Feb.

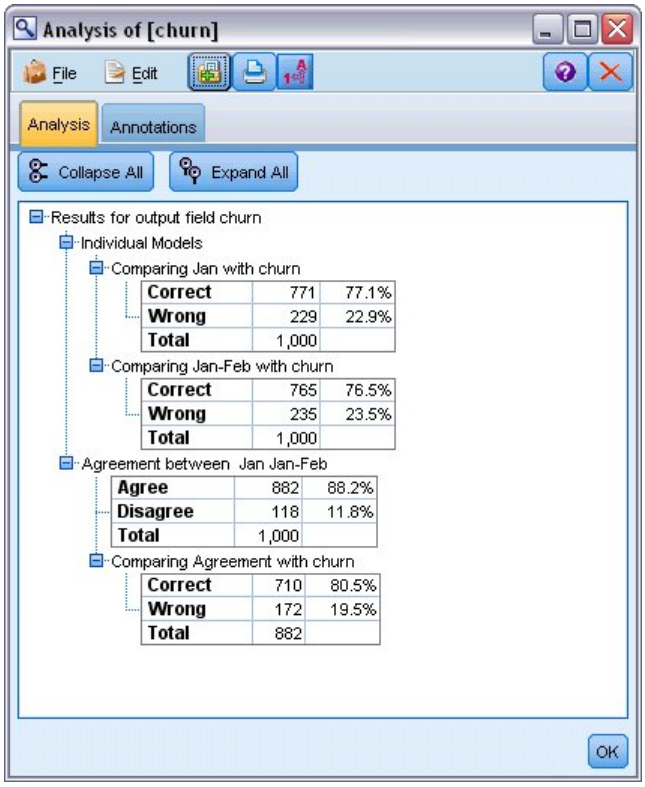


Rysunek 255. Zmiana nazw zmiennych modelu

Aby sprawdzić, jak dobrze każdy model przewiduje odejście, użyj węzła analizy. Pokazuje to dokładność jako wartość procentową dla poprawnych i niepoprawnych predykcji.

9. Załącz węzeł analizy do węzła filtrowania.
10. Otwórz węzeł analizy i kliknij przycisk **Uruchom**.

    Pokazuje to, że oba modele mają podobny stopień dokładności przy przewidywaniu odejścia.



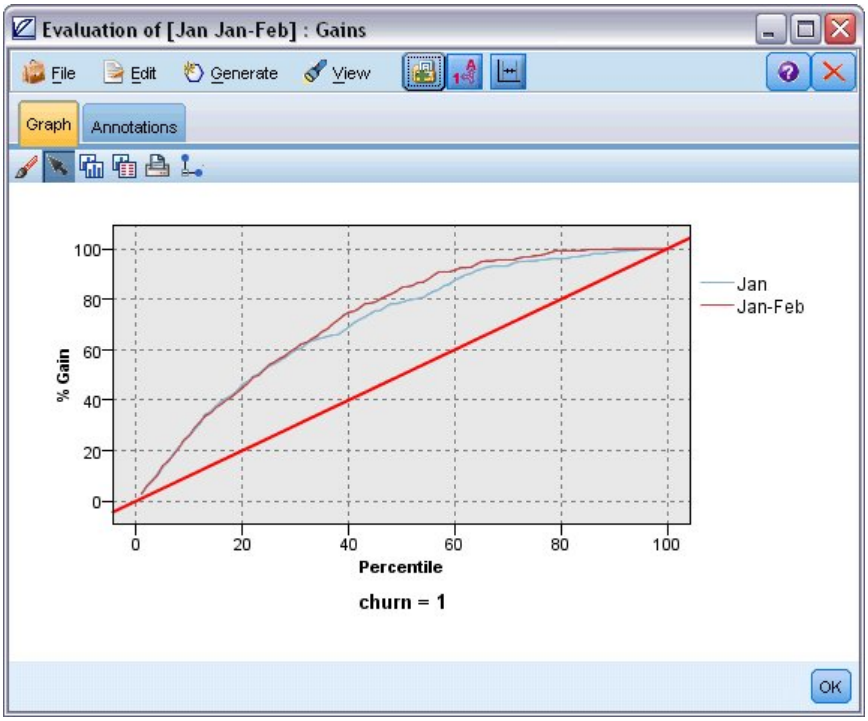
Rysunek 256. Analizowanie dokładności modeli

Jako alternatywę węzła analizy można użyć wykresu ewaluacyjnego, aby porównać przewidywaną dokładność modeli, budując wykres korzyści.

11. Załącz węzeł Wykres ewaluacyjny do węzła filtrowania

i uruchom węzeł wykresu, używając ustawień domyślnych.

Tak jak w przypadku węzła analizy wykres pokazuje, że oba typy modelu generują podobne wyniki. Jednak model używający nauczonego ponownie modelu korzystającego z danych z obu miesięcy jest nieznacznie lepszy, ponieważ ma wyższy poziom ufności w predykcjach.



Rysunek 257. Ocena dokładności modeli

Wyjaśnienia podstaw matematycznych metod modelowania używanych w programie IBM SPSS Modeler przedstawiono w publikacji *IBM SPSS Modeler Algorithms Guide* dostępnej w katalogu *\Documentation* na dysku instalacyjnym.

Należy zauważyć, że wyniki opierają się tylko na danych uczących. Aby ocenić, jak dobrze model uogólnia inne dane w rzeczywistych przypadkach, należałoby użyć węzła podziału na podzbiory, aby zatrzymać podzbiór rekordów do celów testowania i walidacji.

Rozdział 19. Promocja sprzedaży detalicznej (Sieci neuronowe/C&RT)

Ten przykład dotyczy danych, które opisują linie produktów detalicznych oraz efekty promocji na sprzedaż. (Te dane są fikcyjne). Celem użytkownika w tym przykładzie jest przewidzenie efektów przyszłych promocji sprzedaży. Podobnie jak w przykładzie monitorowania warunków proces eksploracji danych składa się z eksploracji, przygotowania danych, uczenia i faz testowych.

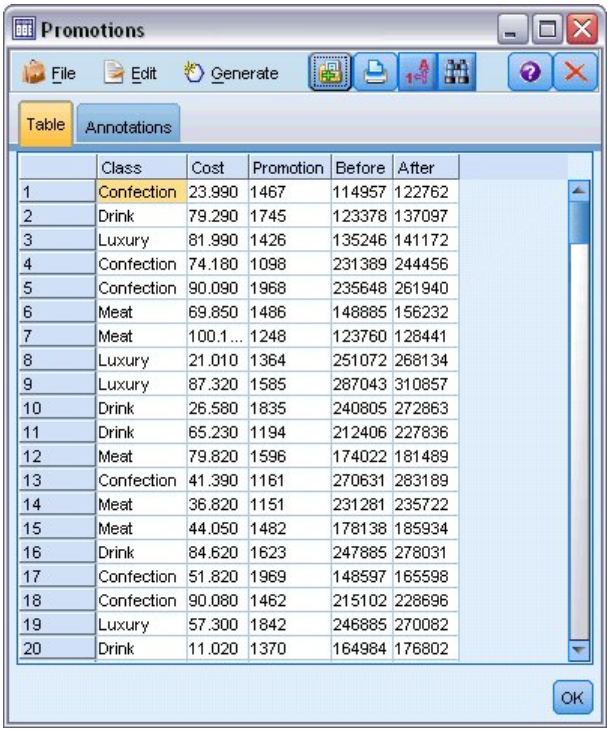
W tym przykładzie używane są strumienie o nazwach *goodsplot.str* i *goodslearn.str*, które odwołują się do plików danych *GOODS1n* i *GOODS2n*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Strumień *goodsplot.str* znajduje się w folderze *streams*, podczas gdy plik *goodslearn.str* znajduje się w katalogu *streams*.

Badanie danych

Każdy rekord zawiera następujące dane:

- *Class*. Typ produktu.
- *Cost*. Cena jednostkowa.
- *Promotion*. Indeks kwoty wydanej na określoną promocję.
- *Before*. Przychody przed promocją.
- *After*. Przychody po promocji.

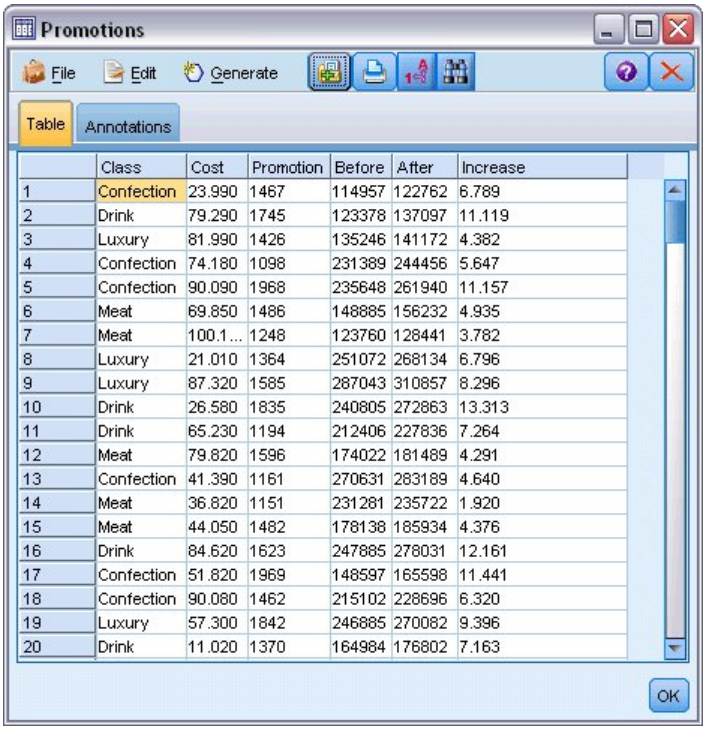
Strumień *goodsplot.str* zawiera prosty strumień wyświetlający dane w tabeli. Dwie zmienne przychodów (*Before* i *After*) są wyrażane jako wartości bezwzględne. Prawdopodobne jest jednak, że zwiększenie przychodów po promocji (i przypuszczalnie w jej wyniku) byłoby bardziej przydatną wartością.



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

Rysunek 258. Efekty promocji na sprzedaż produktów

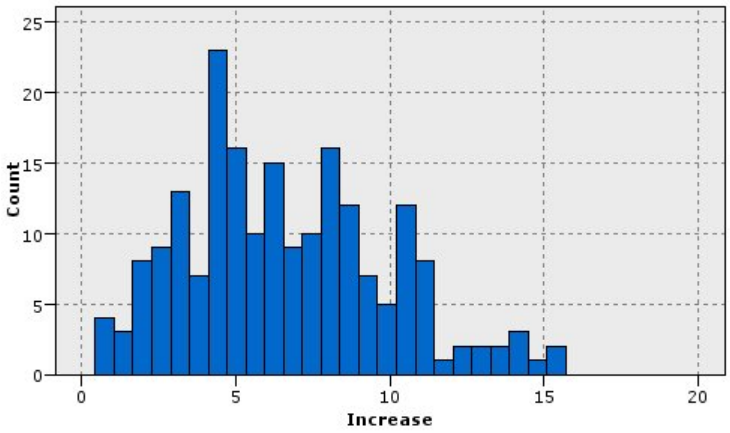
Strumień *goodsplot.str* zawiera również węzeł wyliczający tę wartość, wyrażoną jako wartość procentowa przychodów przed promocją w zmiennej o nazwie *Increase* i wyświetla tabelę przedstawiającą tę zmienną.



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

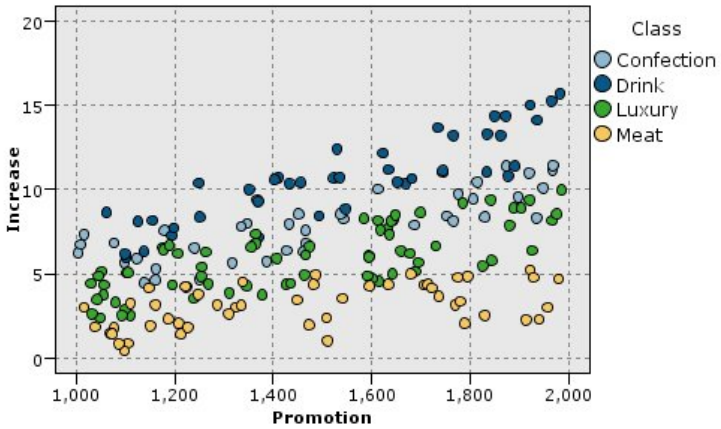
Rysunek 259. Zwiększenie przychodów po promocji

Dodatkowo strumień wyświetla histogram zwiększenia oraz wykres rozrzutu zwiększenia przychodów dla poniesionych kosztów promocji z nałożonymi kategoriami produktów.



Rysunek 260. Histogram zwiększenia przychodów

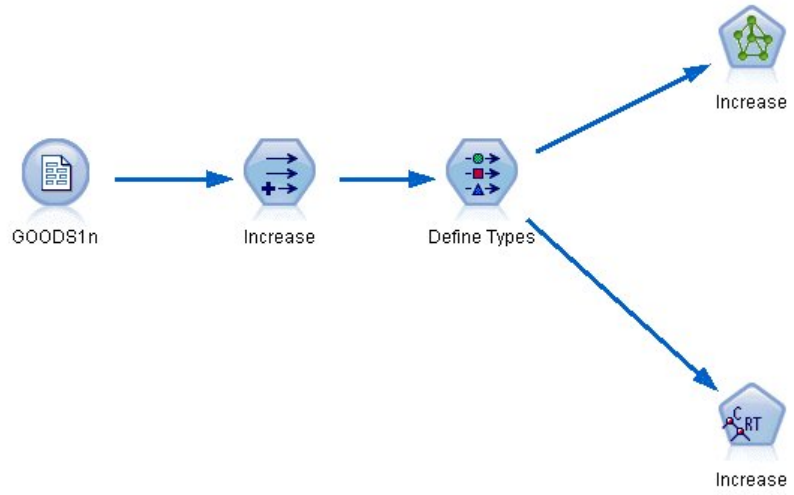
Wykres rozrzutu pokazuje, że dla każdej klasy produktów istnieje prawie liniowy związek pomiędzy zwiększeniem przychodów i kosztem promocji. Dlatego też wydaje się prawdopodobne, że drzewo decyzyjne lub sieć neuronowa mogłyby przewidzieć z rozsądną dokładnością zwiększenie przychodów z innych dostępnych zmiennych.



Rysunek 261. Zwiększenie przychodów w stosunku do wydatków promocyjnych

Nauka i testowanie

Strumień goodslearn.str uczy sieć neuronową i drzewo decyzyjne prognozowania wzrostu przychodów.



Rysunek 262. Modelowanie strumienia goodslearn.str

Po wykonaniu węzłów modelu i wygenerowaniu modeli rzeczywistych można testować wyniki procesu uczenia. Można to zrobić, łącząc drzewo decyzyjne i sieć w szereg pomiędzy węzłem typu i nowym węzłem analizy, zmieniając plik danych wejściowych na GOODS2n i wykonując węzeł analizy. Na wyjściu tego węzła, w szczególności z korelacji liniowej pomiędzy przewidywanym wzrostem a poprawną odpowiedzią, widoczne będzie, że nauczone systemy przewidują zwiększenie przychodów z dużą skutecznością.

Dalsza eksploracja może koncentrować się na przypadkach, w których nauczone systemy robią relatywnie duże błędy. Można je wykryć, rysując wykres przewidywanego wzrostu przychodów dla rzeczywistego wzrostu. Wartości odstające na tym wykresie można wybrać, używając interaktywnych grafik w programie SPSS Modeler i korzystając z ich właściwości możliwe może być dostosowanie opisu danych lub procesu nauki, aby poprawić dokładność.

Rozdział 20. Monitorowanie warunków (Sieci neuronowe/C5.0)

Ten przykład dotyczy monitorowania informacji o statusie maszyny oraz problemów z rozpoznaniem i przewidywaniem stanów awaryjnych. Dane zostały utworzone z fikcyjnej symulacji i składają się z wielu połączonych szeregów zmierzonych w czasie. Każdy rekord to obraz stanu maszyny uwzględniający następujące parametry:

- *Time*. Liczba całkowita.
- *Power*. Liczba całkowita.
- *Temperature*. Liczba całkowita.
- *Pressure*. 0, jeśli normalne, 1 dla chwilowego ostrzeżenia o ciśnieniu.
- *Uptime*. Czas od ostatnich prac serwisowych.
- *Status*. Zazwyczaj 0, zmienia się na kod błędu przy wystąpieniu błędu (101, 202 lub 303).
- *Outcome*. Kod błędu, który pojawia się w tym szeregu czasowym lub 0, jeśli nie występuje żaden błąd. (Te kody są dostępne tylko dzięki wiedzy po fakcie).

W tym przykładzie używane są strumienie o nazwach *condplot.str* i *condlearn.str*, które odwołują się do plików danych *COND1n* i *COND2n*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Pliki *condplot.str* i *condlearn.str* znajdują się w katalogu *streams*.

Dla każdego szeregu czasowego istnieje szereg rekordów z okresu normalnego działania, po którym następuje okres prowadzący do awarii, jak pokazano w poniższej tabeli:

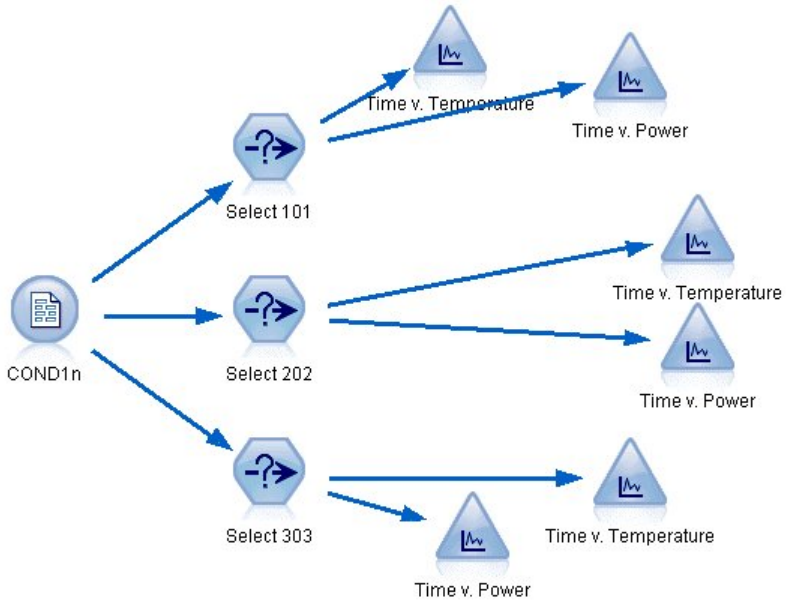
Czas	Moc	Temperatura	Ciśnienie	Czas dostępności	Status	Wynik
0	1059	259	0	404	0	0
1	1059	259	0	404	0	0
			...			
51	1059	259	0	404	0	0
52	1059	259	0	404	0	0
53	1007	259	0	404	0	303
54	998	259	0	404	0	303
			...			
89	839	259	0	404	0	303
90	834	259	0	404	303	303
0	965	251	0	209	0	0
1	965	251	0	209	0	0
			...			
51	965	251	0	209	0	0
52	965	251	0	209	0	0
53	938	251	0	209	0	101
54	936	251	0	209	0	101
			...			
208	644	251	0	209	0	101
209	640	251	0	209	101	101

Następujący proces jest wspólny w większości projektów eksploracji danych:

- Badanie danych, aby określić, które atrybuty mogą być istotne do przewidywania lub rozpoznania stanów badanego obiektu.
- Zachowanie tych atrybutów (jeśli są już obecne) lub wyliczenie i dodanie ich do danych, jeśli istnieje taka potrzeba.
- Użycie wynikowych danych do reguł uczenia i sieci neuronowych.
- Testowanie nauczonych systemów z użyciem niezależnych danych testowych.

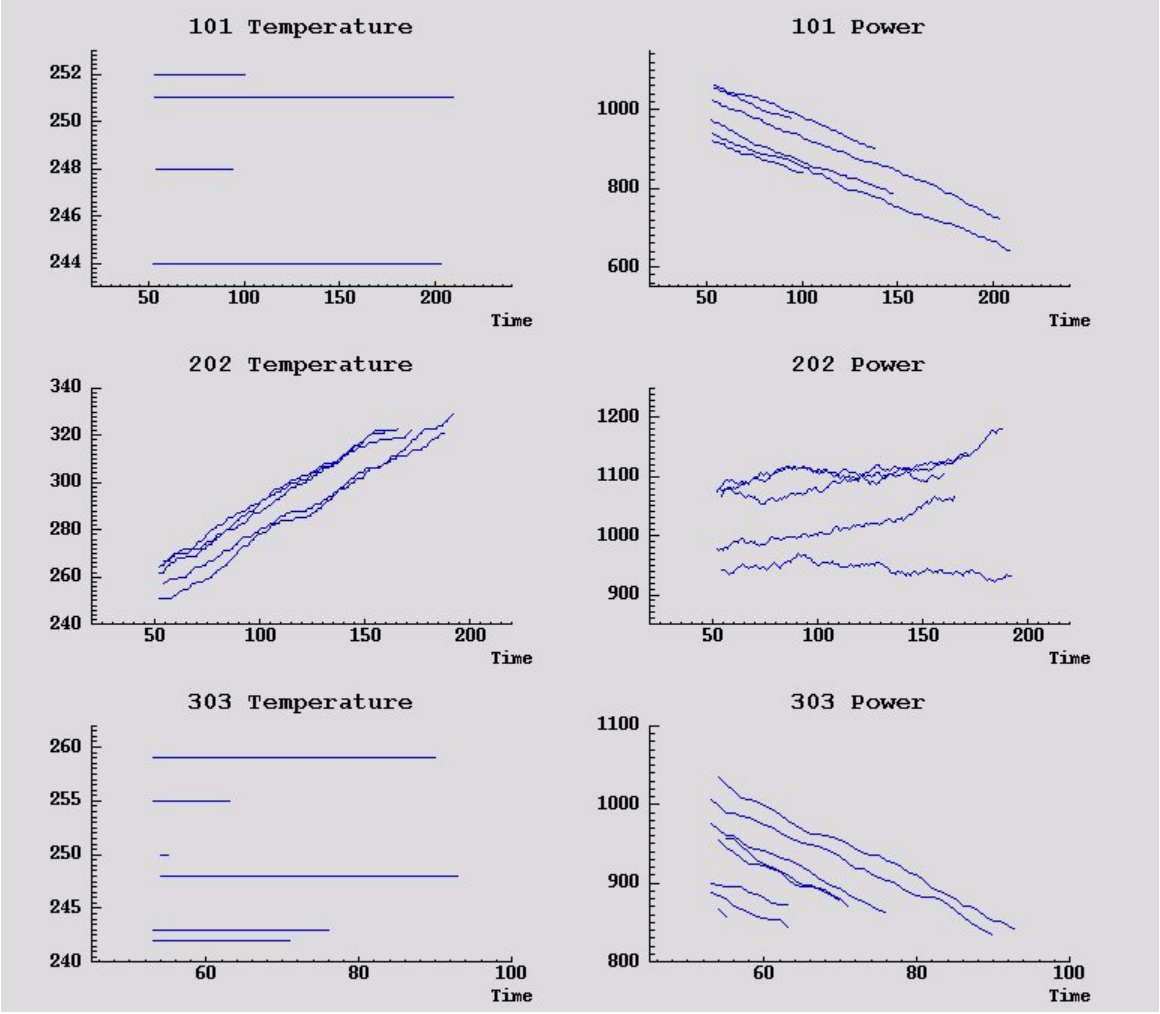
Badanie danych

Plik *condplot.str* ilustruje pierwszą część procesu. Plik zawiera strumień rysujący różne wykresy. Jeśli szereg czasowy temperatury lub mocy zawiera widoczne wzory, możliwe jest rozróżnienie nadchodzących warunków błędu lub przewidzieć ich wystąpienie. Dla temperatury i mocy strumień rysuje wykresy szeregów czasowych powiązanych z trzema różnymi kodami błędów na osobnych wykresach, co daje sześć wykresów. Węzły selekcji rozdzielają dane powiązane z różnymi kodami błędów.



Rysunek 263. Strumień condplot

Wyniki tego strumienia przedstawiono na rysunku.



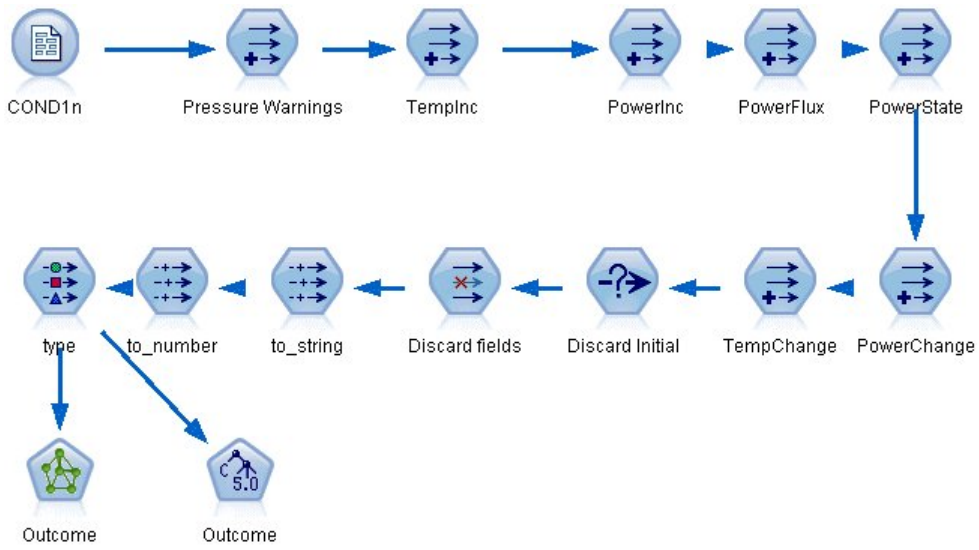
Rysunek 264. Temperatura i moc w czasie

Wykresy wyraźnie przedstawiają wzorce rozróżniające błędy 202 od błędów 101 i 303. Błędy 202 pokazują rosnącą temperaturę i zmienną moc w czasie. Pozostałe błędy nie mają takiej charakterystyki. Wzorce rozróżniające błędy 101 i 303 są już jednak mniej wyraźne. Oba typy błędów wykazują równą temperaturę i spadek mocy, ale spadek mocy wydaje się bardziej stromy dla błędów 303.

Na podstawie tych wykresów widać, że obecność i współczynnik zmiany dla temperatury i mocy, jak również obecność i stopień fluktuacji są istotne do przewidzenia i rozróżnienia awarii. Dlatego też te atrybuty należy dodać do danych przed zastosowaniem systemów uczenia.

Przygotowanie danych

Na podstawie wyników eksplorowania danych strumień *condlearn.str* wylicza powiązane dane i uczy się przewidywania awarii.



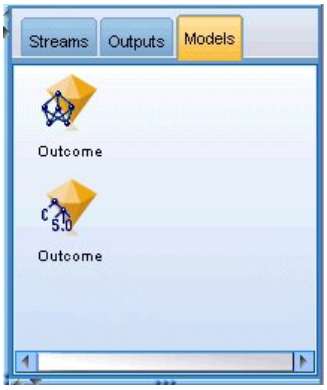
Rysunek 265. Strumień condlearn

Strumień używa wielu węzłów wyliczeń do przygotowania danych do modelowania.

- **Węzeł pliku zmiennych**. Odczytuje plik danych *COND1n*.
- **Wyliczanie Pressure Warnings**. Zlicza liczbę chwilowych ostrzeżeń dotyczących ciśnienia. Wartość resetowana, gdy czas powraca do 0.
- **Wyliczanie TempInc**. Oblicza chwilowy współczynnik zmiany temperatury za pomocą @DIFF1.
- **Wyliczanie PowerInc**. Oblicza chwilowy współczynnik zmiany mocy za pomocą @DIFF1.
- **Wyliczanie PowerFlux**. Flaga, która ma wartość prawda, jeśli moc zmieniła się w przeciwnym kierunku we wcześniejszym rekordzie i w tym rekordzie, czyli szczyt mocy lub spadek.
- **Wyliczanie PowerState**. Stan, który rozpoczyna się jako *Stable* i przełącza na *Fluctuating*, gdy wykryte zostaną dwa kolejne skoki mocy. Przełącza się z powrotem na stan *Stable* tylko, gdy nie wystąpił skok mocy przez pięć przedziałów czasowych lub zmienna *Time* została zresetowana.
- **PowerChange**. Średnia z *PowerInc* w ostatnich pięciu przedziałach czasowych.
- **TempChange**. Średnia z *TempInc* w ostatnich pięciu przedziałach czasowych.
- **Discard Initial (Wybierz)**. Odrzuca pierwszy rekord każdego szeregu czasowego, aby uniknąć dużego (niepoprawnego) skoku wartości *Power* i *Temperature* przy granicach przedziału.
- **Discard fields**. Ogranicza rekordy do *Uptime*, *Status*, *Outcome*, *Pressure Warnings*, *PowerState*, *PowerChange* i *TempChange*.
- **Typ**. Definiuje rolę zmiennej *Outcome* jako **Przewidywana** (zmienna przewidywana). Dodatkowo węzeł definiuje poziom pomiaru zmiennej *Outcome* jako **Nominalna**, *Pressure Warnings* jako **Ilościowa** i *PowerState* jako **Flaga**.

Uczenie

Uruchomienie strumienia w pliku *condlearn.str* uczy regułę C5.0 i sieć neuronową. Uczenie sieci może zająć dużo czasu, ale uczenie można przerwać wcześnie, aby zapisać sieć, która generuje akceptowalne wyniki. Po zakończeniu uczenia karta Modele w prawym górnym rogu okna menedżerów miga, aby powiadomić użytkownika, że utworzono dwa nowe modele użytkowe: jeden reprezentujący sieć neuronową, a drugi reprezentujący regułę.



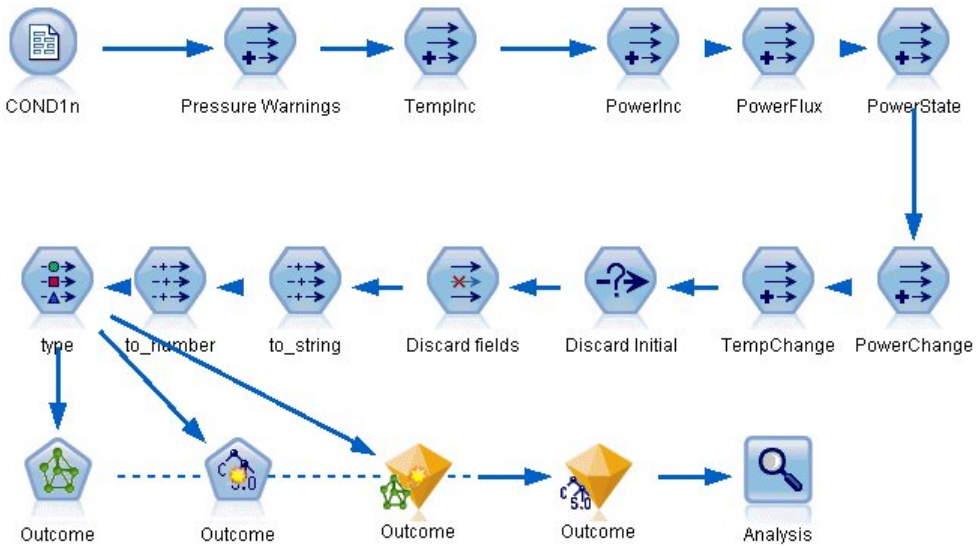
Rysunek 266. Menedżer modeli z modelami użytkowymi

Modele użytkowe są również dodawane do istniejącego strumienia, co pozwala na testowanie systemu lub eksportowanie wyników modelu. W tym przykładzie przetestujemy wyniki modelu.

Testowanie

Modele użytkowe są dodawane do strumienia, oba z nich podłączone do węzła typu.

1. Zmień pozycję modeli użytkowych zgodnie z rysunkiem, aby węzeł typu łączył się z modelem użytkowym sieci neuronowej, który łączy się z modelem użytkowym C5.0.
2. Załącz węzeł analizy do modelu użytkowego C5.0.
3. Edytuj oryginalny węzeł źródłowy, aby odczytał plik *COND2n* (zamiast pliku *COND1n*), ponieważ *COND2n* zawiera nieużywane dane testowe.



Rysunek 267. Testowanie nauczonej sieci

4. Otwórz węzeł analizy i kliknij przycisk Uruchom.

Powoduje to wygenerowanie wartości odzwierciedlających dokładność nauczonej sieci i reguły.

Rozdział 21. Klasyfikowanie klientów usług telekomunikacyjnych (Analiza dyskryminacyjna)

Analiza dyskryminacyjna to technika statystyczna umożliwiająca klasyfikację rekordów na podstawie wartości zmiennych wejściowych. Jest ona analogiczna do regresji liniowej, lecz bazuje na przewidywanej zmiennej jakościowej zamiast na liczbowej.

Na przykład załóżmy, że operator telekomunikacyjny pogrupował bazę klientów wg wzorców korzystania z usług, tworząc cztery kategorie. Jeśli można użyć danych demograficznych do przewidywania członkostwa w grupie, można dostosować oferty dla indywidualnych potencjalnych klientów.

W tym przykładzie zastosowano strumień o nazwie *telco_custcat_discriminant.str*, który odwołuje się do pliku danych o nazwie *telco.sav*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *telco_custcat_discriminant.str* znajduje się w katalogu *streams*.

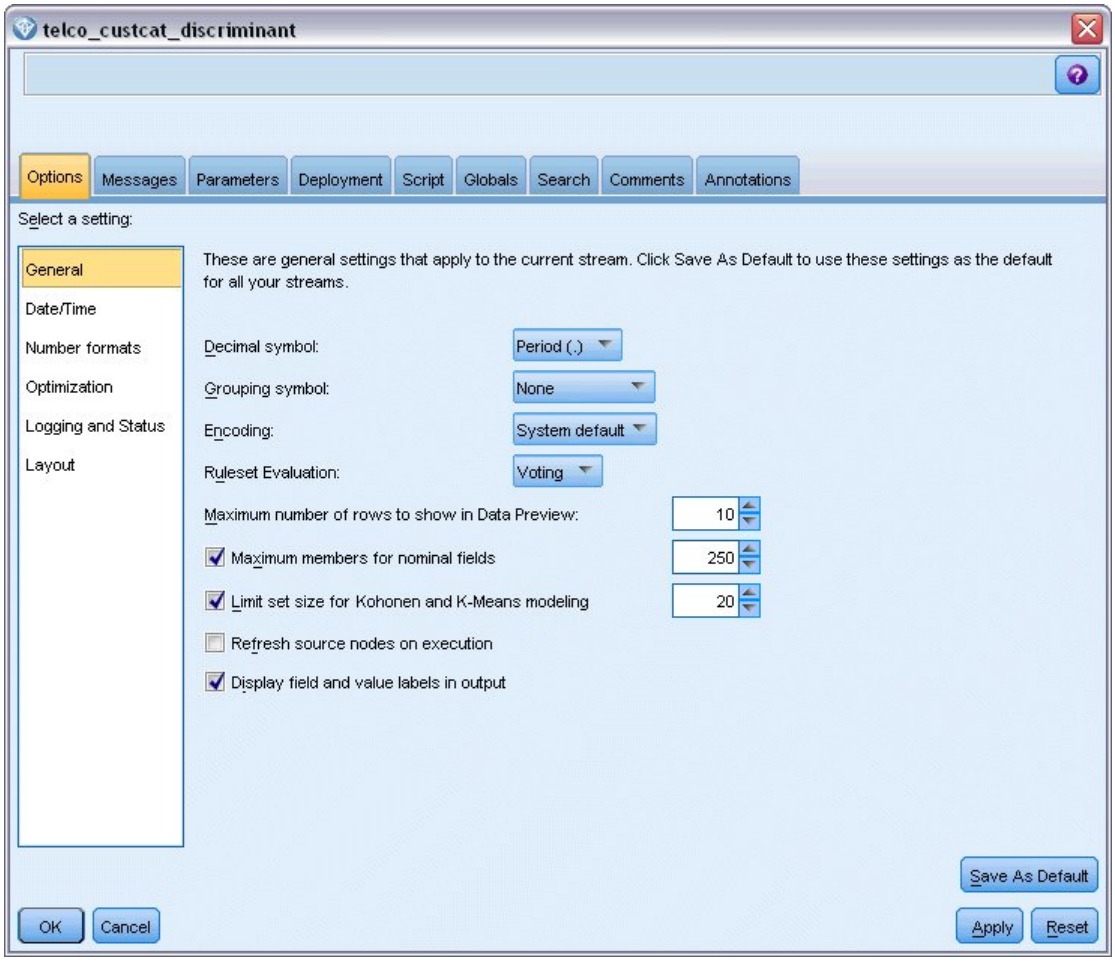
Ten przykład koncentruje się na użyciu danych demograficznych do przewidywania wzorców używania. Zmienna przewidywana *custcat* ma cztery możliwe wartości, które odpowiadają czterem grupom klientów w następujący sposób:

Wartość	**Etykieta**
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

Tworzenie strumienia

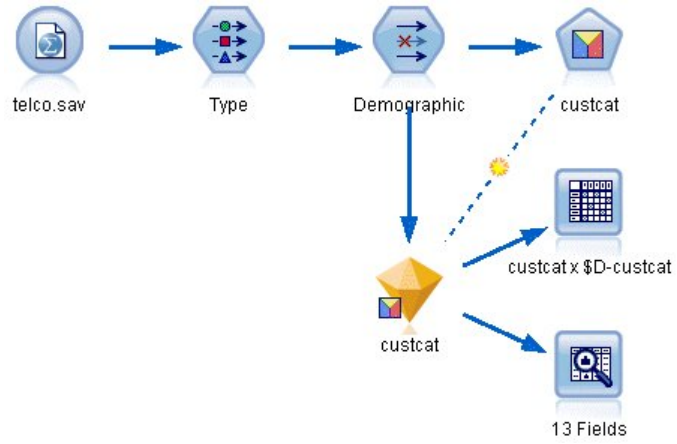
1. Najpierw ustaw właściwości strumienia, aby przedstawiały etykiety zmiennych i wartości w wynikach. Z menu wybierz kolejno następujące pozycje:

   Plik > Właściwości strumienia... > Opcje > Opcje ogólne
2. Upewnij się, że zaznaczona jest opcja **Wyświetlaj w wynikach etykiety zmiennych i wartości** i kliknij przycisk **OK**.



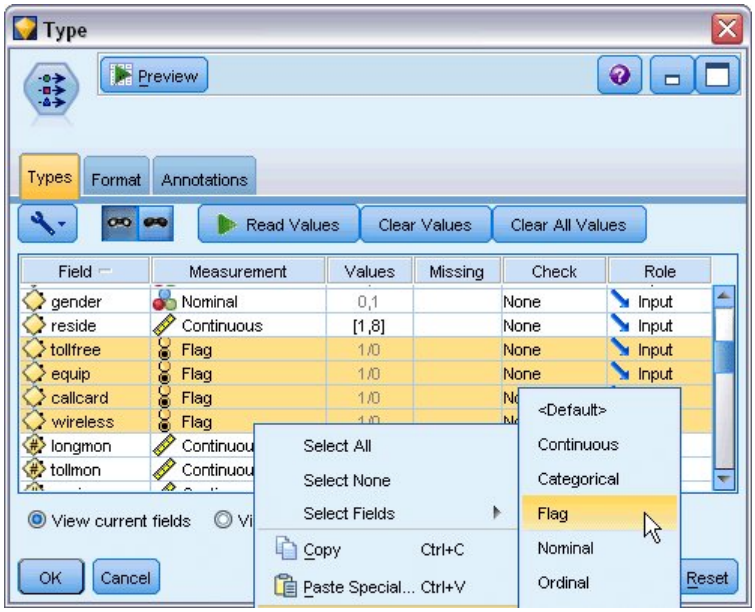
Rysunek 268. Właściwości strumienia

3. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *telco.sav* znajdujący się w folderze *Demos*.



Rysunek 269. Przykładowy strumień klasyfikujący klientów za pomocą analizy dyskryminacyjnej

a. Dodaj węzeł typu i kliknij przycisk **Odczytaj wartości**, upewniając się, że ustawiono poprawnie wszystkie poziomy pomiaru. Na przykład większość zmiennych z wartościami 0 i 1 można traktować jako flagi.

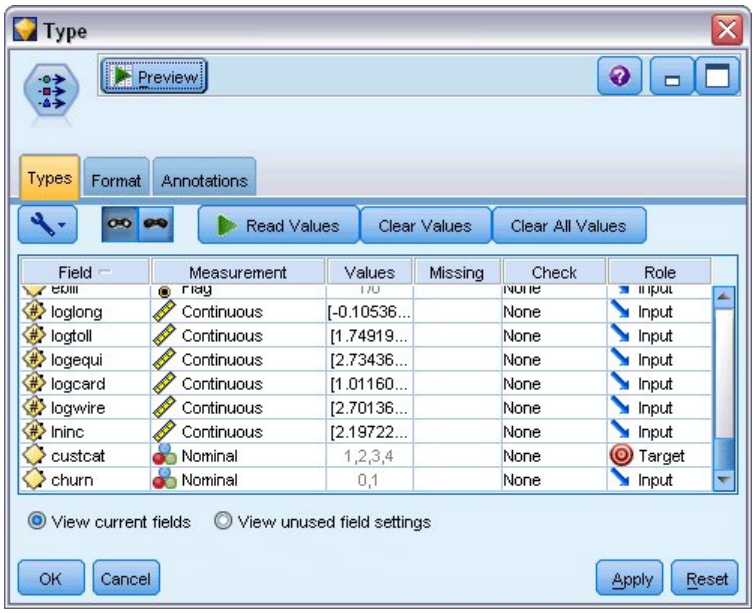


Rysunek 270. Ustawianie poziomu pomiaru dla wielu zmiennych

Wskazówka: Aby zmienić właściwości dla wielu zmiennych z podobnymi wartościami (takimi jak 0/1), kliknij nagłówek kolumny *Wartości*, aby posortować zmienne według wartości, a następnie przytrzymaj naciśnięty klawisz Shift, używając myszy lub klawiszy strzałek, aby wybrać wszystkie zmienne, które chcesz zmienić. Możesz następnie kliknąć wybrany zakres prawym klawiszem myszy, aby zmienić poziom pomiaru lub inne atrybuty dla wybranych zmiennych.

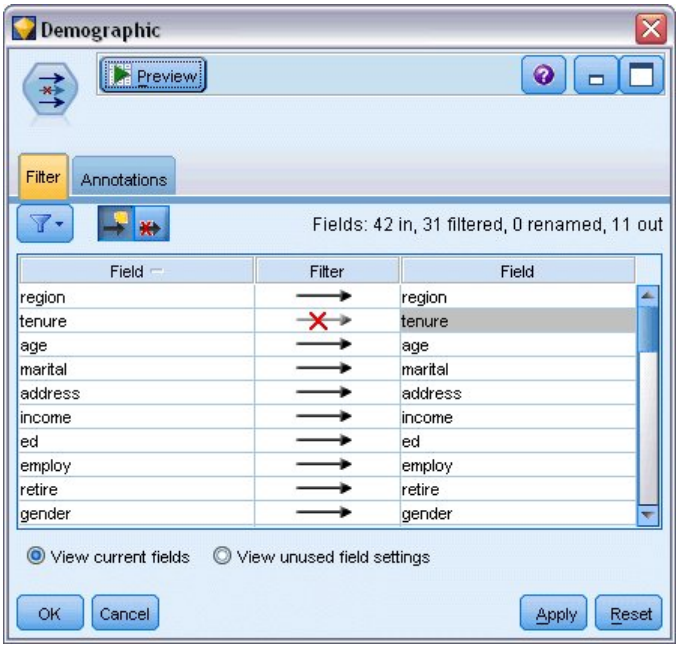
Należy zauważyć, że zmienna *gender* (płeć) bardziej prawidłowo jest uważana za zmienną z dwiema wartościami niż za flagę, więc należy pozostawić wartość kolumny Poziom pomiaru jako **Nominalna**.

b. Ustaw rolę zmiennej *custcat* na **Przewidywana**. Wszystkie pozostałe zmienne powinny mieć ustawioną rolę na **Dane wejściowe**.



Rysunek 271. Ustawianie roli zmiennej

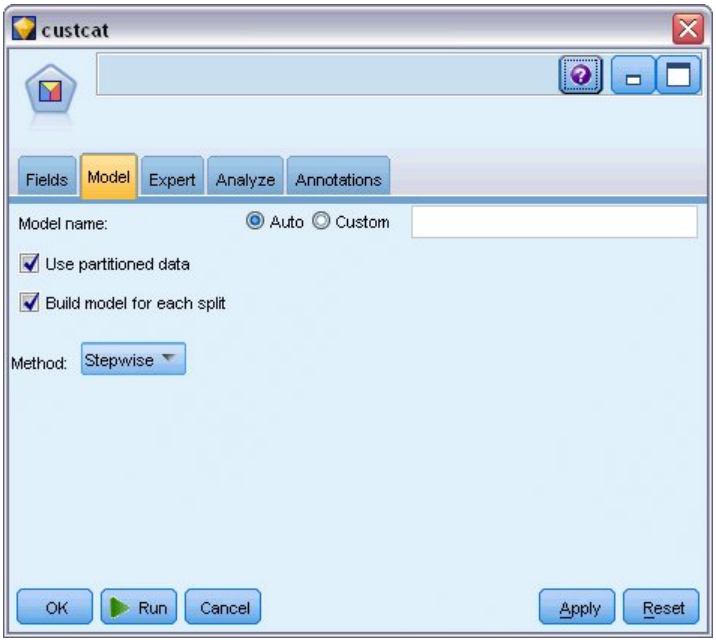
Ponieważ ten przykład koncentruje się na demografii, użyj węzła filtrowania do uwzględnienia tylko istotnych zmiennych (*region*, *age*, *marital*, *address*, *income*, *ed*, *employ*, *retire*, *gender*, *reside* i *custcat*). Inne zmienne można wyłączyć na potrzeby tej analizy.



Rysunek 272. Filtrowanie zmiennych demograficznych

(Można również zmienić rolę tych zmiennych na **Brak**, zamiast je wykluczać lub wybrać zmienne, których chcesz użyć w węźle modelowania).

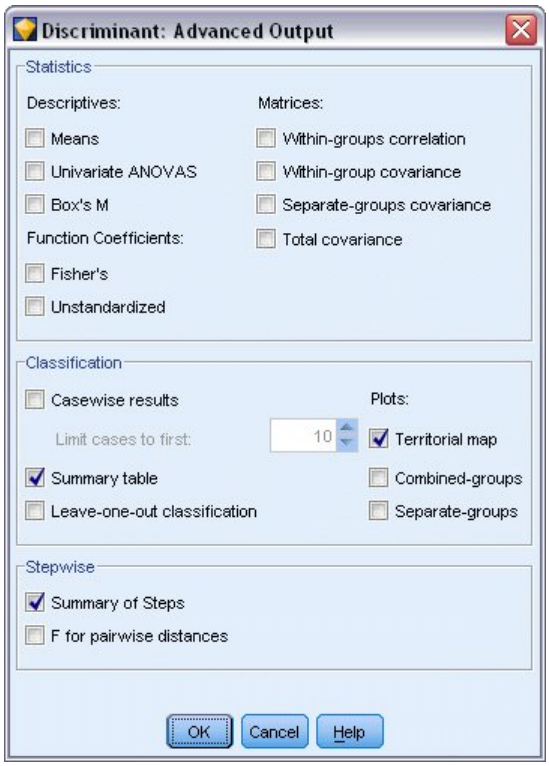
4. W węźle Analiza dyskryminacyjna kliknij kartę Model i wybierz metodę **Krokowa**.



Rysunek 273. Wybieranie opcji modelu

5. Na karcie Zaawansowany ustaw tryb na **Zaawansowany** i kliknij **Wynik**.

6. W oknie dialogowym Zaawansowane dane wyjściowe zaznacz opcje **Tabela podsumowań**, **Mapa terytorialna** i **Podsumowanie dla kolejnych kroków**, a następnie kliknij przycisk **OK**.

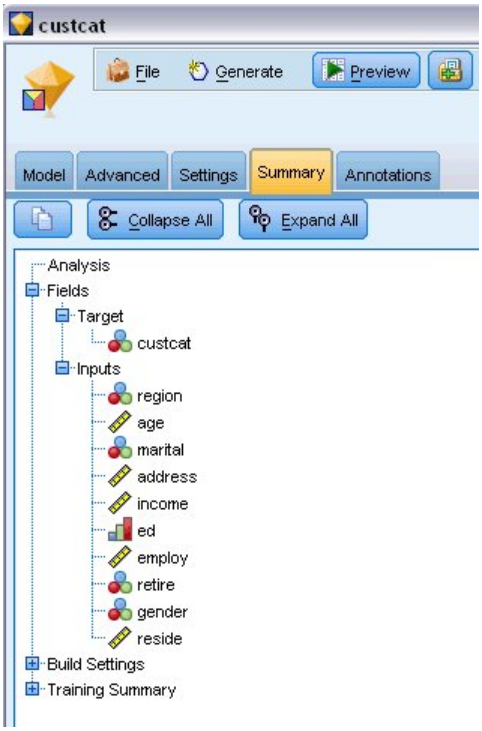


Rysunek 274. Wybieranie opcji wyników

Badanie modelu

1. Kliknij przycisk **Uruchom**, aby utworzyć model, który jest dodawany do strumienia i palety modeli w prawym górnym rogu. Aby wyświetlić jego szczegóły, dwukrotnie kliknij model użytkowy w strumieniu.

   Karta Podsumowanie przedstawia (między innymi) zmienną przewidywaną i pełną listę danych wejściowych (zmiennych predykcyjnych przekazanych do uwzględnienia).



Rysunek 275. Podsumowanie modelu przedstawiające zmienną przewidywaną i zmienne wejściowe

Aby uzyskać szczegółowe informacje o wynikach analizy dyskryminacyjnej:

2. Kliknij kartę Zaawansowane.
3. Kliknij przycisk Uruchom w zewnętrznej przeglądarce (pod kartą Model), aby wyświetlić wyniki w przeglądarce.

Analizowanie wyników analizy dyskryminacyjnej w celu klasyfikacji klientów usług telekomunikacyjnych

Krokowa analiza dyskryminacyjna

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

Rysunek 276. Zmienne poza analizą, krok 0

Kiedy istnieje wiele predyktorów, można użyć metody krokowej, aby automatycznie wybrać „najlepsze” zmienne do użycia w modelu. Metoda krokowa zaczyna się od modelu, który nie zawiera żadnych predyktorów. W każdym kolejnym kroku predyktor z najwyższą wartością *F-wprowadzenia*, która przekracza kryteria wprowadzania (domyślnie 3,84) jest dodawany do modelu.

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

Rysunek 277. Zmienne poza analizą, krok 3

Zmienne pozostawione poza analizą w ostatnim kroku mają wartości *F-wprowadzenia* mniejsze niż 3,84, więc żadna nie zostanie dodana.

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

Rysunek 278. Zmienne w analizie

Ta tabela przedstawia statystyki dla zmiennych, które są uwzględnione w analizie w każdym kroku. *Tolerancja* to proporcja wariancji zmiennej niewyjaśniona wpływem pozostałych zmiennych niezależnych w równaniu.Zmienna o małej tolerancji wnosi do modelu niewiele informacji i może spowodować problemy obliczeniowe.

Wartości *F-usunięcia* są użyteczne do opisu zachowania po usunięciu zmiennej z bieżącego modelu (przy założeniu pozostania pozostałych zmiennych). Wartość *F-usunięcia* dla wprowadzenia zmiennej jest taka sama jak wartość *F-wprowadzenia* w poprzednim kroku (pokazana w tabeli Zmienne poza analizą).

Ostrzeżenie dotyczące metod krokowych

Metody krokowe są wygodne, ale mają swoje ograniczenia. Należy być świadomym, że ponieważ metody krokowe wybierają modele wyłącznie na podstawie ujęcia statystycznego, mogą wybrać predyktory, które nie mają *znaczenia praktycznego*. Jeśli użytkownik posiada pewne doświadczenie w pracy z danymi i ma oczekiwania dotyczące tego, które predyktory są ważne, powinien wykorzystać tę wiedzę i unikać metod krokowych. Jeśli jednak dostępnych jest wiele predyktorów i nie wiadomo, od czego zacząć, uruchomienie analizy krokowej i dostosowanie wybranego modelu jest lepszym rozwiązaniem niż zupełny brak modelu.

Sprawdzanie dopasowania modelu

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

Rysunek 279. Wartości własne

Prawie cała wariancja wyjaśniona przez model jest spowodowana dwiema pierwszymi funkcjami dyskryminacyjnymi. Trzy funkcje są dopasowane automatycznie, ale z powodu małej wartości własnej można bez obaw zignorować trzecią

funkcję.

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

Rysunek 280. Lambda Wilksa

Wartość lambda Wilksa potwierdza, że tylko pierwsze dwie funkcje są przydatne. Dla każdego zestawu funkcji testowana jest hipoteza, że średnie wymienionych funkcji są równe w grupach. Test funkcji 3 ma wartość istotności większą niż 0,10, więc ta funkcja ma mały wkład w model.

Macierz struktury

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

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

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

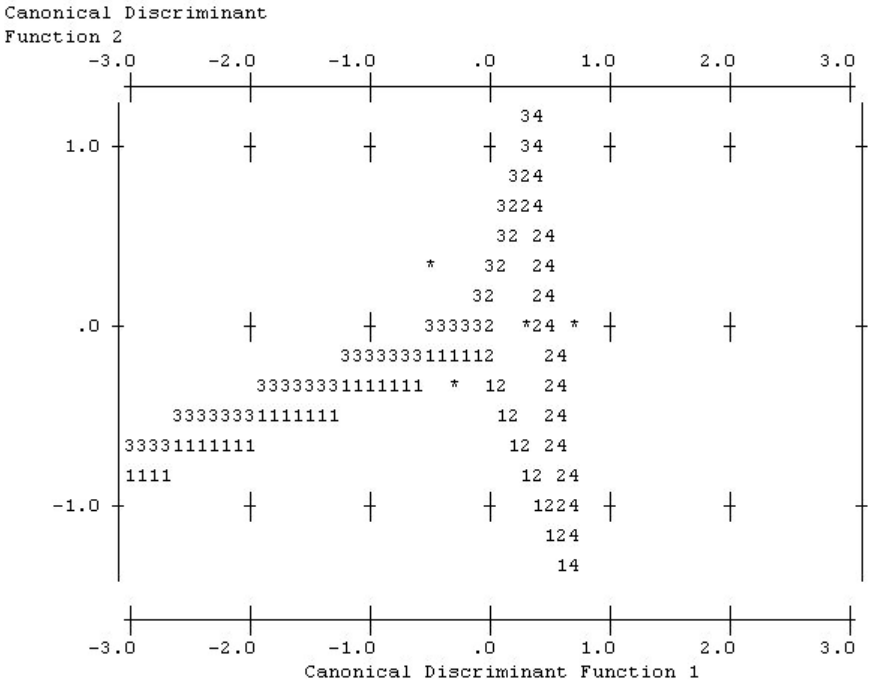
a. This variable not used in the analysis.

Rysunek 281. Macierz struktury

Jeśli istnieje więcej niż jedna funkcja dyskryminacyjna, gwiazdka (*) oznacza największą bezwzględną korelację każdej zmiennej z jedną z funkcji kanonicznych. W ramach każdej funkcji te oznaczone zmienne są następnie szeregowane według wielkości korelacji.

- *Level of education* ma dużą korelację z pierwszą funkcją i jest to jedyna zmienna najsilniej skorelowana z tą funkcją.
- Zmienne *Years with current employer*, *Age in years*, *Household income in thousands*, *Years at current address*, *Retired* i *Gender* są najbardziej skorelowane z drugą funkcją, mimo że zmienne *Gender* i *Retired* są słabiej skorelowane niż pozostałe. Inne zmienne oznaczają tę funkcję jako funkcję „stabilności”.
- Zmienne *Number of people in household* i *Marital status* są najsilniej skorelowane z trzecią funkcją dyskryminacyjną, ale jest to nieprzydatna funkcja, więc są to prawie nieprzydatne predyktory.

Mapa terytorialna



Rysunek 282. Mapa terytorialna

Mapa terytorialna pomaga zbadać relacje pomiędzy grupami i funkcjami dyskryminacyjnymi. W połączeniu z wynikami macierzy struktury zapewnia graficzną interpretację relacji między predyktorami i grupami. Pierwsza funkcja, przedstawiona na osi poziomej, oddziela grupę 4 (klientów usługi *Total service*) od innych. Ponieważ zmienna *Level of education* ma silną dodatnią korelację z pierwszą funkcją, sugeruje to, że klienci usługi *Total service* mają zazwyczaj wyższe wykształcenie. Druga funkcja rozdziela grupy 1 i 3 (klienci usług *Basic service* i *Plus service*). Klienci usługi *Plus service* mają większy staż pracy i są starsi niż klienci usługi *Basic service*. Klienci usługi *E-service* nie są rozdzieleni od innych, mimo że mapa sugeruje, że zazwyczaj mają wyższe wykształcenie ze średnim stażem pracy.

Zazwyczaj bliskość środków ciężkości oznaczonych gwiazdkami (*) do linii terytorialnych sugeruje, że separacja między grupami nie jest bardzo silna.

Wykres zawiera tylko pierwsze dwie funkcje dyskryminacyjne, ale ponieważ uznano trzecią funkcję za nieistotną, mapa terytorialna zapewnia wszechstronny wgląd w model dyskryminacyjny.

Wyniki klasyfikacji

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

a. 39.5% of original grouped cases correctly classified.

Rysunek 283. Wyniki klasyfikacji

Dzięki wartości lambda Wilksa wiadomo, że model działa lepiej niż przy zgadywaniu, ale należy sprawdzić klasyfikację, aby dowiedzieć się, o ile lepiej. Biorąc pod uwagę zaobserwowane dane, model pusty (tzn. bez predyktorów) zaklasyfikowałby wszystkich klientów do grupy modalnej, *Plus service*. W ten sposób model zerowy byłby poprawny w 281/1000 = 28,1% przypadków. Opracowany model uzyskuje o 11,4% więcej, czyli 39,5% klientów. Model uzyskuje w szczególności bardzo dobre wyniki przy identyfikacji klientów usługi *Total service*. Model ma jednak bardzo złe wyniki przy klasyfikacji klientów usługi *E-service*. Możliwe, że konieczne będzie znalezienie kolejnego predyktora, aby rozdzielić tych klientów.

Podsumowanie

Użytkownik utworzył model dyskryminacyjny klasyfikujący klientów w jednej z czterech predefiniowanych grup użycia usług na podstawie informacji demograficznych o każdym kliencie. Używając macierzy struktury i mapy terytorialnej, użytkownik zidentyfikował, które zmienne są najbardziej przydatne do segmentacji bazy klientów. Wyniki klasyfikacji wykazały też, że model ma słabe wyniki przy klasyfikacji klientów usługi *E-service*. Wymagane są dalsze poszukiwania, aby określić inną zmienną predyktora, która lepiej zaklasyfikuje tych klientów, ale w zależności od tego, co należy przewidzieć, model może być idealnie dopasowany do potrzeb użytkownika. Na przykład, jeśli użytkownikowi nie zależy na identyfikacji klientów usługi *E-service*, model może być wystarczająco dokładny. Może być tak w przypadku, gdy usługa E-service ma największe straty i daje mały zysk. Jeśli najwyższy zwrot z inwestycji pochodzi od klientów usługi *Plus service* lub *Total service*, model może zapewnić wymagane informacje.

Należy również zauważyć, że wyniki opierają się tylko na danych uczących. Aby ocenić, jak dobrze model uogólnia inne dane, należałoby użyć węzła podziału na podzbiory, aby zatrzymać podzbiór rekordów do celów testowania i walidacji.

Wyjaśnienia podstaw matematycznych metod modelowania używanych w programie IBM SPSS Modeler przedstawiono w publikacji IBM SPSS Modeler Algorithms Guide. Publikacja ta jest dostępna w katalogu *\Documentation* na dysku instalacyjnym.

Rozdział 22. Analizowanie danych przeżycia cenzurowanych interwałowo (Uogólnione modele liniowe)

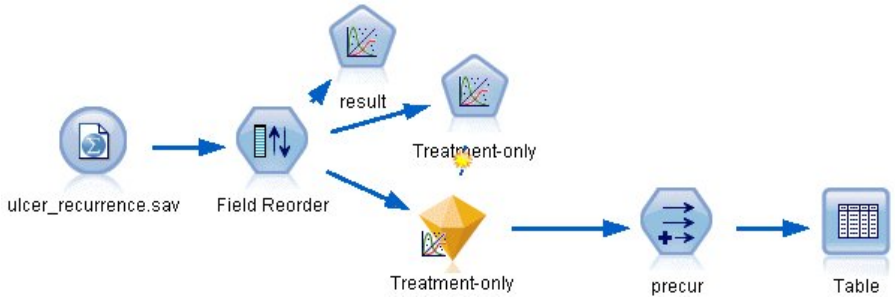
Podczas analizowania danych przeżycia cenzurowanych interwałowo (tzn. gdy nie jest znany dokładny czas zdarzenia, ale znany jest przedział czasowy), zastosowanie modelu Coxa dla zagrożeń zdarzeń w przedziałach powoduje powstanie komplementarnego modelu regresji log-log.

Częściowe informacje pochodzące z badania, którego celem było porównanie skuteczności dwóch terapii zapobiegających nawrotom wrzodów zgromadzono w pliku *ulcer_recurrence.sav*. Ten zbiór danych został przedstawiony i przeanalizowany w innej pracy[1]. Używając uogólnionych modeli liniowych, można zreplikować wyniki komplementarnego modelu regresji log-log.

W tym przykładzie zastosowano strumień o nazwie *ulcer_genlin.str*, który odwołuje się do pliku danych o nazwie *ulcer_recurrence.sav*. Plik danych znajduje się w folderze *Demos*, a plik strumienia w podfolderze *streams*.

Tworzenie strumienia

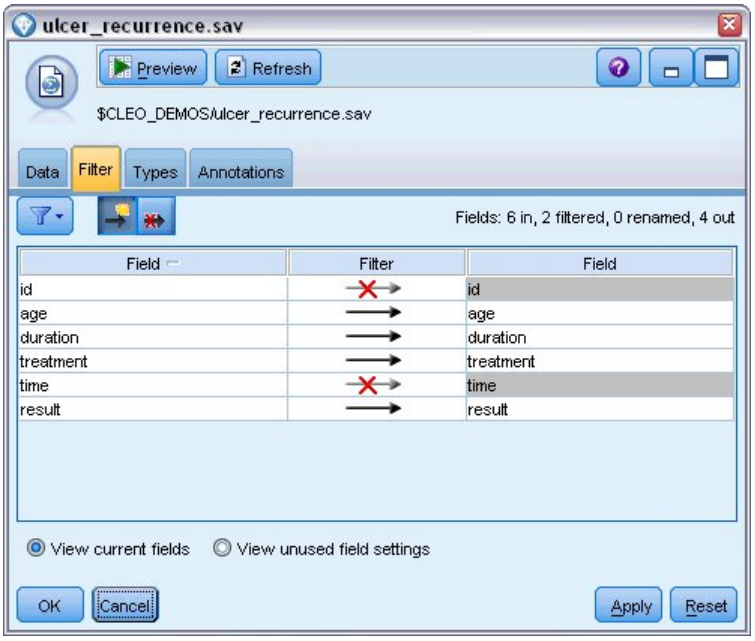
1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *ulcer_recurrence.sav* znajdujący się w folderze *Demos*.



Rysunek 284. Przykładowy strumień pozwalający przewidzieć nawroty wrzodów

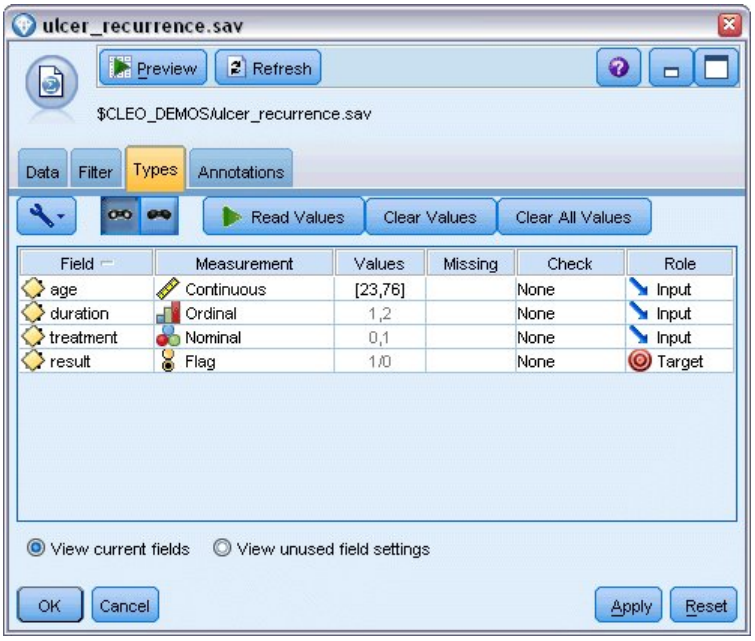
2. Na karcie Filtr węzła źródłowego odfiltruj zmienne *id* i *time*.

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



Rysunek 285. Filtrowanie niechcianych zmiennych

3. Na karcie Typy węzła źródłowego ustaw rolę zmiennej *result* na **Przewidywana** i ustaw jej poziom pomiaru na **Flaga**. Wynik 1 wskazuje na nawrót wrzodów. Wszystkie pozostałe zmienne powinny mieć ustawioną rolę na **Dane wejściowe**.
4. Kliknij przycisk **Odczytaj wartości**, aby zrealizować dane.

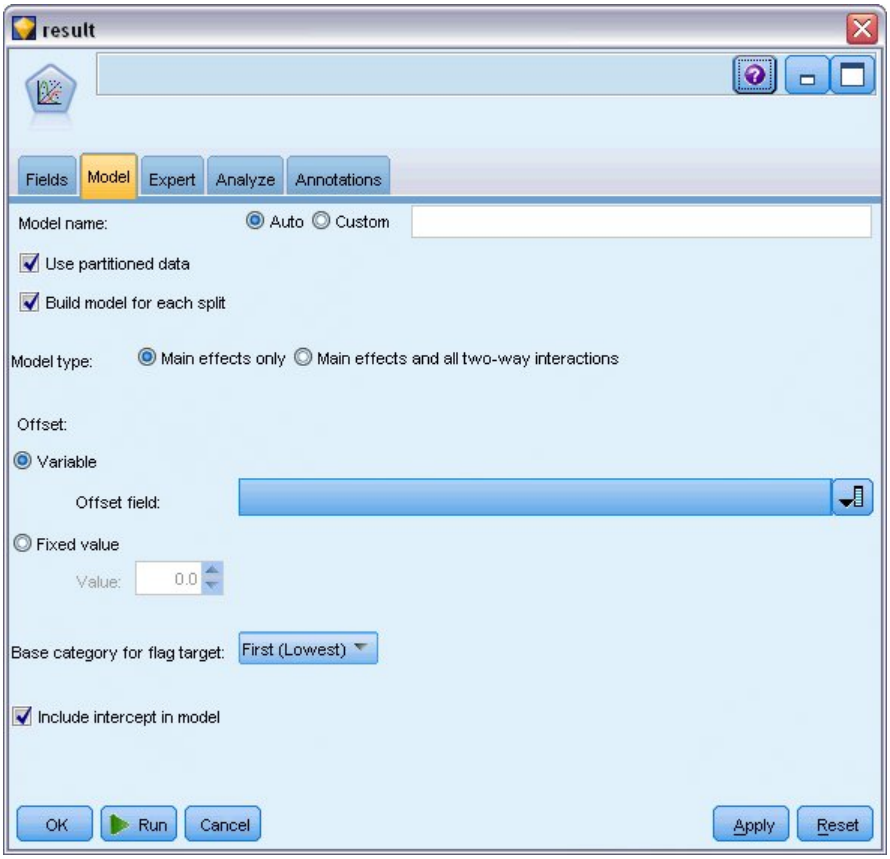


Rysunek 286. Ustawianie roli zmiennej

5. Dodaj węzeł Reorganizacja i określ kolejność danych wejściowych jako *duration*, *treatment* i *age*. Określa to kolejność wprowadzania zmiennych do modelu, co pozwoli zreplikować wyniki Colletta.

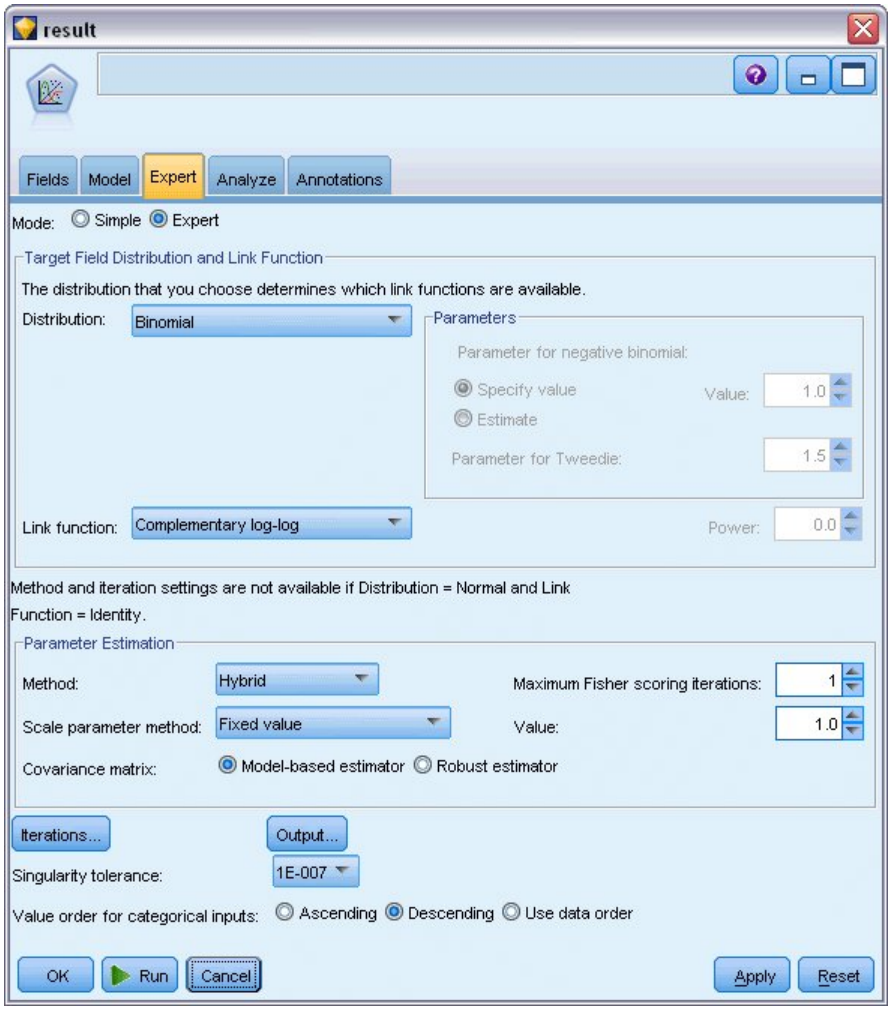
Rysunek 287. Zmiana kolejności zmiennych, aby były wprowadzane do modelu w wymagany sposób

6. Dołącz węzeł Modele uogólnione do węzła źródłowego. W węźle Modele uogólnione kliknij kartę **Model**.
7. Wybierz pozycję **Pierwsza (Najniższa wartość)** jako kategorię odniesienia dla zmiennej przewidywanej. Wskazuje to, że druga kategoria jest badanym zdarzeniem i jej wpływ na model jest interpretacją oszacowań parametrów. Predyktor ciągły z dodatnim współczynnikiem wskazuje zwiększone prawdopodobieństwo nawrotu przy rosnących wartościach predyktora. Kategorie predyktora nominalnego z większymi współczynnikami wskazują na większe prawdopodobieństwo nawrotu pod względem innych kategorii zbioru.



Rysunek 288. Wybieranie opcji modelu

8. Kliknij kartę **Zaawansowany** i wybierz opcję **Zaawansowany**, aby aktywować zaawansowane opcje modelowania.
9. Wybierz rozkład **Dwumianowy** oraz **Komplementarny log-log** jako funkcję łączenia.
10. Wybierz pozycję **Wartość ustalona** jako metodę oceny parametru skali i pozostaw wartość domyślną 1,0.
11. Wybierz opcję **Malejąco** jako kolejność kategorii dla czynników. Wskazuje to, że pierwszą kategorią dla każdego czynnika będzie jego kategoria odniesienia. Wpływ tego wyboru na model jest interpretacją oszacowań parametrów.



Rysunek 289. Wybieranie opcji zaawansowanych

12. Uruchom strumień, aby utworzyć model użytkowy, które jest dodawany do obszaru roboczego strumienia i palety modeli w prawym górnym rogu. Aby wyświetlić szczegóły modelu, kliknij go prawym przyciskiem myszy i wybierz opcję **Edytuj** lub **Przeglądaj**.

Testy efektów modelu

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

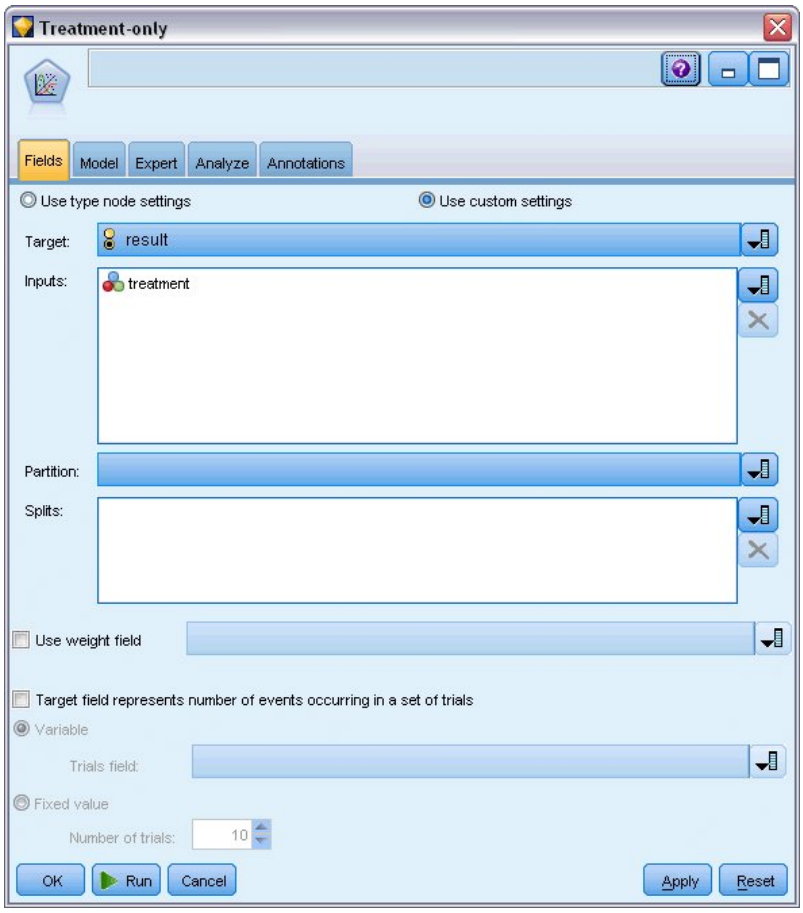
Dependent Variable: Result
Model: (Intercept), duration, treatment, age

Rysunek 290. Testy efektów modelu dla modelu efektów głównych

Żaden z efektów modelu nie jest statystycznie istotny. Wszystkie obserwowane różnice w efektach leczenia są przedmiotem zainteresowania klinicznego, więc dopasujemy model zredukowany tylko ze zmienną treatment jako składnikiem modelu.

Dopasowanie modelu tylko dla zmiennej treatment

1. Na karcie Zmienne węzła Modele uogólnione kliknij opcję **Użyj ustawień użytkownika**.
2. Wybierz *result* jako zmienną przewidywaną.
3. Wybierz zmienną *treatment* jako jedyne dane wejściowe.



Rysunek 291. Wybieranie opcji zmiennej

4. Uruchom strumień i otwórz wynikowy model użytkowy.

W modelu użytkowym wybierz kartę **Zaawansowane** i przewiń do końca.

Oszacowania parametrów

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

Dependent Variable: Result
Model: (Intercept), treatment

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

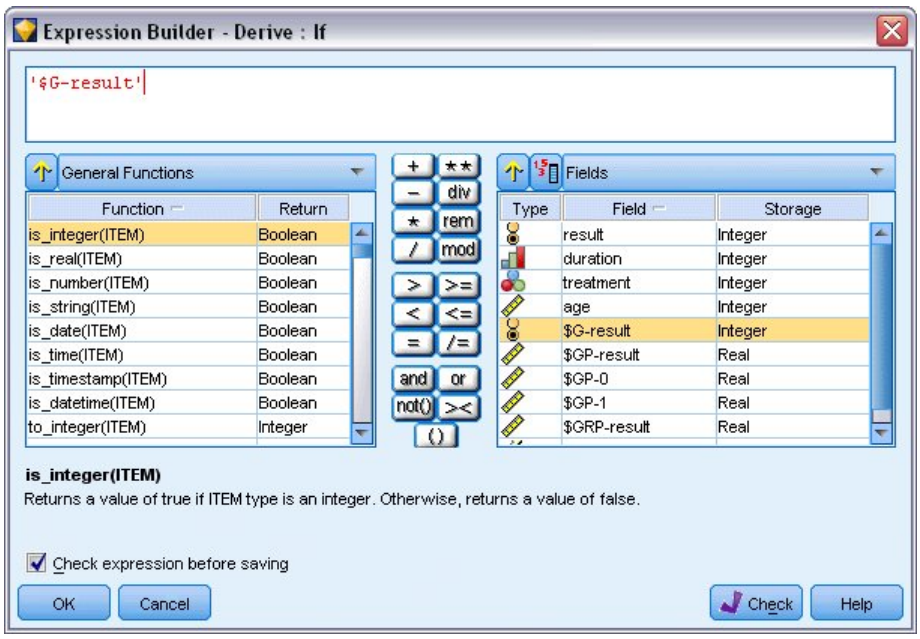
Rysunek 292. Oszacowania parametrów dla modelu uwzględniającego tylko zmienną treatment

Efekt leczenia (różnica predyktora liniowego pomiędzy dwoma poziomami leczenia, tzn. współczynnik dla *[treatment=1]*) wciąż nie jest statystycznie istotny, ale tylko sugeruje, że terapia *A [treatment=0]* może być lepsza niż terapia *B [treatment=1]*, ponieważ oszacowanie parametru dla terapii *B* jest większe niż dla terapii *A* i jest w ten sposób powiązane ze zwiększonym prawdopodobieństwem nawrotu w pierwszych 12 miesiącach. Predyktor liniowy (wyraz wolny + efekt terapii) jest oszacowaniem wyrażenia $\log(-\log(1-P(recur_{12,t}))$, gdzie $P(recur_{12,t})$ to prawdopodobieństwo nawrotu po 12 miesiącach dla zmiennej treatment t(=*A* lub *B*). Przewidywane prawdopodobieństwo jest generowane dla każdej obserwacji w zbiorze danych.

Przewidywany nawrót i prawdopodobieństwa przeżycia

Rysunek 293. Opcje ustawień węzła wyliczeń

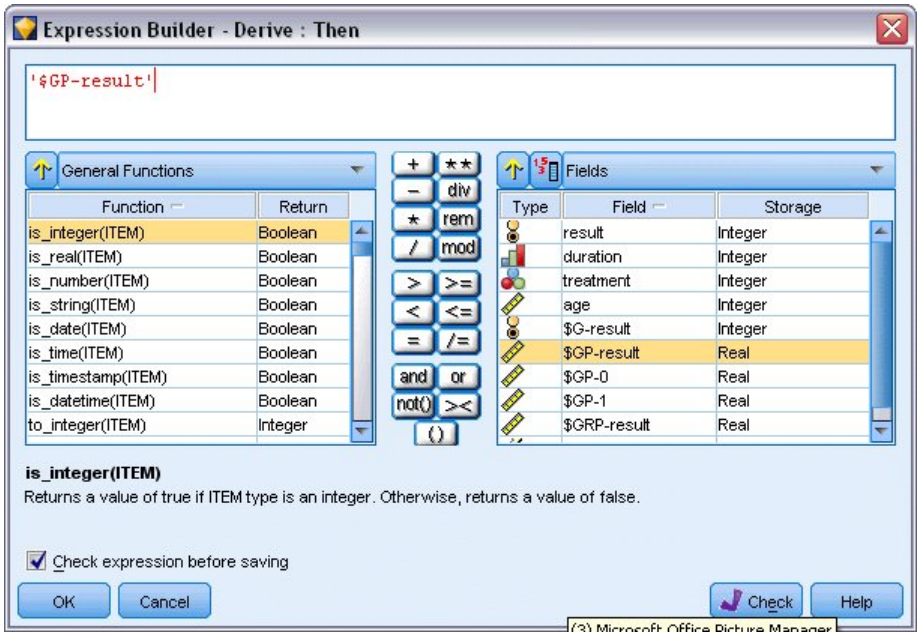
1. Dla każdego pacjenta model ocenia przewidywany wynik oraz prawdopodobieństwo tego przewidywanego wyniku. W celu zobaczenia przewidywanych prawdopodobieństw nawrotu skopiuj wygenerowany model na paletę i dołącz węzeł wyliczeń.
2. Na karcie Ustawienia wpisz precur jako zmienną wyliczaną.
3. Wybierz, aby wyliczyć zmienną jako **Warunkowe**.
4. Kliknij przycisk kalkulatora, aby otworzyć konstruktora wyrażeń dla wyrażenia **Jeżeli**.



Rysunek 294. Węzeł wyliczeń: Konstruktor wyrażeń dla wyrażenia Jeżeli

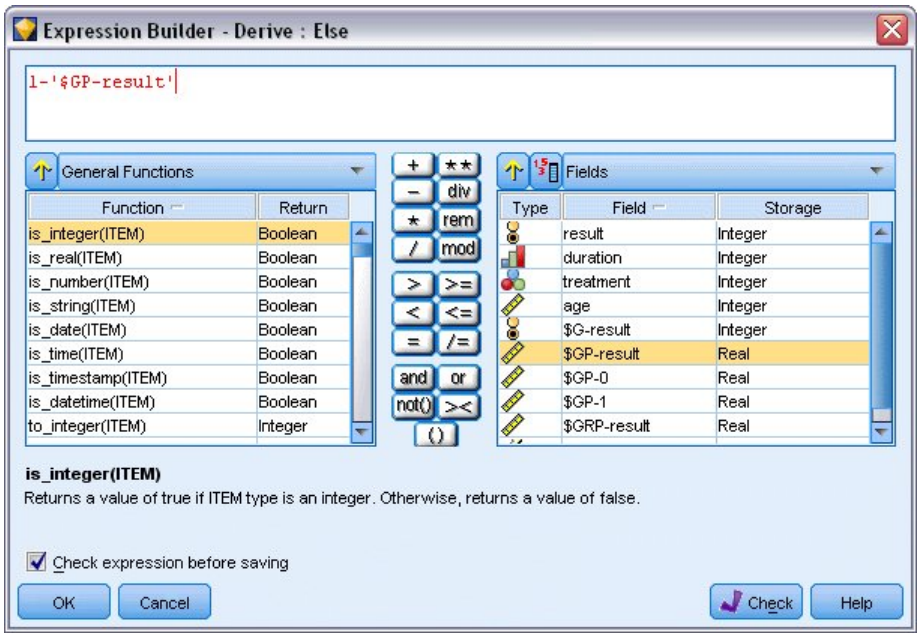
5. Wstaw zmienną *$G-result* do wyrażenia.
6. Kliknij przycisk **OK**.

   Zmienna wyliczana *precur* przyjmie wartość wyrażenia **To**, gdy zmienna *$G-result* jest równa 1, i wartość wyrażenia **Inaczej**, gdy jest równa 0.



Rysunek 295. Węzeł wyliczeń: Konstruktor wyrażeń dla wyrażenia To

7. Kliknij przycisk kalkulatora, aby otworzyć konstruktora wyrażeń dla wyrażenia **To**.
8. Wstaw zmienną *$GP-result* do wyrażenia.
9. Kliknij przycisk **OK**.

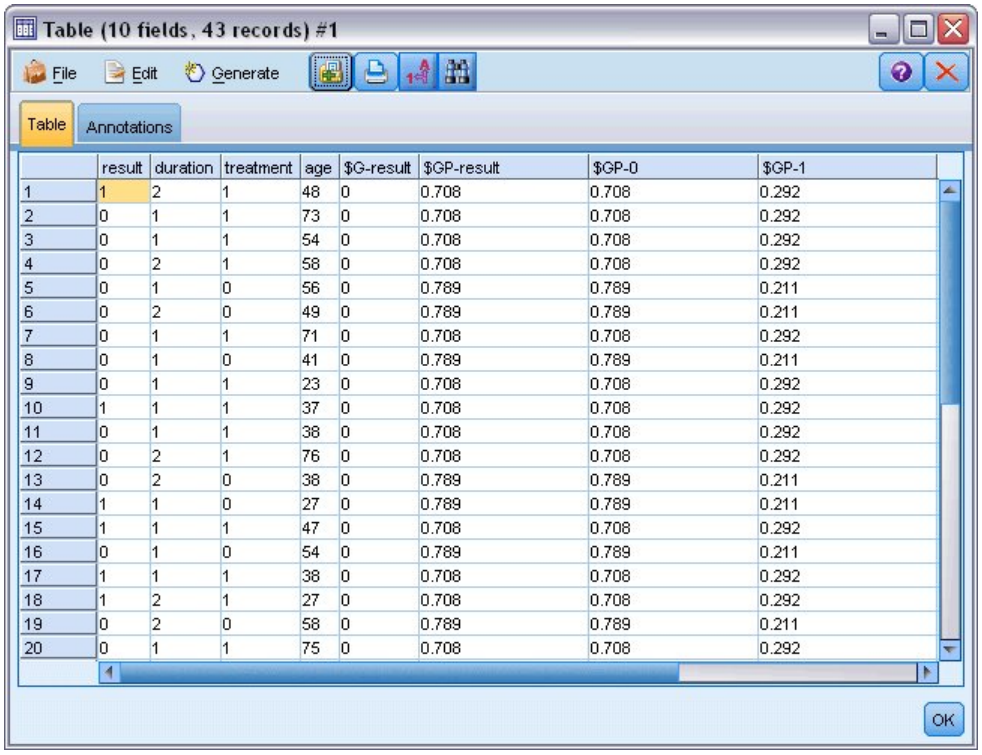


Rysunek 296. Węzeł wyliczeń: Konstruktor wyrażeń dla wyrażenia Inaczej

10. Kliknij przycisk kalkulatora, aby otworzyć konstruktora wyrażeń dla wyrażenia **Inaczej**.
11. Wpisz 1- w wyrażeniu, a następnie wstaw zmienną *$GP-result* w wyrażeniu.
12. Kliknij przycisk **OK**.

Rysunek 297. Opcje ustawień węzła wyliczeń

13. Dołącz węzeł tabeli do węzła wyliczeń i uruchom go.



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

Rysunek 298. Przewidywane prawdopodobieństwa

Istnieje oszacowane prawdopodobieństwo 0,211, że pacjenci przypisani do terapii *A* doświadczą nawrotu w pierwszych 12 miesiącach, i 0,292 dla terapii *B*. Należy zauważyć, że $1-P(recur_{12,\ t})$ to prawdopodobieństwo przeżycia po 12 miesiącach, co może być bardziej istotne dla analityka zajmującego się szansami przeżycia.

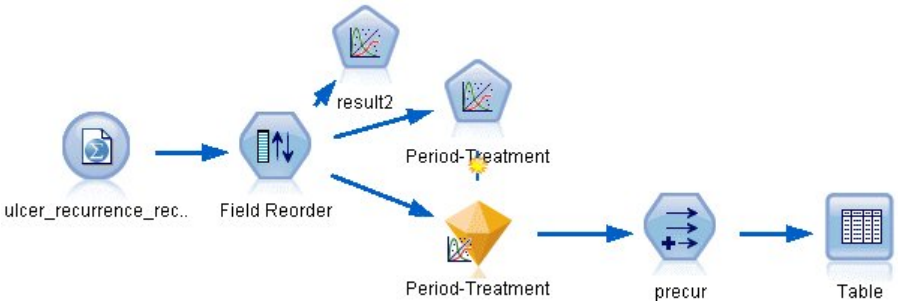
Modelowanie prawdopodobieństwa nawrotu według okresu

Problemem modelu jest to, że ignoruje informacje zgromadzone podczas pierwszego badania, tzn. wielu pacjentów nie doświadczyło nawrotu w pierwszych sześciu miesiącach. Lepszy model uwzględniłby binarną odpowiedź rejestrującą, czy zdarzenie wystąpiło podczas pierwszego okresu. Dopasowanie tego modelu wymaga rekonstrukcji oryginalnego zbioru danych, który można znaleźć w pliku *ulcer_recurrence_recoded.sav*. Plik zawiera dwie dodatkowe zmienne:

- *Period* — rejestruje, czy przypadek dotyczy pierwszego okresu badania, czy drugiego.
- *Result by period* — rejestruje, czy wystąpił nawrot dla określonego pacjenta w określonym okresie.

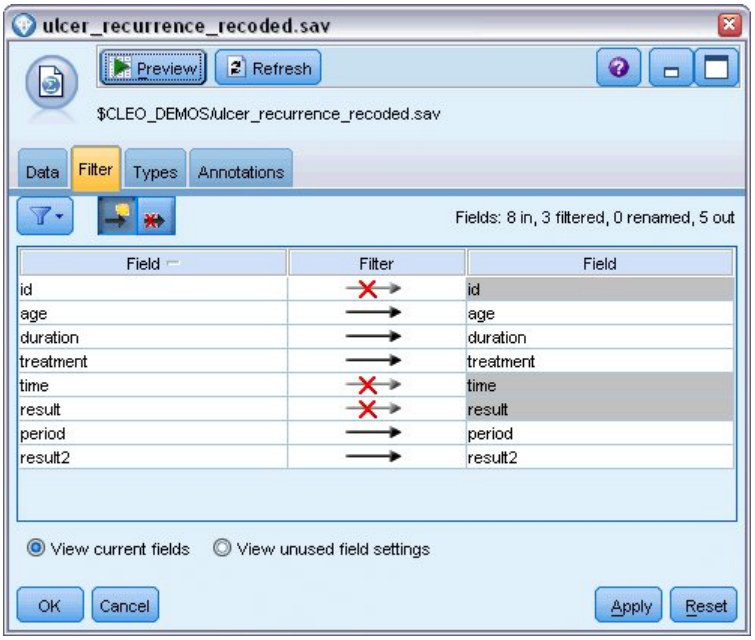
Każdy oryginalny przypadek (pacjent) daje jedną obserwację dla przedziału, w którym pozostaje w zbiorze ryzyka. W ten sposób np. pacjent 1 daje dwie obserwacje: jedną w pierwszym okresie badania, w którym nie wystąpił nawrót, i jedną w drugim okresie badania, w którym zarejestrowano nawrót. Z kolei pacjent 10 dostarcza jedną obserwację, ponieważ nawrót zarejestrowano w pierwszym okresie. Pacjenci 16, 28 i 34 opuścili badanie po sześciu miesiącach, więc dostarczają tylko jedną obserwację do nowego zbioru danych.

1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *ulcer_recurrence_recoded.sav* znajdujący się w folderze *Demos*.



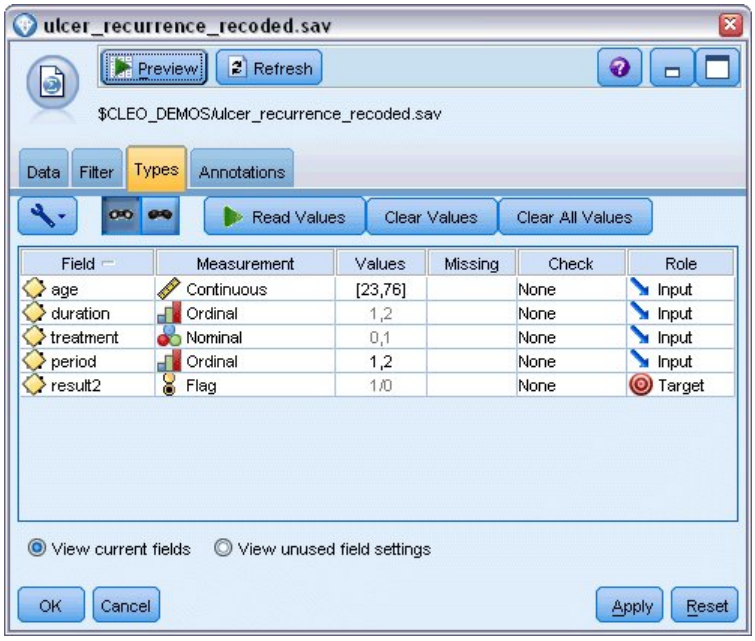
Rysunek 299. Przykładowy strumień pozwalający przewidzieć nawroty wrzodów

2. Na karcie Filtr węzła źródłowego odfiltruj zmienne *id*, *time* i *result*.



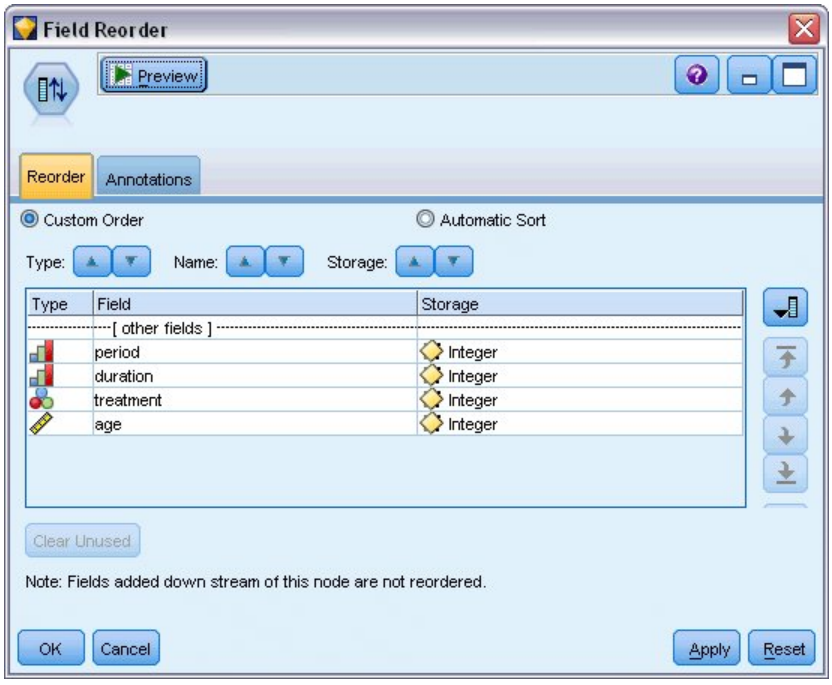
Rysunek 300. Filtrowanie niechcianych zmiennych

3. Na karcie Typy węzła źródłowego ustaw rolę zmiennej *result2* na **Przewidywana** i ustaw jej poziom pomiaru na **Flaga**. Wszystkie pozostałe zmienne powinny mieć ustawioną rolę na **Dane wejściowe**.



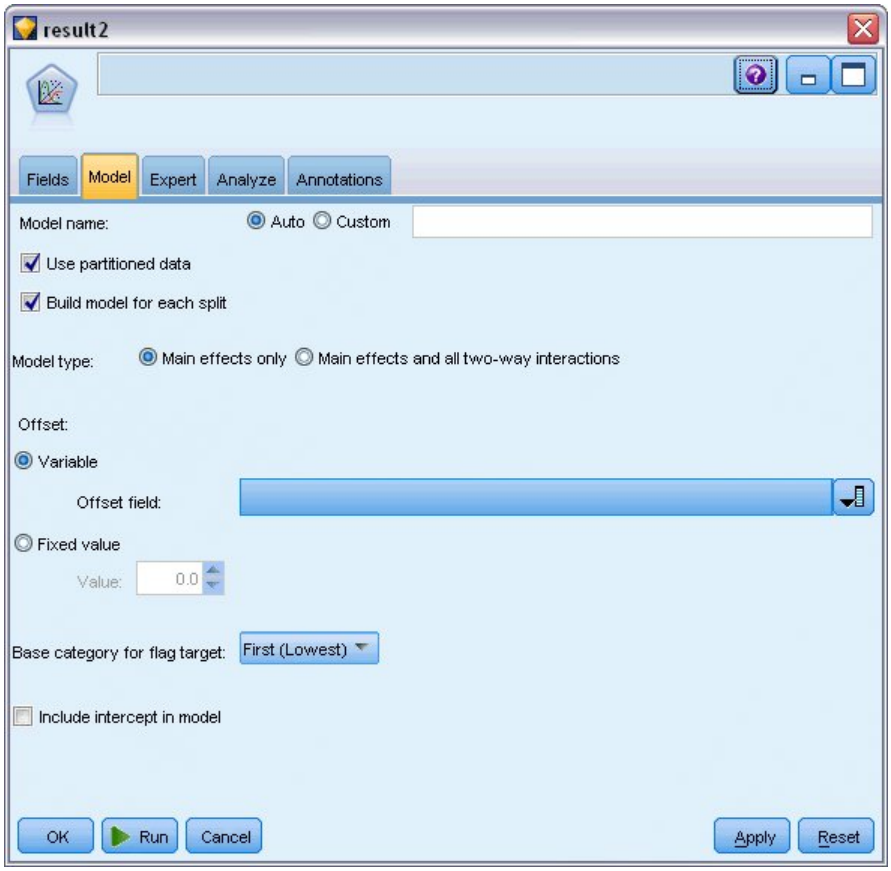
Rysunek 301. Ustawianie roli zmiennej

4. Dodaj węzeł Reorganizacja i określ kolejność danych wejściowych jako *period*, *duration*, *treatment* i *age*. Ustawienie zmiennej *period* jako pierwszej pozycji wejściowej (nie uwzględniając składnika stałej w modelu) pozwoli dopasować pełny zestaw zmiennych fikcyjnych, aby przechwycić skutki okresu.



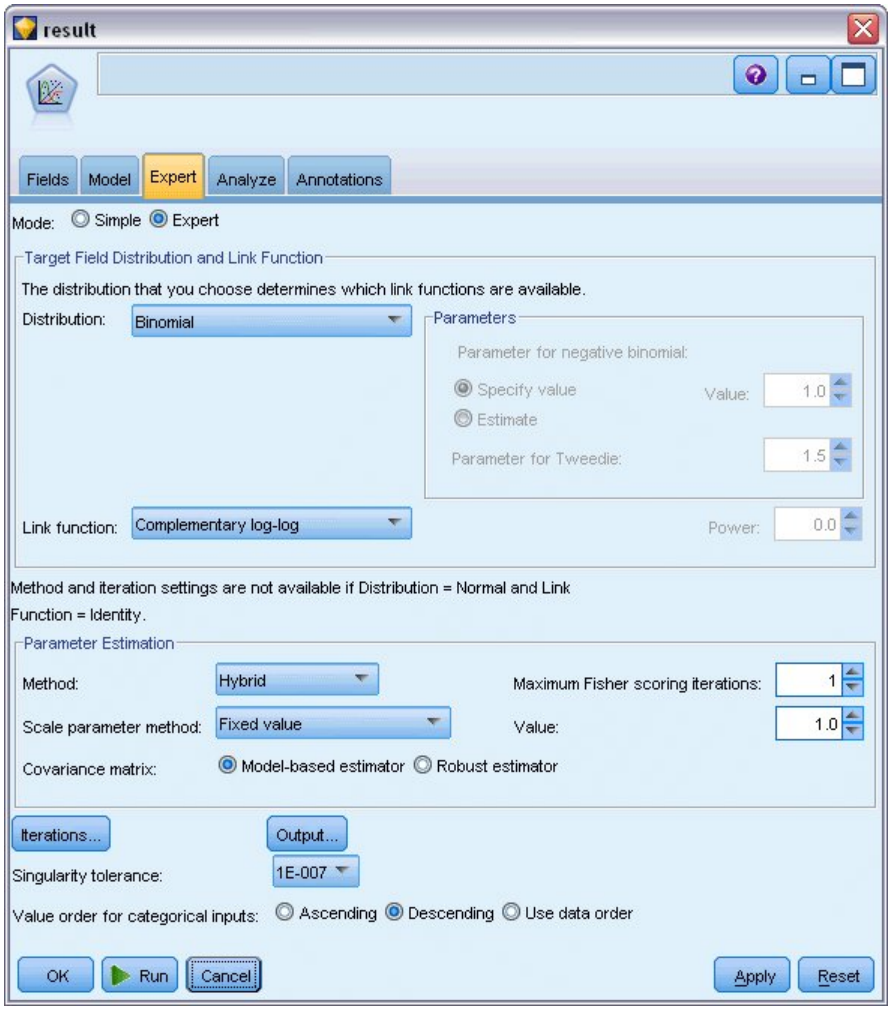
Rysunek 302. Zmiana kolejności zmiennych, aby były wprowadzane do modelu w wymagany sposób

5. W węźle Modele uogólnione kliknij kartę **Model**.



Rysunek 303. Wybieranie opcji modelu

6. Wybierz pozycję **Pierwsza (Najniższa wartość)** jako kategorię odniesienia dla zmiennej przewidywanej. Wskazuje to, że druga kategoria jest badanym zdarzeniem i jej wpływ na model jest interpretacją oszacowań parametrów.
7. Usuń zaznaczenie opcji **Uwzględnij wyraz wolny w modelu**.
8. Kliknij kartę **Zaawansowany** i wybierz opcję **Zaawansowany**, aby aktywować zaawansowane opcje modelowania.



Rysunek 304. Wybieranie opcji zaawansowanych

9. Wybierz rozkład **Dwumianowy** oraz **Komplementarny log-log** jako funkcję łączenia.
10. Wybierz pozycję **Wartość ustalona** jako metodę oceny parametru skali i pozostaw wartość domyślną 1,0.
11. Wybierz opcję **Malejąco** jako kolejność kategorii dla czynników. Wskazuje to, że pierwszą kategorią dla każdego czynnika będzie jego kategoria odniesienia. Wpływ tego wyboru na model jest interpretacją oszacowań parametrów.
12. Uruchom strumień, aby utworzyć model użytkowy, które jest dodawany do obszaru roboczego strumienia i palety modeli w prawym górnym rogu. Aby wyświetlić szczegóły modelu, kliknij go prawym przyciskiem myszy i wybierz opcję **Edytuj** lub **Przeglądaj**.

Testy efektów modelu

Source	Type III		
	Wald Chi-Square	df	Sig.
period	.464	1	.496
duration	.000	1	.988
treatment	.117	1	.732
age	.314	1	.575

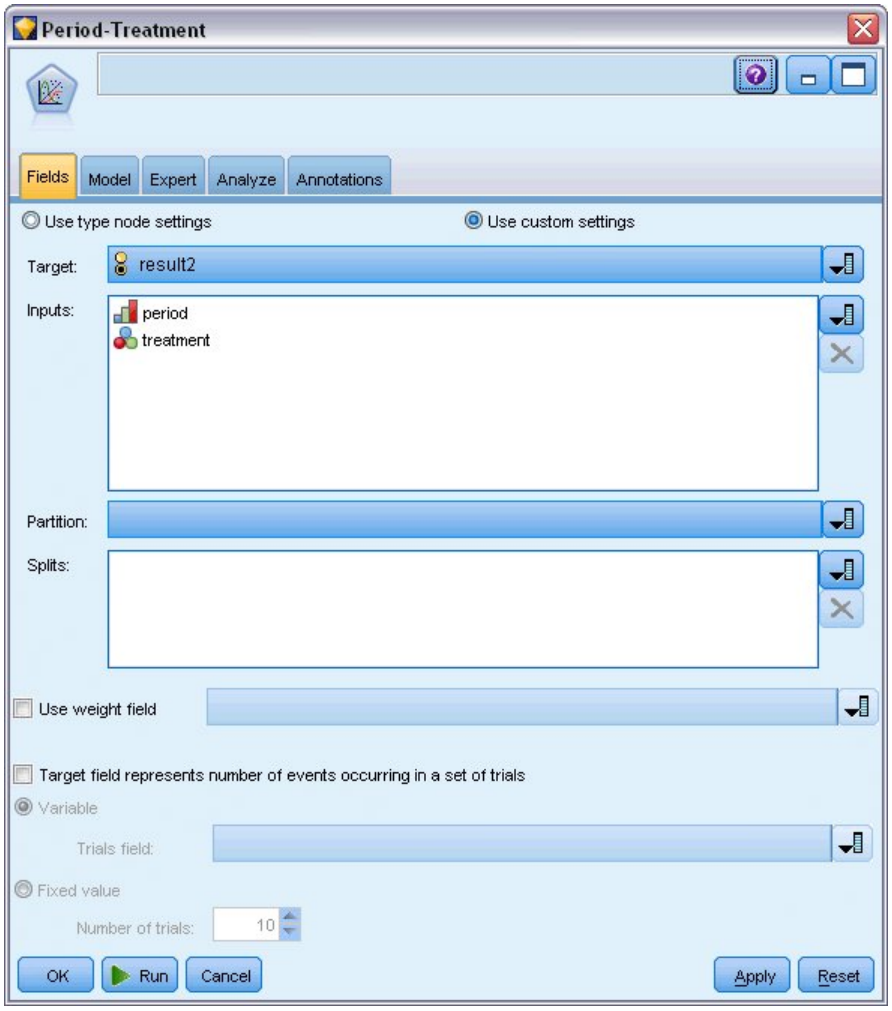
Dependent Variable: Result by period
Model: period, duration, treatment, age

Rysunek 305. Testy efektów modelu dla modelu efektów głównych

Żaden z efektów modelu nie jest statystycznie istotny. Wszystkie obserwowane różnice w okresie i efekty leczenia są przedmiotem zainteresowania klinicznego, więc dopasujemy model zredukowany tylko z tymi składnikami modelu.

Dopasowanie modelu zredukowanego

1. Na karcie Zmienne węzła Modele uogólnione kliknij opcję **Użyj ustawień użytkownika**.
2. Wybierz *result2* jako zmienną przewidywaną.
3. Wybierz zmienne *period* i *treatment* jako dane wejściowe.



Rysunek 306. Wybieranie opcji zmiennej

4. Wykonaj węzeł i przejrzyj wygenerowany model, a następnie skopiuj wygenerowany model do palety, dołącz węzeł tabeli i wykonaj go.

Oszacowania parametrów

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
[period=2]	-1.794	.5792	-2.929	-.659	9.597	1	.002
[period=1]	-2.206	.5912	-3.365	-1.047	13.926	1	.000
[treatment=1]	.195	.6279	-1.035	1.426	.097	1	.756
[treatment=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result by period
Model: period, treatment

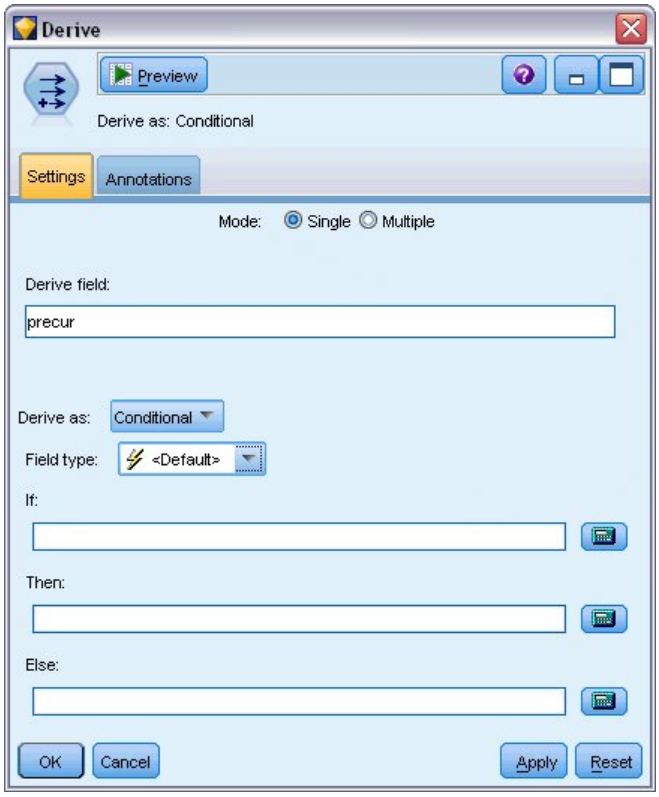
a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Rysunek 307. Oszacowania parametrów dla modelu uwzględniającego tylko zmienną treatment

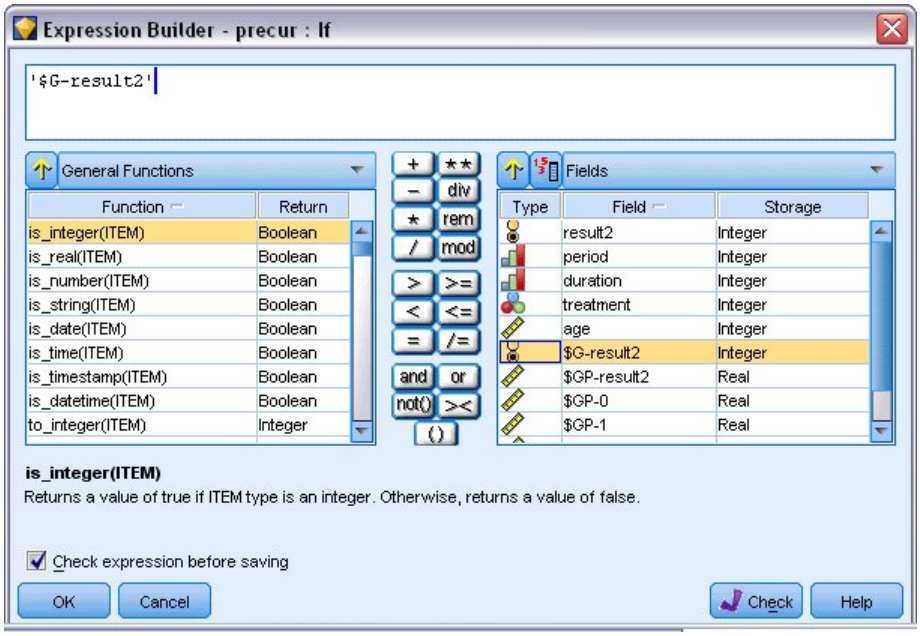
Efekt leczenia wciąż nie jest statystycznie istotny, ale tylko sugeruje, że terapia *A* może być lepsza niż *B*, ponieważ oszacowanie parametru dla terapii *B* jest powiązane ze zwiększonym prawdopodobieństwem nawrotu w pierwszych 12 miesiącach. Wartości okresu różnią się od 0 ze statystyczną istotnością, ale jest tak dlatego, że składnik składowej nie jest dopasowany. Efekt okresu (różnica pomiędzy wartościami predyktora liniowego dla *[period=1]* i *[period=2]*) nie jest statystycznie istotny, jak można zobaczyć w testach efektów modelu. Predyktor liniowy (efekt okresu + efekt leczenia) jest oszacowaniem wyrażenia $\log(-\log(1-P(recur_{p,\,t}))$, gdzie $P(recur_{p,\,t})$ to prawdopodobieństwo nawrotu w okresie p(=1 lub 2, reprezentującym 6 lub 12 miesięcy) przy terapii t(=*A* lub *B*). Przewidywane prawdopodobieństwo jest generowane dla każdej obserwacji w zbiorze danych.

Przewidywany nawrót i prawdopodobieństwa przeżycia



Rysunek 308. Opcje ustawień węzła wyliczeń

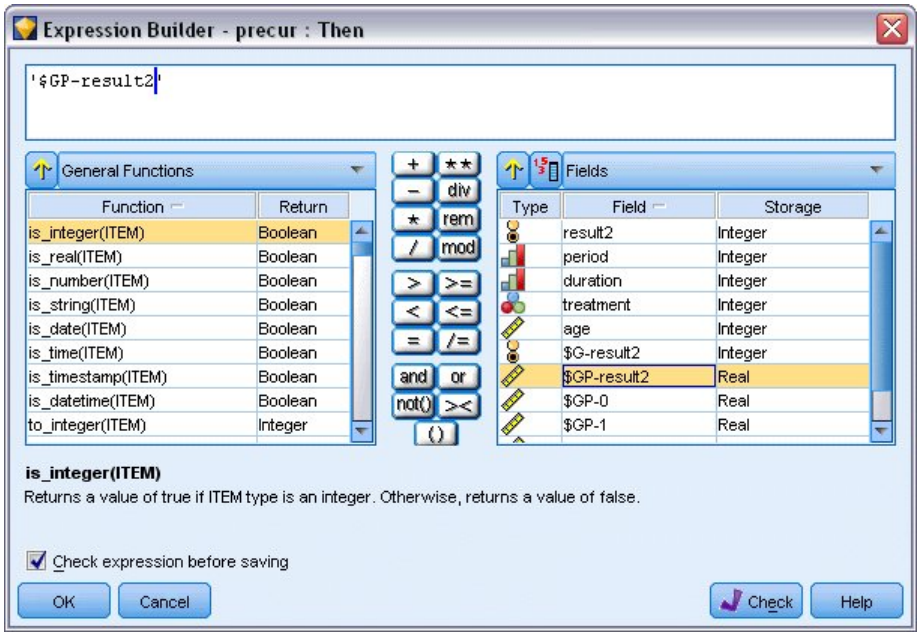
1. Dla każdego pacjenta model ocenia przewidywany wynik oraz prawdopodobieństwo tego przewidywanego wyniku. W celu zobaczenia przewidywanych prawdopodobieństw nawrotu skopiuj wygenerowany model na paletę i dołącz węzeł wyliczeń.
2. Na karcie Ustawienia wpisz precur jako zmienną wyliczaną.
3. Wybierz, aby wyliczyć zmienną jako **Warunkowe**.
4. Kliknij przycisk kalkulatora, aby otworzyć konstruktora wyrażeń dla wyrażenia **Jeżeli**.



Rysunek 309. Węzeł wyliczeń: Konstruktor wyrażeń dla wyrażenia Jeżeli

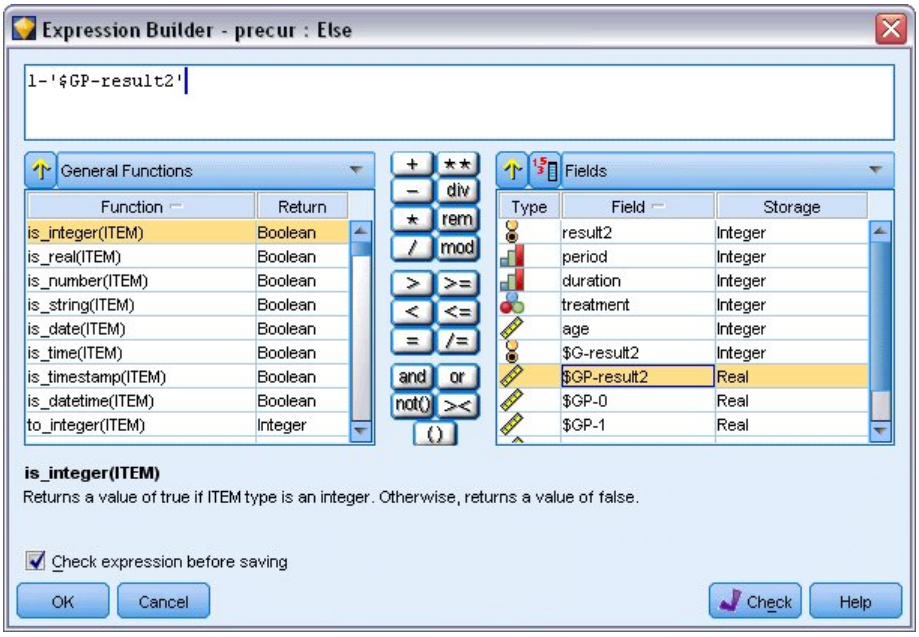
5. Wstaw zmienną *$G-result2* do wyrażenia.
6. Kliknij przycisk **OK**.

   Zmienna wyliczana *precur* przyjmie wartość wyrażenia **To**, gdy zmienna *$G-result2* jest równa 1, i wartość wyrażenia **Inaczej**, gdy jest równa 0.



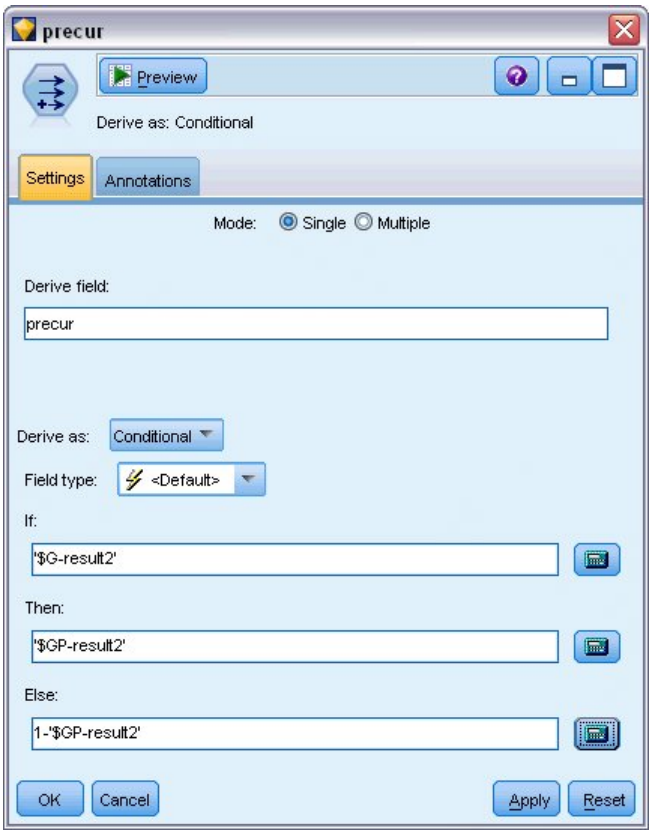
Rysunek 310. Węzeł wyliczeń: Konstruktor wyrażeń dla wyrażenia To

7. Kliknij przycisk kalkulatora, aby otworzyć konstruktora wyrażeń dla wyrażenia **To**.
8. Wstaw zmienną *$GP-result2* do wyrażenia.
9. Kliknij przycisk **OK**.



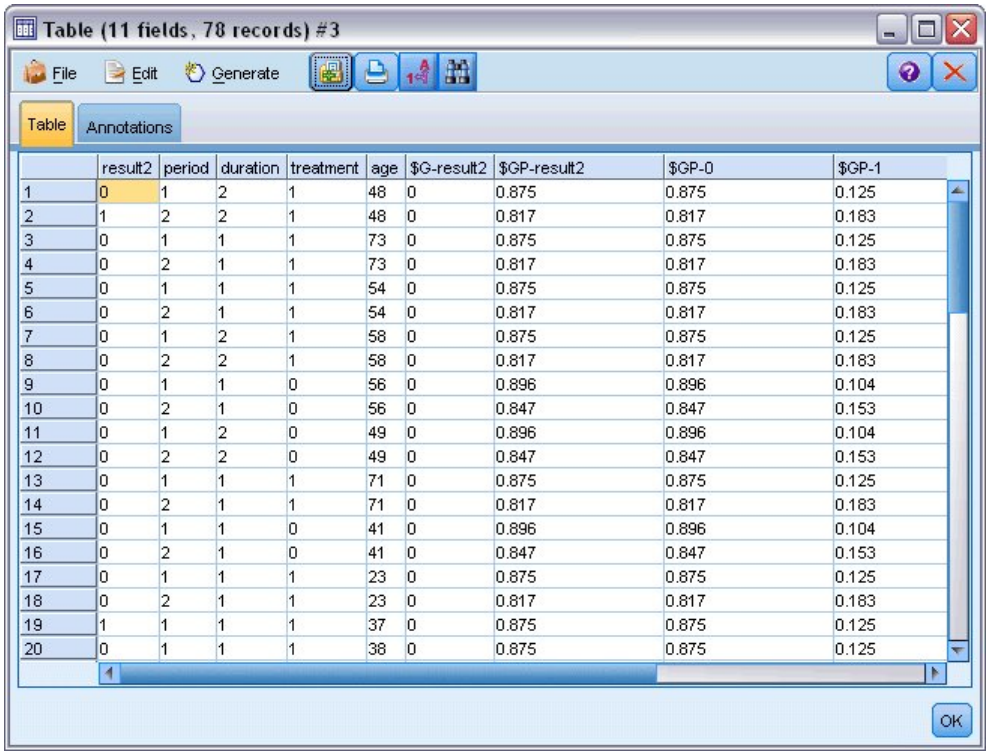
Rysunek 311. Węzeł wyliczeń: Konstruktor wyrażeń dla wyrażenia Inaczej

10. Kliknij przycisk kalkulatora, aby otworzyć konstruktora wyrażeń dla wyrażenia **Inaczej**.
11. Wpisz 1- w wyrażeniu, a następnie wstaw zmienną *$GP-result2* w wyrażeniu.
12. Kliknij przycisk **OK**.



Rysunek 312. Opcje ustawień węzła wyliczeń

13. Dołącz węzeł tabeli do węzła wyliczeń i uruchom go.



	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

Rysunek 313. Przewidywane prawdopodobieństwa

Tabela 3. Szacowane prawdopodobieństwa nawrotu

Terapia	**6 miesięcy**	**12 miesięcy**
A	0,104	0,153
B	0,125	0,183

Na podstawie szacowanych prawdopodobieństw nawrotu prawdopodobieństwo przeżycia po 12 miesiącach można oszacować jako $1-(P(recur_{1,\,t}) + P(recur_{2,\,t})\times(1-P(recur_{1,\,t})))$; więc dla każdej terapii:

A: 1 − (0,104 + 0,153*0,896)= 0,759

B: 1 − (0,125 + 0,183*0,875)= 0,715

co ponownie bez statystycznej istotności wskazuje na terapię *A* jako lepszą metodę leczenia.

Podsumowanie

Używając uogólnionych modeli liniowych, użytkownik dopasował szereg komplementarnych modeli regresji log-log do danych przeżycia ocenzurowanych interwałowo. Mimo że istnieje pewne wsparcie wyboru terapii *A*, osiągnięcie statystycznie istotnego wyniku może wymagać większego badania. Dostępne są jednak dalsze drogi badania istniejących danych.

- Być może warto dopasować ponownie model, uwzględniając efekty interakcji, zwłaszcza pomiędzy grupami *Period* i *Treatment*.

Wyjaśnienia podstaw matematycznych metod modelowania używanych w programie IBM SPSS Modeler przedstawiono w publikacji *IBM SPSS Modeler Algorithms Guide*.

Procedury pokrewne

Procedura Uogólnione modele liniowe jest przydatnym narzędziem dopasowania różnych modeli.

- Procedura Uogólnione równania estymujące rozszerza ogólny model liniowy w taki sposób, że pozwala na powtarzane pomiary.
- Procedura Liniowe modele mieszane pozwala dopasować modele do ilościowych zmiennych zależnych ze składnikiem losowym i/lub powtarzanymi pomiarami.

Zalecana literatura

Zapoznaj się z poniższymi tekstami, aby uzyskać więcej informacji o uogólnionych modelach liniowych:

Cameron, A. C. i P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, wyd. 2. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W. i J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P. i J. A. Nelder. 1989. *Generalized Linear Models*, wyd. 2. London: Chapman & Hall.

Rozdział 23. Korzystanie z regresji Poissona w celu analizy wskaźników uszkodzeń w transporcie (Uogólnione modele liniowe)

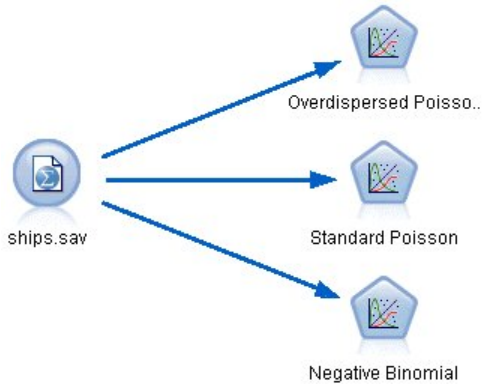
Uogólnionego modelu liniowego można użyć, aby dopasować regresję Poissona dla analizy danych liczebnościowych. Na przykład: zbiór danych przedstawiony i przeanalizowany w innej pracy [2], dotyczący uszkodzeń statków towarowych spowodowanych przez fale. Liczebność wypadków można zamodelować jako zmienną o rozkładzie Poissona, uwzględniając wartości predyktorów, a wynikowy model może pomóc określić, jakie typy statków są najbardziej podatne na uszkodzenia.

W tym przykładzie zastosowano strumień o nazwie *ships_genlin.str*, który odwołuje się do pliku danych o nazwie *ships.sav*. Plik danych znajduje się w folderze *Demos*, a plik strumienia w podfolderze *streams*.

Modelowanie surowych liczebności komórek może być mylące w tej sytuacji, ponieważ zmienna *Aggregate months of service* jest różna w zależności od typu statku. Takie zmienne, które mierzą wartość „narażenia" na ryzyko są obsługiwane przez uogólniony model liniowy jako zmienne przesunięcia. Dodatkowo regresja Poissona zakłada, że logarytm zmiennej zależnej jest liniowy w predyktorach. Dlatego też, aby używać uogólnionych modeli liniowych w celu dopasowania regresji Poissona do częstości wypadków, należy użyć zmiennej *Logarithm of aggregate months of service*.

Dopasowanie „rozproszonej" regresji Poissona

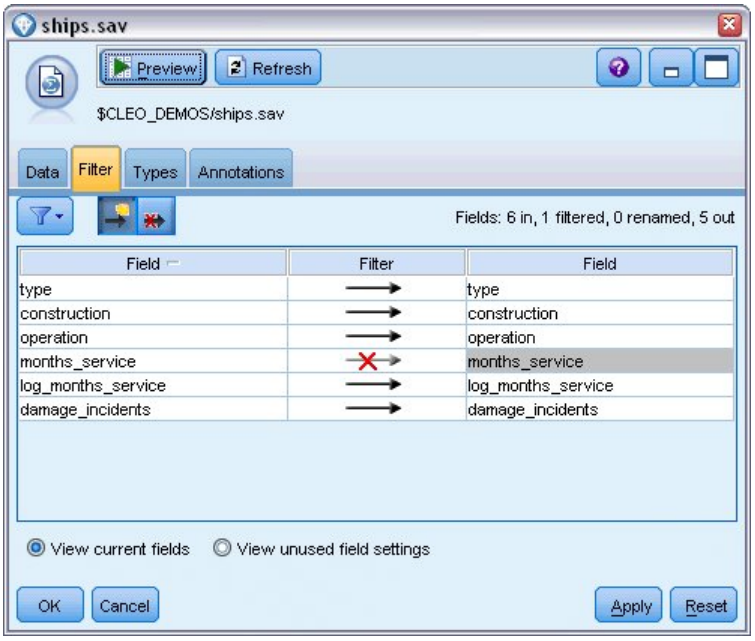
1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *ships.sav* znajdujący się w folderze *Demos*.



Rysunek 314. Przykładowy strumień analizujący wskaźniki uszkodzeń

2. Na karcie Filtr węzła źródłowego wyłącz zmienną *months_service*. Przekształcone logarytmicznie zmienne są zawarte w zmiennej *log_months_service*, która będzie używana w analizie.

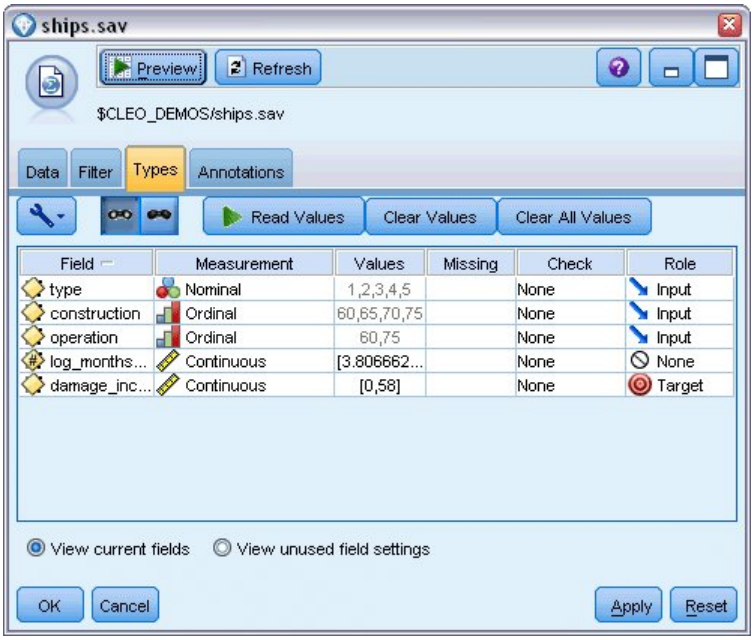
2. McCullagh, P. i J. A. Nelder. 1989. *Generalized Linear Models*, wyd. 2. London: Chapman & Hall.



Rysunek 315. Filtrowanie niepotrzebnej zmiennej

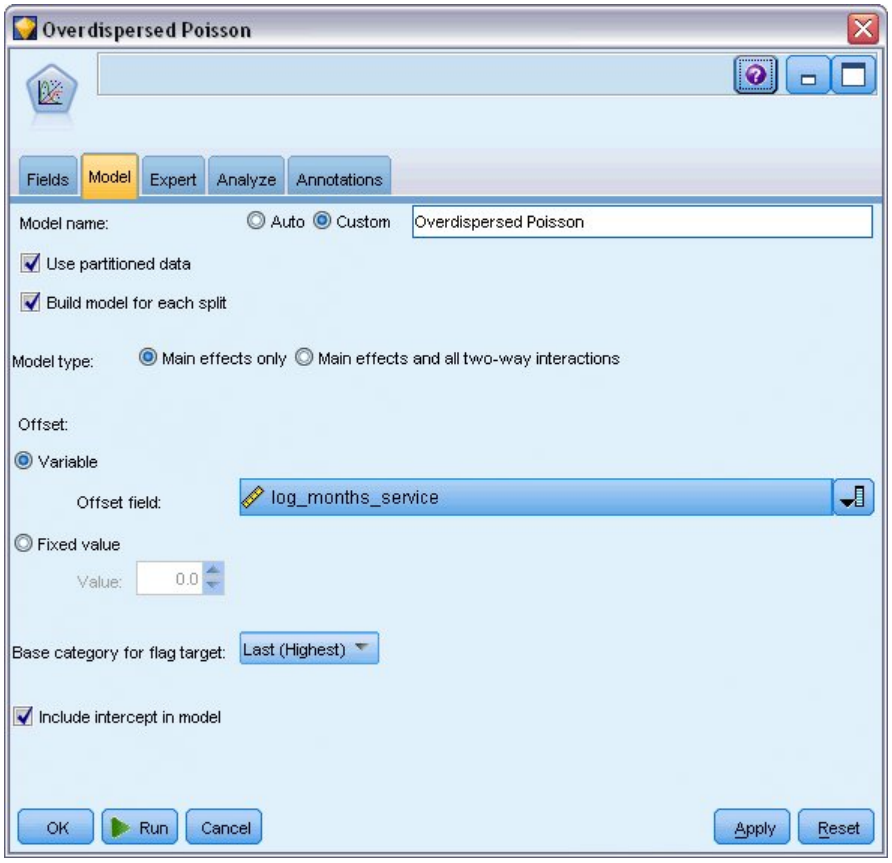
(Można również zmienić rolę tej zmiennej na **Brak** zamiast wykluczać ją lub wybrać zmienne, których chcesz użyć w węźle modelowania).

3. Na karcie Typy węzła źródłowego ustaw rolę zmiennej *damage_incidents* na **Przewidywana**. Wszystkie pozostałe zmienne powinny mieć ustawioną rolę na **Dane wejściowe**.
4. Kliknij przycisk **Odczytaj wartości**, aby zrealizować dane.



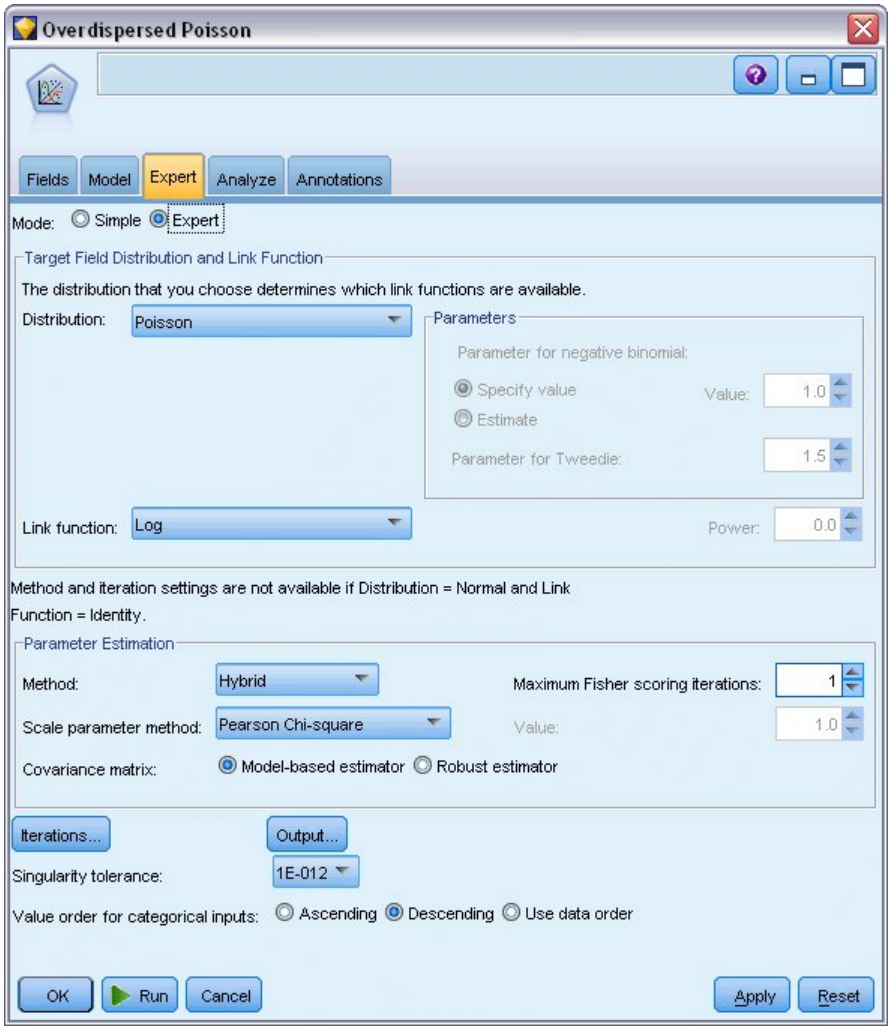
Rysunek 316. Ustawianie roli zmiennej

5. Dołącz węzeł Modele uogólnione do węzła źródłowego. W węźle Modele uogólnione kliknij kartę **Model**.
6. Ustaw *log_months_service* jako zmienną przesunięcia.



Rysunek 317. Wybieranie opcji modelu

7. Kliknij kartę **Zaawansowany** i wybierz opcję **Zaawansowany**, aby aktywować zaawansowane opcje modelowania.



Rysunek 318. Wybieranie opcji zaawansowanych

8. Wybierz rozkład **Poissona** dla odpowiedzi oraz **Log** jako funkcję łączenia.
9. Wybierz metodę **Chi-kwadrat Pearsona** jako metodę oceny parametru skali. Zazwyczaj zakłada się, że parametr skali jest równy 1 w regresji Poissona, ale McCullagh and Nelder używają oszacowania chi-kwadrat Pearsona do uzyskania bardziej konserwatywnych oszacowań wariancji oraz poziomów istotności.
10. Wybierz opcję **Malejąco** jako kolejność kategorii dla czynników. Wskazuje to, że pierwszą kategorią dla każdego czynnika będzie jego kategoria odniesienia. Wpływ tego wyboru na model jest interpretacją oszacowań parametrów.
11. Kliknij przycisk **Uruchom**, aby utworzyć model użytkowy, które jest dodawany do obszaru roboczego strumienia i palety modeli w prawym górnym rogu. Aby wyświetlić szczegóły modelu, kliknij model użytkowy prawym przyciskiem myszy i wybierz opcję **Edytuj** lub **Przeglądaj**, a następnie kliknij kartę **Zaawansowane**.

Statystyki dobroci dopasowania

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	22.883	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	25.000	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Rysunek 319. Statystyka dobroci dopasowania

Tabela statystyki dobroci dopasowania udostępnia miary, które są przydatne do porównywania konkurujących modeli. Dodatkowo współczynnik *Wartość/df* dla statystyk Dewiancja i Chi-kwadrat Pearsona zapewnia powiązane oszacowania parametru skali. Te wartości powinny być w pobliżu 1,0 dla regresji Poissona. Fakt, że wartość jest większa niż 1,0 wskazuje, że przydatne może być dostosowanie rozproszonego modelu.

Test typu omnibus

Likelihood Ratio Chi-Square	df	Sig.
107.633	8	.000

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. Compares the fitted model against the intercept-only model.

Rysunek 320. Test typu omnibus

Test typu omnibus to test ilorazu wiarygodności chi-kwadrat bieżącego modelu w porównaniu do modelu zerowego (w tym przypadku wyrazu wolnego). Wartość istotności mniejsza niż 0,05 wskazuje, że bieżący model ma lepsze wyniki niż model zerowy.

Testy efektów modelu

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	2138.657	1	.000
type	15.415	4	.004
construction	17.242	3	.001
operation	6.249	1	.012

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

Rysunek 321. Testy efektów modelu

Każdy składnik modelu jest testowany pod kątem tego, czy ma jakiś efekt. Składniki z wartością istotności mniejszą niż 0,05 mają pewien zauważalny efekt. Każdy ze składników efektów głównych ma swój wkład w model.

Oszacowania parametrów

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-6.406	.2828	-6.960	-5.852	513.238	1	.000
[type=5]	.326	.3067	-.276	.927	1.127	1	.288
[type=4]	-.076	.3779	-.817	.665	.040	1	.841
[type=3]	-.687	.4279	-1.526	.151	2.581	1	.108
[type=2]	-.543	.2309	-.996	-.091	5.536	1	.019
[type=1]	0[a]	.	.	.	.	.	.
[construction=75]	.453	.3032	-.141	1.048	2.236	1	.135
[construction=70]	.818	.2208	.386	1.251	13.743	1	.000
[construction=65]	.697	.1946	.316	1.079	12.835	1	.000
[construction=60]	0[a]	.	.	.	.	.	.
[operation=75]	.384	.1538	.083	.686	6.249	1	.012
[operation=60]	0[a]	.	.	.	.	.	.
(Scale)	1.691[b]						

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

Rysunek 322. Oszacowania parametrów

Tabela oszacowania parametrów podsumowuje efekt każdego predyktora. Podczas gdy interpretacja współczynników w tym modelu jest trudna z powodu charakteru funkcji łączenia, znaki współczynników dla współzmiennych i powiązanych wartości współczynników dla poziomów czynników mogą zapewnić istotne informacje o efektach predyktorów na model.

- Dla współzmiennych dodatnie (ujemne) współczynniki wskazują na dodatnie (odwrotne) relacje pomiędzy predyktorami i wynikiem. Rosnąca wartość współzmiennej z dodatnim współczynnikiem odpowiada rosnącemu współczynnikowi uszkodzeń.
- Dla czynników poziom czynnika z większym współczynnikiem wskazuje na większą częstość uszkodzeń. Znak współczynnika dla poziomu czynnika zależy od efektu poziomu czynnika powiązanego z kategorią odniesienia.

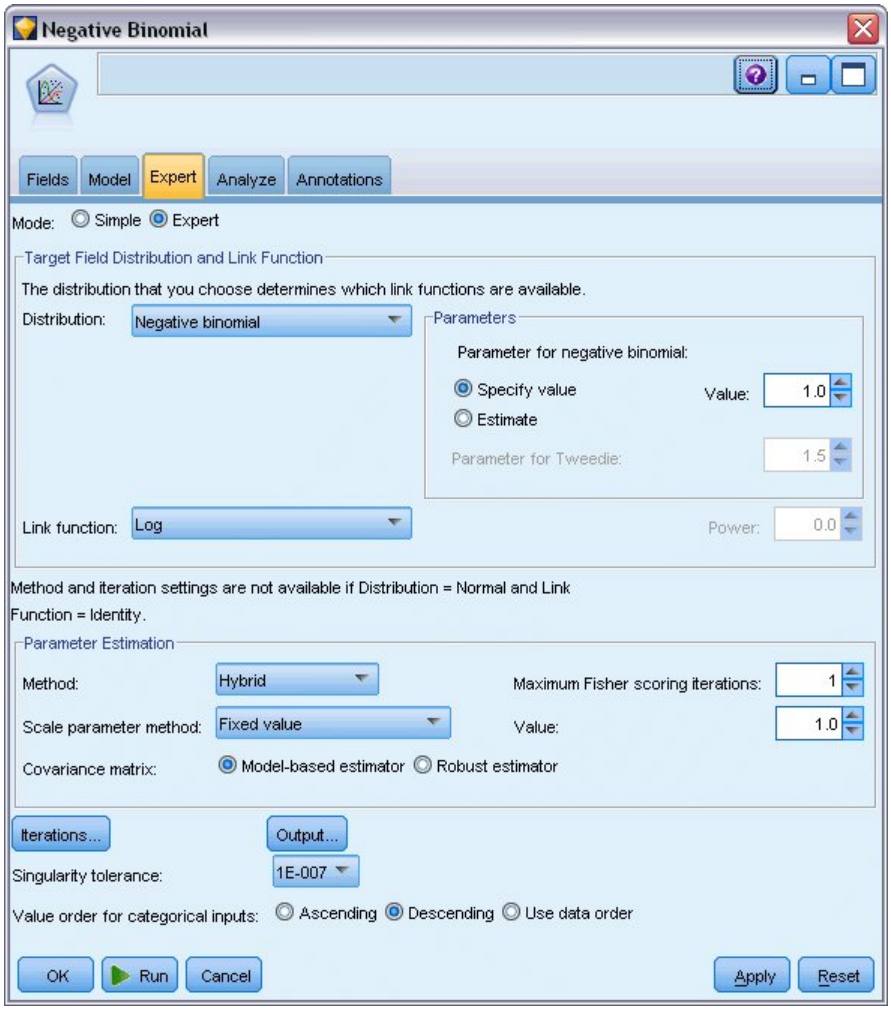
Można dokonać następujących interpretacji na podstawie oszacowań parametru:

- Typ statku *B [type=2]* ma statystycznie niższy (wartość *p* 0,019) współczynnik uszkodzeń (oszacowany współczynnik -0,543) niż typ *A [type=1]*, kategoria odniesienia. Typ *C [type=3]* ma oszacowany parametr niższy niż *B*, ale zmienność w oszacowaniu *C* przesłania efekt. Zobacz oszacowane zmienne brzegowe dla wszystkich relacji pomiędzy poziomami czynnika.
- Statki zbudowane w latach 1965–69 *[construction=65]* i 1970–74 *[construction=70]* mają statystycznie istotnie wyższe (wartość *p* <0,001) współczynniki uszkodzeń (oszacowane współczynniki odpowiednio 0,697 i 0,818) niż statki zbudowane w latach 1960–64 *[construction=60]*, kategoria odniesienia. Zobacz oszacowane zmienne brzegowe dla wszystkich relacji pomiędzy poziomami czynnika.
- Statki używane w latach 1975–79 *[operation=75]* mają statystycznie istotnie wyższe (wartość *p* 0,012) współczynniki uszkodzeń (oszacowany współczynnik 0,384) niż statki używane w latach 1960–1974 *[operation=60]*.

Dopasowanie modelu alternatywnego

Problemem w przypadku „rozproszonej" regresji Poissona jest to, że nie ma formalnego sposobu na jej testowanie w przeciwieństwie do standardowej regresji Poissona. Jednym z sugerowanych formalnych testów, aby określić, czy występuje rozproszenie, jest wykonanie testu ilorazu wiarygodności pomiędzy standardową regresją Poissona i regresją ujemną dwumianową z jednakowymi wszystkimi ustawieniami. Jeśli nie występuje rozproszenie w regresji Poissona, to statystyka −2×(logarytm wiarygodności dla modelu Poissona − logarytm wiarygodności dla modelu ujemnego dwumianowego) powinna mieć rozkład mieszanki z połową masy prawdopodobieństwa na 0 i resztą w rozkładzie chi-kwadrat z 1 stopniem swobody.

1. Wybierz metodę **Wartość ustalona** jako metodę oceny parametru skali. Domyślnie ta wartość to 1.



Rysunek 323. Karta Zaawansowany

2. Aby dopasować regresję ujemną dwumianową, skopiuj i wklej węzeł Modele uogólnione, załącz go do węzła źródłowego, otwórz nowy węzeł i kliknij kartę **Zaawansowany**.
3. Wybierz rozkład **Ujemny dwumianowy**. Pozostaw wartość domyślną 1 dla parametru pomocniczego.
4. Uruchom strumień i przejrzyj kartę Zaawansowane w nowo utworzonym modelu użytkowym.

Statystyki dobroci dopasowania

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	38.695	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	42.275	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Rysunek 324. Statystyki dobroci dopasowania dla standardowej regresji Poissona

Logarytm wiarygodności wyliczony dla standardowej regresji Poissona to -68,281. Porównaj to z modelem ujemnym dwumianowym.

	Value	df	Value/df
Deviance	11.145	25	.446
Scaled Deviance	11.145	25	
Pearson Chi-Square	8.815	25	.353
Scaled Pearson Chi-Square	8.815	25	
Log Likelihood[a]	-83.725		
Akaike's Information Criterion (AIC)	185.450		
Finite Sample Corrected AIC (AICC)	192.950		
Bayesian Information Criterion (BIC)	199.187		
Consistent AIC (CAIC)	208.187		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Rysunek 325. Statystyki dobroci dopasowania regresji ujemnej dwumianowej

Logarytm wiarygodności wyliczony dla regresji ujemnej dwumianowej to -83,725. Wartość jest *mniejsza* niż wartość logarytmu wiarygodności dla regresji Poissona, co wskazuje na to (bez potrzeby testu ilorazu wiarygodności), że ta regresja ujemna dwumianowa nie oferuje lepszych wyników niż regresja Poissona.

Wybrana wartość 1 dla parametru pomocniczego ujemnego rozkładu dwumianowego może jednak nie być optymalna dla tego zbioru danych. Kolejnym sposobem, w jaki można przetestować rozproszenie, jest dopasowanie modelu ujemnego dwumianowego z parametrem dodatkowym równym 0 i wykonanie testu mnożnikiem Lagrange'a na karcie okna dialogowego Wynik karty Zaawansowany. Jeśli test nie jest istotny, rozproszenie nie powinno być problemem dla tego zbioru danych.

Podsumowanie

Używając uogólnionych modeli liniowych, dopasowano trzy różne modele do danych liczebnościowych. Regresja ujemna dwumianowa nie wykazała poprawy w stosunku do regresji Poissona. Rozproszona regresja Poissona wydaje się oferować rozsądną alternatywę dla standardowego modelu Poissona, ale nie istnieje formalny test wyboru pomiędzy nimi.

Wyjaśnienia podstaw matematycznych metod modelowania używanych w programie IBM SPSS Modeler przedstawiono w publikacji *IBM SPSS Modeler Algorithms Guide*.

Procedury pokrewne

Procedura Uogólnione modele liniowe jest przydatnym narzędziem dopasowania różnych modeli.

- Procedura Uogólnione równania estymujące rozszerza ogólny model liniowy w taki sposób, że pozwala na powtarzane pomiary.
- Procedura Liniowe modele mieszane pozwala dopasować modele do ilościowych zmiennych zależnych ze składnikiem losowym i/lub powtarzanymi pomiarami.

Zalecana literatura

Zapoznaj się z poniższymi tekstami, aby uzyskać więcej informacji o uogólnionych modelach liniowych:

Cameron, A. C. i P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, wyd. 2. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W. i J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P. i J. A. Nelder. 1989. *Generalized Linear Models*, wyd. 2. London: Chapman & Hall.

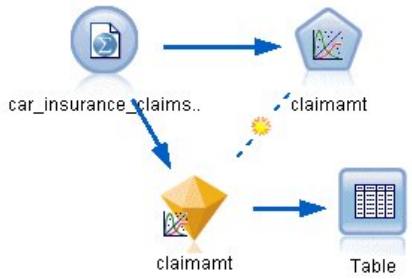
Rozdział 24. Dopasowywanie regresji Gamma do roszczeń z tytułu ubezpieczenia pojazdu (Uogólnione modele liniowe)

Uogólnionego modelu liniowego można użyć, aby dopasować regresję gamma dla analizy danych dodatniego zakresu. Na przykład zbiór danych przedstawiony i przeanalizowany w innej pracy [3] dotyczy roszczeń o odszkodowania motoryzacyjne. Średnią kwotę roszczenia można zamodelować jako zmienną o rozkładzie gamma, używając odwróconej funkcji łączenia w celu powiązania średniej zmiennej zależnej z liniową kombinacją predyktorów. W celu uwzględnienia zmiennej liczby roszczeń używanych do obliczenia średnich kwot roszczeń, zmienna *Number of claims* określana jest jako waga skalowania.

W tym przykładzie zastosowano strumień o nazwie *insurance_genlin.str*, który odwołuje się do pliku danych o nazwie *car_insurance_claims.sav*. Plik danych znajduje się w folderze *Demos*, a plik strumienia w podfolderze *streams*.

Tworzenie strumienia

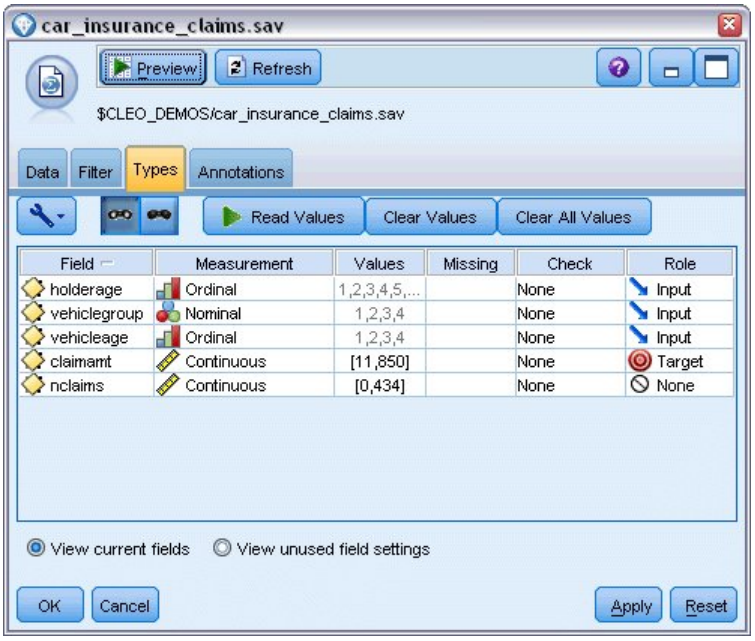
1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *car_insurance_claims.sav* znajdujący się w folderze *Demos*.



Rysunek 326. Przykładowy strumień pozwalający na przewidywanie roszczeń z tytułu ubezpieczenia pojazdu

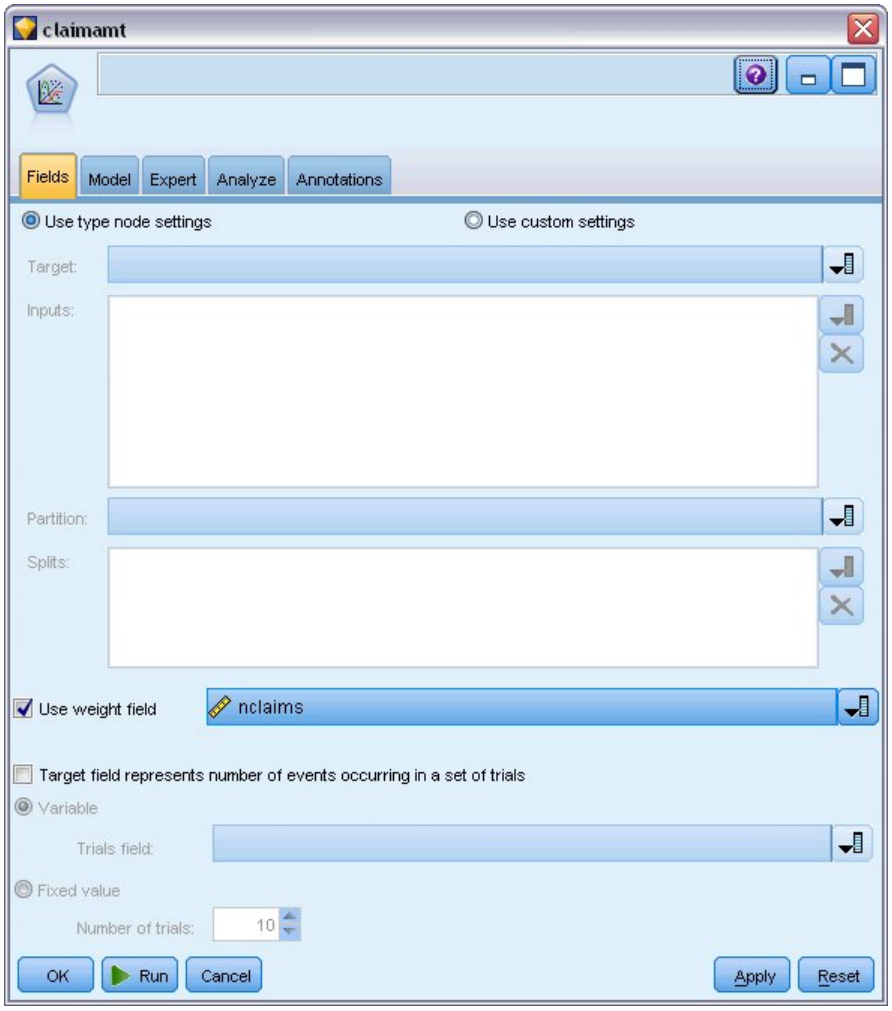
2. Na karcie Typy węzła źródłowego ustaw rolę zmiennej *claimamt* na **Przewidywana**. Wszystkie pozostałe zmienne powinny mieć ustawioną rolę na **Dane wejściowe**.
3. Kliknij przycisk **Odczytaj wartości**, aby zrealizować dane.

3. McCullagh, P. i J. A. Nelder. 1989. *Generalized Linear Models*, wyd. 2. London: Chapman & Hall.



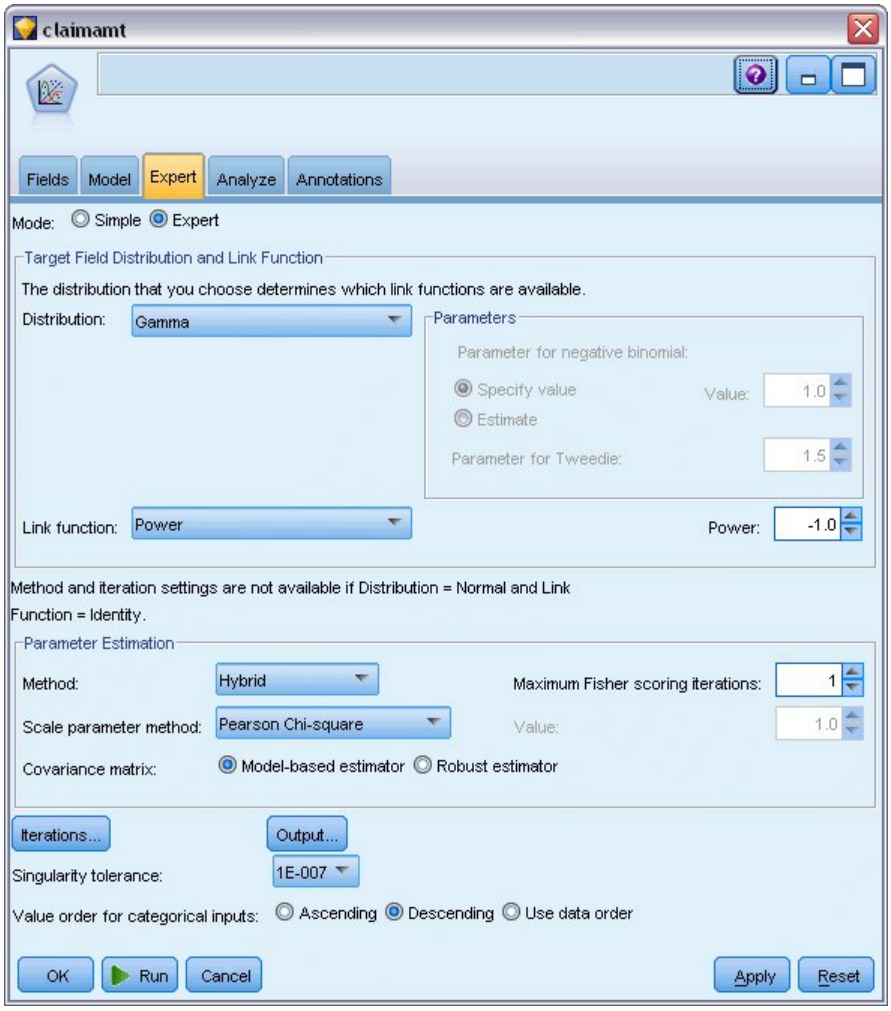
Rysunek 327. Ustawianie roli zmiennej

4. Dołącz węzeł Modele uogólnione do węzła źródłowego. W węźle Modele uogólnione kliknij kartę Zmienne.
5. Wybierz zmienną *nclaims* jako zmienną ważącą skalę.



Rysunek 328. Wybieranie opcji zmiennej

6. Kliknij kartę Zaawansowany i wybierz opcję **Zaawansowany**, aby aktywować zaawansowane opcje modelowania.



Rysunek 329. Wybieranie opcji zaawansowanych

7. Wybierz opcję **Gamma** jako rozkład reakcji.
8. Wybierz **Potęgowy** jako funkcję łączenia i wpisz -1,0 jako wykładnik funkcji wykładniczej. Jest to łącze odwrotne.
9. Wybierz metodę **Chi-kwadrat Pearsona** jako metodę oceny parametru skali. Jest to metoda używana przez McCullagha i Neldera, więc używamy jej tutaj, aby zreplikować ich wyniki.
10. Wybierz opcję **Malejąco** jako kolejność kategorii dla czynników. Wskazuje to, że pierwszą kategorią dla każdego czynnika będzie jego kategoria odniesienia. Wpływ tego wyboru na model jest interpretacją oszacowań parametrów.
11. Kliknij przycisk **Uruchom**, aby utworzyć model użytkowy, które jest dodawany do obszaru roboczego strumienia i palety modeli w prawym górnym rogu. Aby wyświetlić szczegóły modelu, kliknij model użytkowy prawym przyciskiem myszy i wybierz opcję **Edytuj** lub **Przeglądaj**, a następnie wybierz kartę Zaawansowane.

Oszacowania parametrów

Parameter	B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	Hypothesis Test df	Hypothesis Test Sig.
(Intercept)	.003411	.000418	.002591	.004230	66.593	1	.000
[holderage=8]	.000920	.000416	.000105	.001735	4.898	1	.027
[holderage=7]	.000916	.000408	.000117	.001716	5.046	1	.025
[holderage=6]	.000969	.000405	.000176	.001763	5.740	1	.017
[holderage=5]	.001370	.000419	.000548	.002192	10.682	1	.001
[holderage=4]	.000462	.000411	-.000342	.001267	1.268	1	.260
[holderage=3]	.000350	.000412	-.000458	.001158	.720	1	.396
[holderage=2]	.000101	.000436	-.000754	.000956	.054	1	.816
[holderage=1]	.000000[a]	.	.	.	.	.	.
[vehiclegroup=4]	-.001421	.000181	-.001775	-.001067	61.883	1	.000
[vehiclegroup=3]	-.000614	.000170	-.000947	-.000281	13.039	1	.000
[vehiclegroup=2]	.000038	.000169	-.000293	.000368	.050	1	.823
[vehiclegroup=1]	.000000[a]	.	.	.	.	.	.
[vehicleage=4]	.004154	.000442	.003287	.005021	88.175	1	.000
[vehicleage=3]	.001651	.000227	.001207	.002096	53.013	1	.000
[vehicleage=2]	.000366	.000101	.000169	.000564	13.191	1	.000
[vehicleage=1]	.000000[a]	.	.	.	.	.	.
(Scale)	1.209[b]			.001	.0004	.000	.002

Dependent Variable: Average cost of claims
Model: (Intercept), holderage, vehiclegroup, vehicleage

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

Rysunek 330. Oszacowania parametrów

Test typu omnibus i testy efektów modelu (niepokazane) wskazują, że model ma lepszą skuteczność niż model zerowy i każdy ze składników efektów głównych ma wpływ na model. Tabela oszacowania parametrów pokazuje takie same wartości, jak uzyskane przez McCullagha i Neldera dla poziomów czynników i parametru skali.

Podsumowanie

Używając uogólnionych modeli liniowych, dopasowano regresję gamma do danych roszczeń. Należy zauważyć, że mimo że w tym modelu użyto kanonicznej funkcji łączenia dla rozkładu gamma, logarytmiczna funkcja łączenia również zapewni dobre wyniki. Zazwyczaj trudno jest bezpośrednio porównać modele z różnymi funkcjami łączenia. Logarytmiczna funkcja łączenia jest jednak wyjątkowym przypadkiem wykładniczej funkcji łączenia, w której wykładnik to 0. Można więc porównać odchylenia modelu z logarytmiczną funkcją łączenia i modelu z wykładniczą funkcją łączenia, aby określić, który model zapewnia lepsze dopasowanie (zobacz np. sekcję 11.3 publikacji McCullagha i Neldera).

Wyjaśnienia podstaw matematycznych metod modelowania używanych w programie IBM SPSS Modeler przedstawiono w publikacji *IBM SPSS Modeler Algorithms Guide*.

Procedury pokrewne

Procedura Uogólnione modele liniowe jest przydatnym narzędziem dopasowania różnych modeli.

- Procedura Uogólnione równania estymujące rozszerza ogólny model liniowy w taki sposób, że pozwala na powtarzane pomiary.
- Procedura Liniowe modele mieszane pozwala dopasować modele do ilościowych zmiennych zależnych ze składnikiem losowym i/lub powtarzanymi pomiarami.

Zalecana literatura

Zapoznaj się z poniższymi tekstami, aby uzyskać więcej informacji o uogólnionych modelach liniowych:

Cameron, A. C. i P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.
Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, wyd. 2. Boca Raton, FL: Chapman & Hall/CRC.
Hardin, J. W. i J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P. i J. A. Nelder. 1989. *Generalized Linear Models*, wyd. 2. London: Chapman & Hall.

Rozdział 25. Klasyfikowanie próbek komórek (SVM)

Algorytm SVM to technika klasyfikacji i regresji, która jest wyjątkowo dostosowana do szerokich zbiorów danych. Szeroki zbiór danych to zbiór z dużą liczbą predyktorów, taki jak można napotkać w obszarze bioinformatyki (zastosowanie informatyki do danych biochemicznych i biologicznych).

Pracownik naukowo-badawczy zgromadził zbiór danych zawierający charakterystykę pewnej liczby prób komórek ludzkich pobranych od pacjentów z podejrzeniem nowotworu. Jak wykazała analiza oryginalnych danych, wiele z charakterystyk próbek z nowotworem złośliwym różniło się istotnie od próbek z nowotworem niezłośliwym. Badacz chce opracować model SVM wykorzystujący wartości charakterystyk komórek w próbkach od innych pacjentów w celu wstępnego określenia, czy ich próbki zawierają nowotwór złośliwy, czy niezłośliwy.

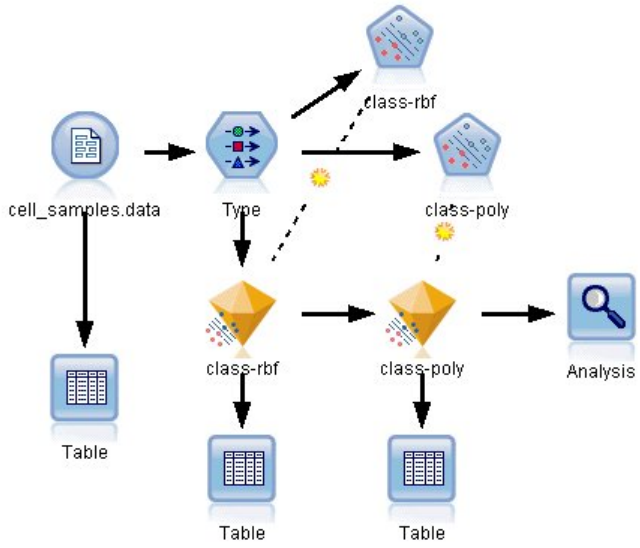
W tym przykładzie zastosowano strumień o nazwie *svm_cancer.str*, który jest dostępny w folderze *Demos*, podfolder *streams*. Plik danych to *cell_samples.data*. Więcej informacji można znaleźć w temacie “Folder Demos” na stronie 5.

Ten przykład opiera się na zbiorze danych udostępnionym publicznie przez UCI Machine Learning Repository. Zbiór danych składa się z kilkuset rekordów próbek komórek ludzkich, z których każda zawiera wartości zbioru cech komórki. Zmienne w każdym rekordzie to:

Nazwa zmiennej	**Opis**
ID	Identyfikator pacjenta
Clump	Grubość skupień
UnifSize	Jednorodność rozmiaru komórki
UnifShape	Jednorodność kształtu komórki
MargAdh	Przyleganie komórek
SingEpiSize	Rozmiar pojedynczej komórki nabłonka
BareNuc	Odsłonięte jądra
BlandChrom	Jednorodna chromatyna
NormNucl	Jądra prawidłowe
Mit	Mitozy
Class	Łagodny lub złośliwy

Do celów tego przykładu używamy zbioru danych, który ma relatywnie małą liczbę predyktorów w każdym rekordzie.

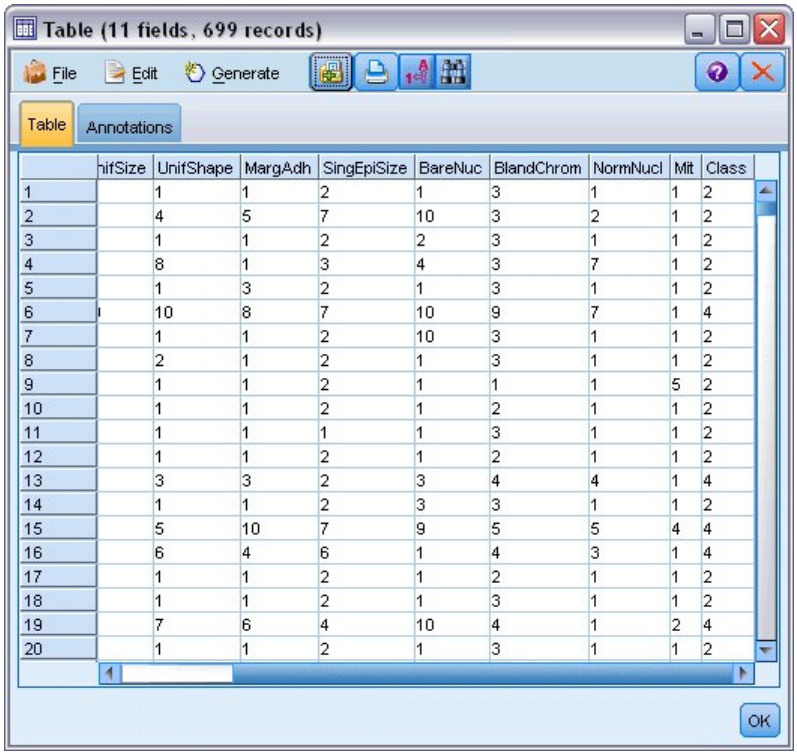
Tworzenie strumienia



Rysunek 331. Przykładowy strumień przedstawiający modelowanie SVM

1. Utwórz nowy strumień i dodaj węzeł źródłowy pliku zmiennych wskazujący na plik *cell_samples.data* w folderze *Demos* instalacji programu IBM SPSS Modeler.

   Spójrzmy na dane w pliku źródłowym.
2. Do strumienia dodaj węzeł tabeli.
3. Dołącz węzeł tabeli do węzła pliku zmiennych i uruchom strumień.

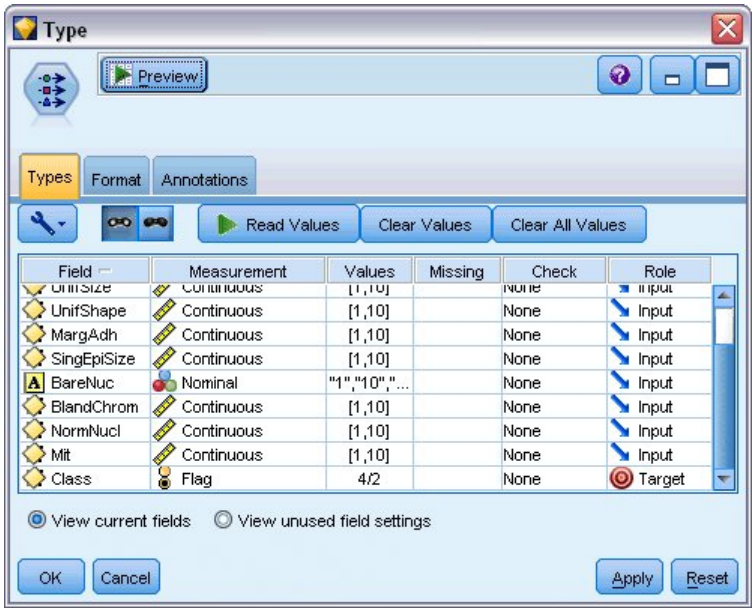


	hifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1		1	1	2	1	3	1	1	2
2		4	5	7	10	3	2	1	2
3		1	1	2	2	3	1	1	2
4		8	1	3	4	3	7	1	2
5		1	3	2	1	3	1	1	2
6		10	8	7	10	9	7	1	4
7		1	1	2	10	3	1	1	2
8		2	1	2	1	3	1	1	2
9		1	1	2	1	1	1	5	2
10		1	1	2	1	2	1	1	2
11		1	1	1	1	3	1	1	2
12		1	1	2	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

Rysunek 332. Dane źródłowe dla modelu SVM

Zmienna *ID* zawiera identyfikatory pacjentów. Charakterystyki próbek komórek od każdego pacjenta zawarte są w zmiennych od *Clump* do *Mit*. Wartości są oceniane w skali od 1 do 10, gdzie 1 to wartość najbliższa łagodnej próbce.

Zmienna *Class* zawiera diagnozę potwierdzoną osobnymi procedurami medycznymi, czy próbka jest łagodna (wartość = 2), czy złośliwa (wartość = 4).



Field	Measurement	Values	Missing	Check	Role
UnifSize	Continuous	[1,10]		None	Input
UnifShape	Continuous	[1,10]		None	Input
MargAdh	Continuous	[1,10]		None	Input
SingEpiSize	Continuous	[1,10]		None	Input
BareNuc	Nominal	"1","10",...		None	Input
BlandChrom	Continuous	[1,10]		None	Input
NormNucl	Continuous	[1,10]		None	Input
Mit	Continuous	[1,10]		None	Input
Class	Flag	4/2		None	Target

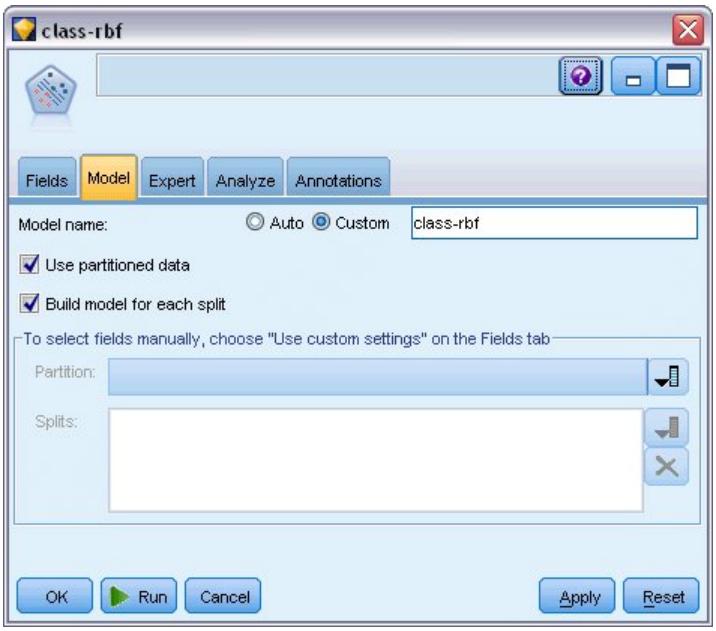
Rysunek 333. Ustawienia węzła typu

4. Dodaj węzeł typu i załącz go do węzła pliku zmiennych.

5. Otwórz węzeł typu.

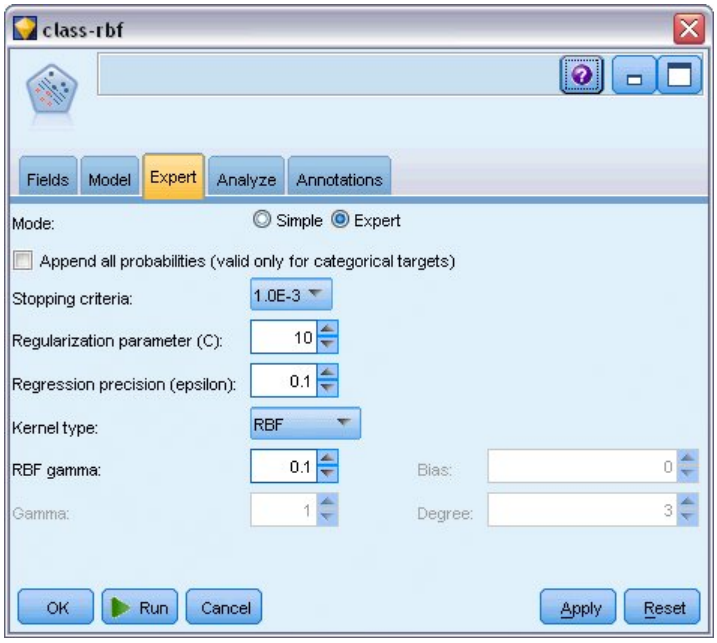
   Chcemy, aby model przewidział wartość zmiennej *Class* (tzn. łagodny (=2) lub złośliwy (=4)). Ponieważ ta zmienna może mieć jedną z dwóch możliwych wartości, musimy zmienić poziom pomiaru, aby to odzwierciedlić.
6. W kolumnie **Poziom pomiaru** dla zmiennej *Class* (ostatnia pozycja na liście) kliknij wartość **Ilościowa** i zmień ją na **Flaga**.
7. Kliknij przycisk **Odczytaj wartości**.
8. W kolumnie **Rola** ustaw rolę dla zmiennej *ID* (identyfikator pacjenta) na **Brak**, ponieważ ta zmienna nie będzie używana ani jako predyktor, ani jako zmienna przewidywana dla modelu.
9. Ustaw rolę dla zmiennej przewidywanej *Class* na **Przewidywana** i pozostaw role pozostałych zmiennych (predyktorów) ustawione na **Dane wejściowe**.
10. Kliknij przycisk **OK**.

    Węzeł SVM oferuje wybór funkcji algorytmu domyślnego do wykonywania jego przetwarzania. Ponieważ nie istnieje prosty sposób określenia, która funkcja działa najlepiej z dowolnym zbiorem danych, wybierzemy kolejno różne funkcje i porównamy wyniki. Rozpoczniemy od domyślnej opcji, którą jest radialna funkcja bazowa (RBF).



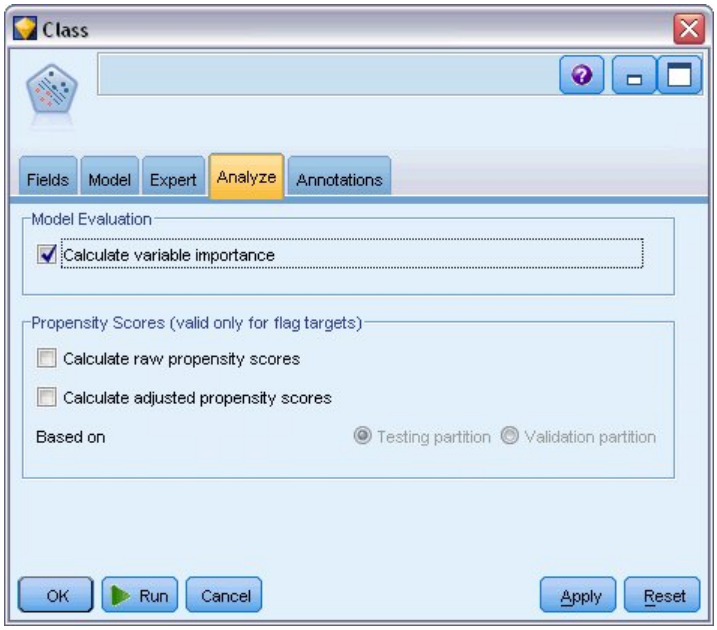
Rysunek 334. Ustawienia karty Model

11. Z poziomu palety Modelowanie dołącz węzeł SVM do węzła typu.
12. Otwórz węzeł SVM. Na karcie **Model** kliknij opcję **Użytkownika** dla pozycji **Nazwa modelu** i wpisz *class-rbf* w sąsiednim polu tekstowym.



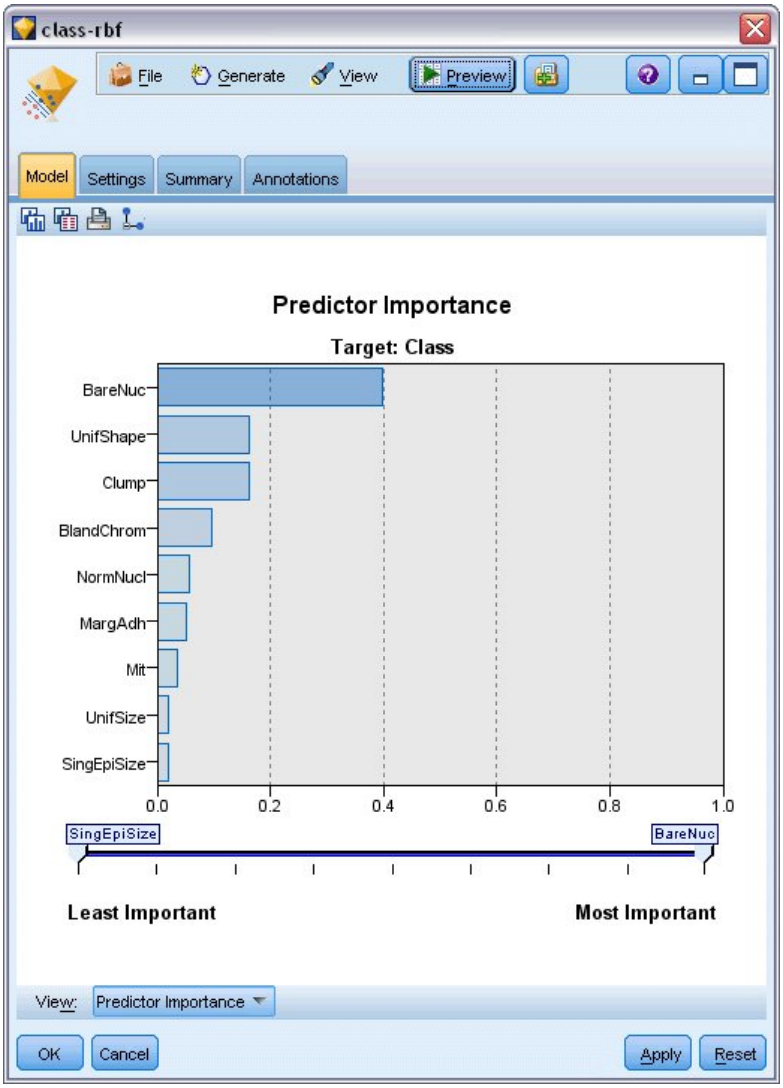
Rysunek 335. Domyślne ustawienia karty Zaawansowany

13. Na karcie **Zaawansowany** ustaw **Tryb** na **Zaawansowany**, aby zapewnić czytelność, ale pozostaw wszystkie domyślne opcje tak, jak są ustawione. Zauważ, że opcja **Typ jądra** jest ustawiona domyślnie na **RBF**. W trybie prostym wszystkie opcje są wyszarzone.



Rysunek 336. Ustawienia karty Analiza

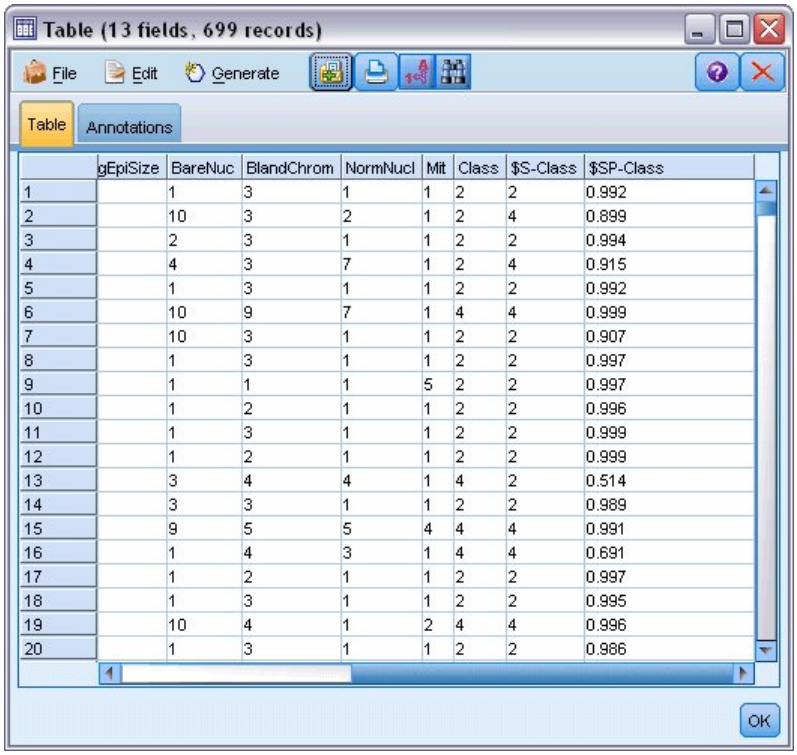
14. Na karcie **Analiza** zaznacz pole wyboru **Oblicz ważność zmiennych**.
15. Kliknij przycisk **Uruchom**. Model użytkowy jest umieszczany w strumieniu i w palecie modeli w prawym górnym rogu okna.
16. Dwukrotnie kliknij model użytkowy w strumieniu.



Rysunek 337. Wykres Ważność predyktorów

Wykres Ważność predyktorów na karcie Model przedstawia względny efekt różnych zmiennych na predykcję. Pokazuje to, że zmienna *BareNuc* ma zdecydowanie największy efekt, podczas gdy zmienne *UnifShape* i *Clump* są również istotne.

1. Kliknij przycisk **OK**.
2. Dołącz węzeł tabeli do modelu użytkowego *class-rbf*.
3. Otwórz węzeł Tabela i kliknij przycisk **Uruchom**.



	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

Rysunek 338. Dodane zmienne predykcji i współczynnika ufności

4. Model utworzył dwie dodatkowe zmienne. Przewiń tabelę w prawo, aby je zobaczyć:

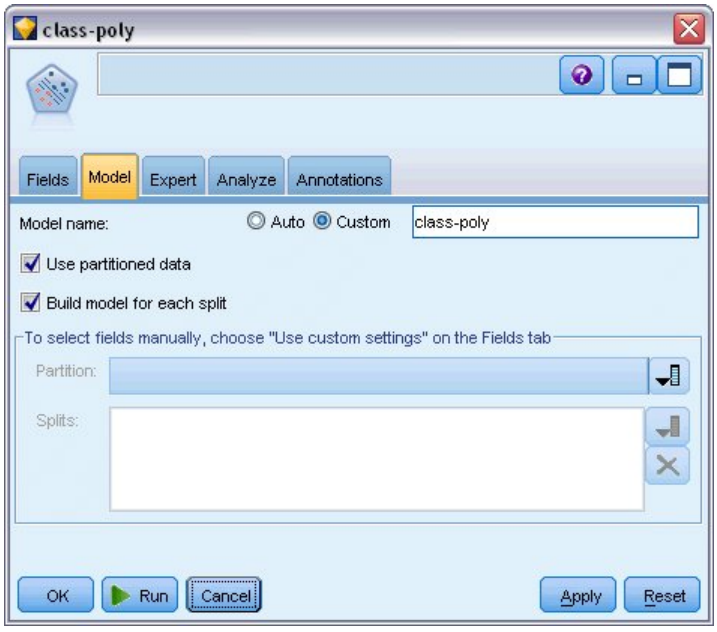
Nazwa nowej zmiennej	**Opis**
$S-Class	Wartość zmiennej *Class* przewidziana przez model.
$SP-Class	Ocena skłonności dla tej predykcji (prawdopodobieństwo, że ta predykcja będzie prawdziwa — wartość od 0,0 do 1,0).

Tylko patrząc na tabelę można zobaczyć, że oceny skłonności (w kolumnie *$SP-Class*) dla większości rekordów są raczej wysokie.

Zdarzają się jednak istotne wyjątki. Na przykład pacjent 1041801 w wierszu 13, gdzie wartość 0,514 jest niedopuszczalnie niska. Również porównanie zmiennych *Class* i *$S-Class* pokazuje wyraźnie, że ten model dokonał wiele niepoprawnych predykcji, nawet jeśli ocena skłonności była relatywnie wysoka (na przykład wiersze 2 i 4).

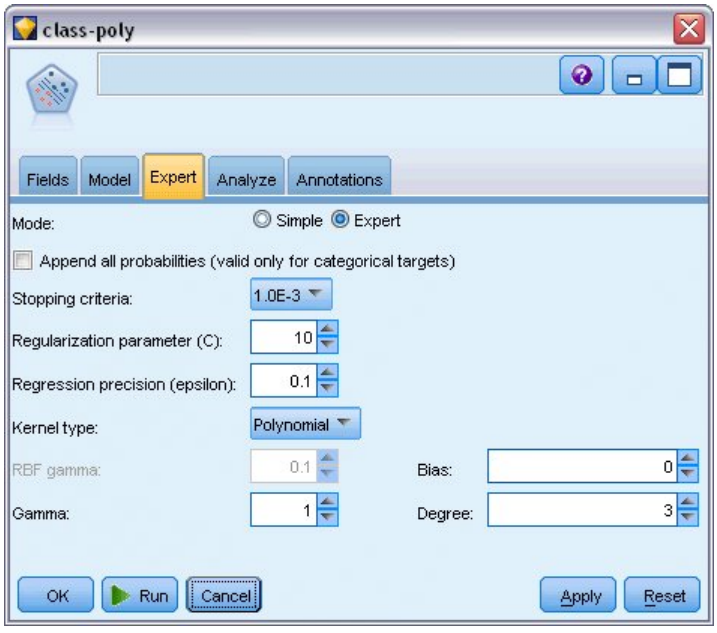
Zobaczmy, czy możemy osiągnąć lepsze wyniki, wybierając inny typ funkcji.

Próbowanie innej funkcji



Rysunek 339. Ustawianie nowej nazwy modelu

1. Zamknij okno wyników tabeli.
2. Załącz drugi węzeł modelowania SVM do węzła typu.
3. Otwórz nowy węzeł SVM.
4. Na karcie **Model** wybierz opcję Użytkownika i wpisz *class-poly* jako nazwę modelu.

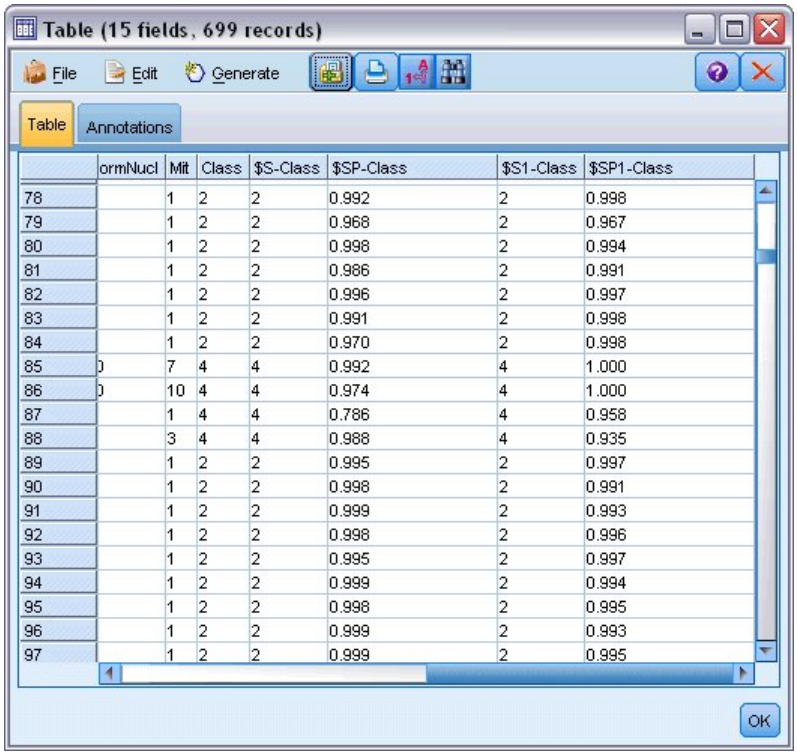


Rysunek 340. Ustawienia karty Zaawansowany dla opcji Wielomian

5. Na karcie **Zaawansowany** ustaw **Tryb** na **Zaawansowany**.

6. Ustaw opcję **Typ jądra** na **Wielomian** i kliknij przycisk **Uruchom**. Model użytkowy *class-poly* jest umieszczany w strumieniu i w palecie modeli w prawym górnym rogu okna.
7. Podłącz model użytkowy *class-rbf* do modelu użytkowego *class-poly* (wybierz opcję **Zamień** w oknie ostrzeżenia).
8. Dołącz węzeł tabeli do modelu użytkowego *class-poly*.
9. Otwórz węzeł Tabela i kliknij przycisk **Uruchom**.

Porównywanie wyników



	ormNucl	Mit	Class	$S-Class	$SP-Class	$S1-Class	$SP1-Class
78		1	2	2	0.992	2	0.998
79		1	2	2	0.968	2	0.967
80		1	2	2	0.998	2	0.994
81		1	2	2	0.986	2	0.991
82		1	2	2	0.996	2	0.997
83		1	2	2	0.991	2	0.998
84		1	2	2	0.970	2	0.998
85	0	7	4	4	0.992	4	1.000
86	0	10	4	4	0.974	4	1.000
87		1	4	4	0.786	4	0.958
88		3	4	4	0.988	4	0.935
89		1	2	2	0.995	2	0.997
90		1	2	2	0.998	2	0.991
91		1	2	2	0.999	2	0.993
92		1	2	2	0.998	2	0.996
93		1	2	2	0.995	2	0.997
94		1	2	2	0.999	2	0.994
95		1	2	2	0.998	2	0.995
96		1	2	2	0.999	2	0.993
97		1	2	2	0.999	2	0.995

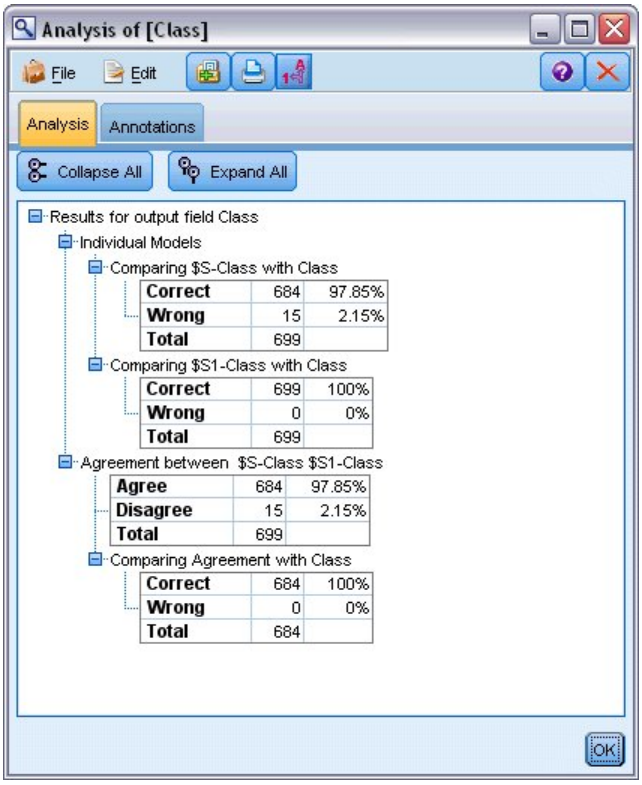
Rysunek 341. Dodane zmienne funkcji Wielomian

1. Przewiń tabelę w prawo, aby zobaczyć nowo dodane zmienne.

   Wygenerowane zmienne typu funkcji Wielomian mają nazwy *$S1-Class* i *$SP1-Class*.

   Wyniki dla funkcji Wielomian wyglądają dużo lepiej. Wiele ocen skłonności ma wartość 0,995 lub lepszą, co jest bardzo zachęcające.
2. Aby potwierdzić poprawę modelu, załącz węzeł analizy do modelu użytkowego *class-poly*.

Otwórz węzeł analizy i kliknij przycisk **Uruchom**.



Rysunek 342. Węzeł Analiza

Ta technika użycia węzła analizy pozwala na porównanie dwóch lub wielu modeli użytkowych jednocześnie. Wyniki z węzła analizy pokazują, że funkcja RBF przewiduje poprawnie 97,85% przypadków, co jest całkiem dobrym wynikiem. Wyniki pokazują jednak, że funkcja Wielomian przewidziała poprawnie diagnozę w każdym przypadku. W praktyce mało prawdopodobne jest uzyskanie 100-procentowej dokładności, ale można użyć węzła analizy, aby pomóc określić, czy model jest odpowiednio dokładny dla konkretnego zastosowania.

Żadna z pozostałych typów funkcji (Funkcja sigmoidalna i Liniowa) nie osiąga tak dobrych wyników jak funkcja Wielomian dla tego konkretnego zbioru danych. Jednak przy innym zbiorze danych wyniki mogłyby być inne, więc zawsze warto sprawdzić pełny zakres opcji.

Podsumowanie

Użyto różnych typów funkcji algorytmu domyślnego SVM, aby przewidzieć klasyfikację wielu atrybutów. Zauważyliśmy, że różne algorytmy domyślne dają różne wyniki dla tego samego zbioru danych, i poznaliśmy sposób pomiaru poprawy jednego modelu względem drugiego.

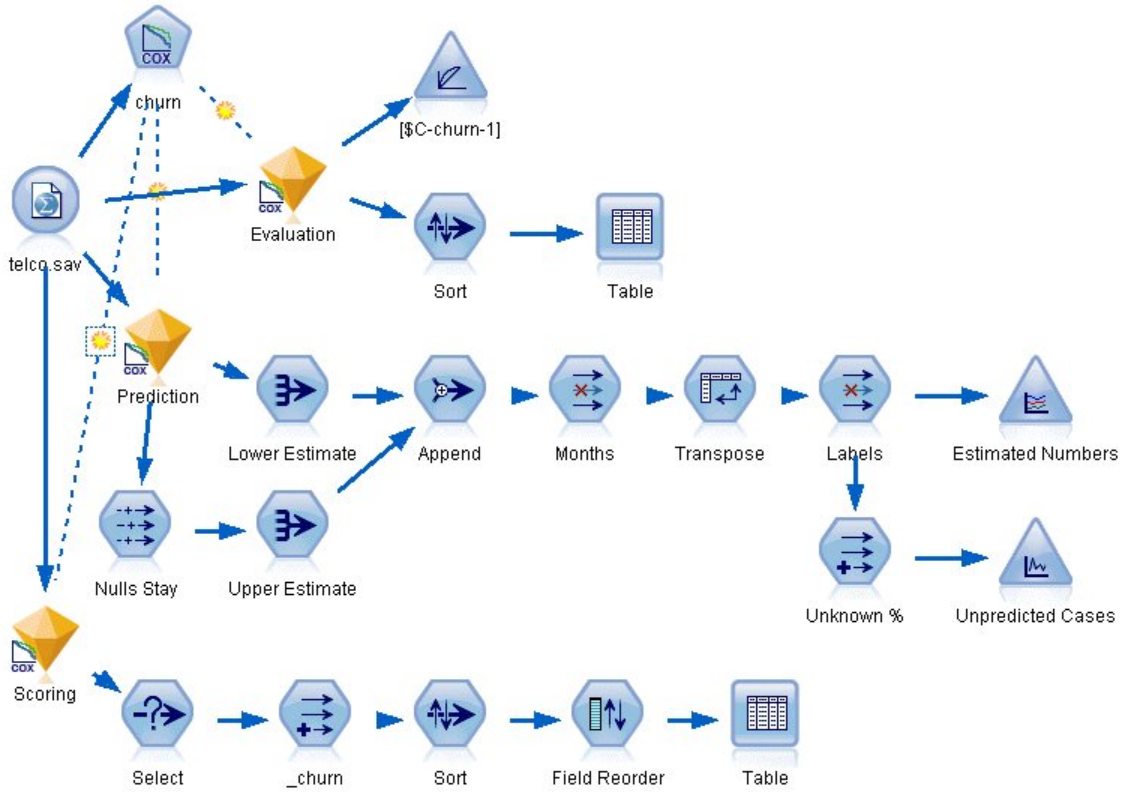
Rozdział 26. Użycie regresji Coxa do modelowania czasu do odejścia klienta

W ramach strategii zapobiegania odejściom klientów operator telekomunikacyjny jest zainteresowany modelowaniem „czasu do odejścia” w celu określenia czynników charakterystycznych dla klientów, którzy szybko zmieniają operatora. W tym celu wybierana jest losowa próba klientów i z bazy danych pobierane są informacje o czasie współpracy (czy wciąż są aktywnymi klientami) oraz różne inne zmienne.

W tym przykładzie zastosowano strumień o nazwie *telco_coxreg.str*, który odwołuje się do pliku danych *telco.sav*. Plik danych znajduje się w folderze *Demos*, a plik strumienia w podfolderze *streams*. Więcej informacji można znaleźć w temacie “Folder Demos” na stronie 5.

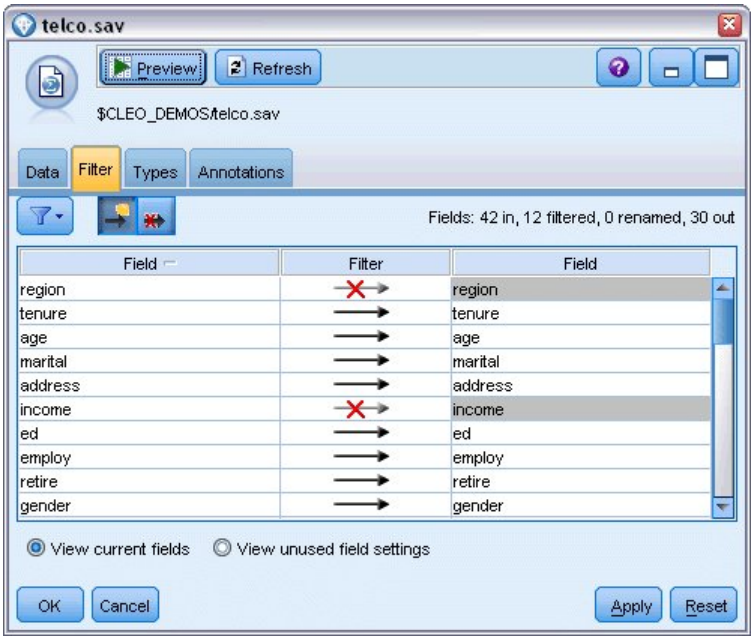
Budowanie odpowiedniego modelu

1. Dodaj węzeł źródłowy Plik Statistics wskazujący na plik *telco.sav* znajdujący się w folderze *Demos*.



Rysunek 343. Przykładowy strumień analizujący czas do odejścia

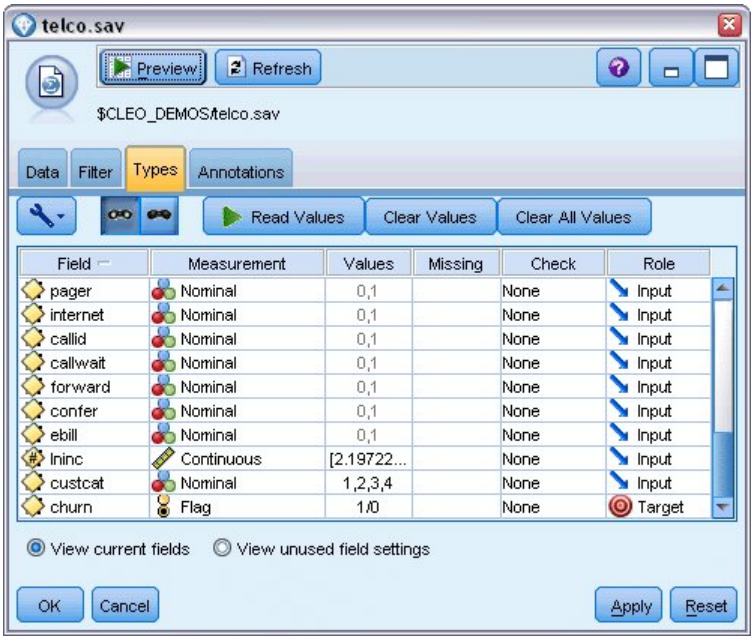
2. Na karcie Filtrowanie węzła źródłowego wyłącz zmienne *region*, *income*, od *longten* przez *wireten* oraz zmienne od *loglong* do *logwire*.



Rysunek 344. Filtrowanie niepotrzebnych zmiennych

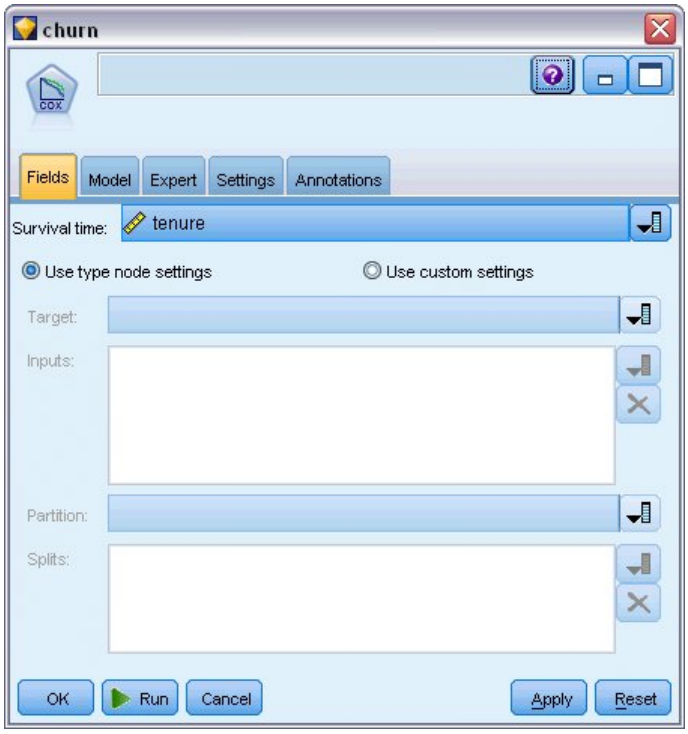
(Można również zmienić rolę tych zmiennych na **Brak**, zamiast je wykluczać, lub wybrać zmienne, których chcesz użyć w węźle modelowania).

3. Na karcie Typy węzła źródłowego ustaw rolę zmiennej *churn* na **Przewidywana** i ustaw jej poziom pomiaru na **Flaga**. Wszystkie pozostałe zmienne powinny mieć ustawioną rolę na **Dane wejściowe**.
4. Kliknij przycisk **Odczytaj wartości**, aby zrealizować dane.



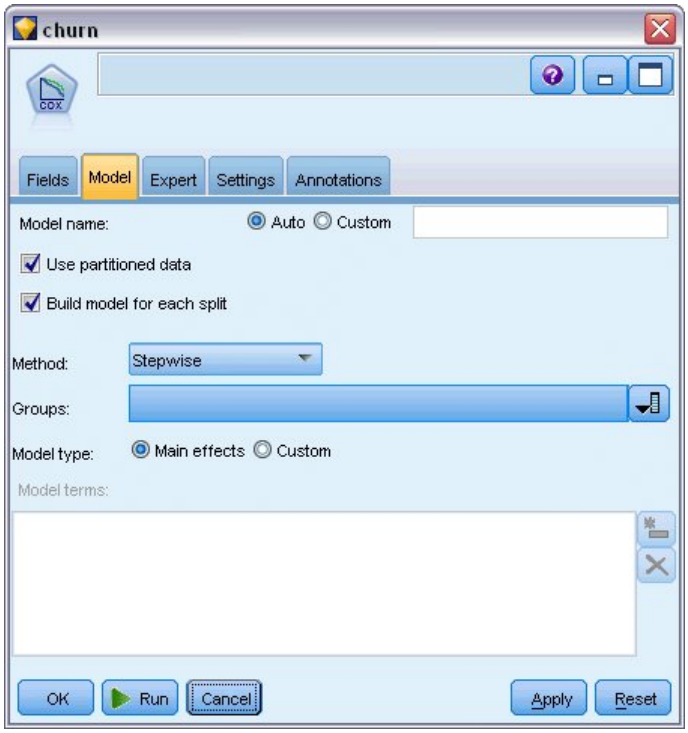
Rysunek 345. Ustawianie roli zmiennej

5. Załącz węzeł Model Coxa do węzła źródłowego. Na karcie **Zmienne** wybierz zmienną *tenure* jako zmienną czasu przeżycia.



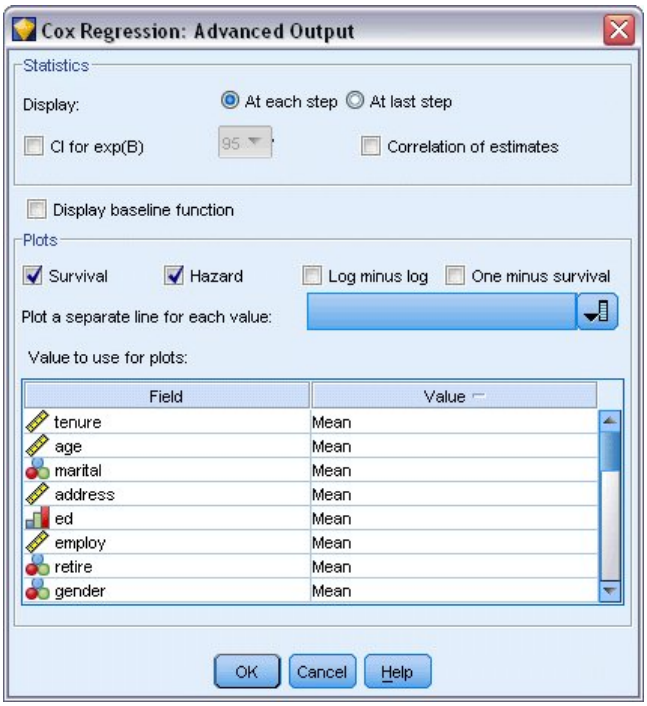
Rysunek 346. Wybieranie opcji zmiennej

6. Kliknij kartę **Model** .
7. Wybierz pozycję **Krokowa** jako metodę wyboru zmiennej.



Rysunek 347. Wybieranie opcji modelu

8. Kliknij kartę **Zaawansowany** i wybierz opcję **Zaawansowany**, aby aktywować zaawansowane opcje modelowania.
9. Kliknij opcję **Wynik**.



Rysunek 348. Wybieranie zaawansowanych opcji wyników

10. Zaznacz wykresy **Analiza przeżycia** i **Funkcja hazardu**, które zostaną wygenerowane, a następnie kliknij przycisk **OK**.
11. Kliknij przycisk **Uruchom**, aby utworzyć model użytkowy, który jest dodawany do strumienia i palety modeli w prawym górnym rogu. Aby wyświetlić jego szczegóły, dwukrotnie kliknij model użytkowy w strumieniu. Najpierw popatrz na kartę Zaawansowane.

Obserwacje ocenzurowane

		N	Percent
Cases available in analysis	Event[a]	274	27.4%
	Censored	726	72.6%
	Total	1000	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		1000	100.0%

a. Dependent Variable: Months with service

Rysunek 349. Informacja o analizowanych danych

Zmienna statusu identyfikuje, czy zdarzenie wystąpiło dla danej obserwacji. Jeśli zdarzenie nie wystąpiło, obserwacja jest określana jako ocenzurowana. Obserwacje ocenzurowane nie są używane w obliczaniu współczynników regresji, ale są używane do obliczania linii bazowej hazardu. Informacje o analizowanych danych pokazują, że 726 obserwacji jest ocenzurowanych. Są to klienci, którzy nie odeszli.

Kodowanie zmiennych jakościowych

		Frequency	(1)[b]	(2)	(3)	(4)
marital[a]	0=Unmarried	505	1			
	1=Married	495	0			
ed[a]	1=Did not complete high school	204	1	0	0	0
	2=High school degree	287	0	1	0	0
	3=Some college	209	0	0	1	0
	4=College degree	234	0	0	0	1
	5=Post-undergraduate degree	66	0	0	0	0
retire[a]	.00=No	953	1			
	1.00=Yes	47	0			
gender[a]	0=Male	483	1			
	1=Female	517	0			
tollfree[a]	0=No	526	1			
	1=Yes	474	0			
equip[a]	0=No	614	1			
	1=Yes	386	0			
callcard[a]	0=No	322	1			
	1=Yes	678	0			
wireless[a]	0=No	704	1			
	1=Yes	296	0			
multline[a]	0=No	525	1			
	1=Yes	475	0			
voice[a]	0=No	696	1			
	1=Yes	304	0			
pager[a]	0=No	739	1			
	1=Yes	261	0			
internet[a]	0=No	632	1			
	1=Yes	368	0			
callid[a]	0=No	519	1			
	1=Yes	481	0			
callwait[a]	0=No	515	1			
	1=Yes	485	0			
forward[a]	0=No	507	1			
	1=Yes	493	0			
confer[a]	0=No	498	1			
	1=Yes	502	0			
ebill[a]	0=No	629	1			
	1=Yes	371	0			
custcat[a]	1=Basic service	266	1	0	0	
	2=E-service	217	0	1	0	
	3=Plus service	281	0	0	1	
	4=Total service	236	0	0	0	

Rysunek 350. Kodowanie zmiennych jakościowych

Kodowanie zmiennych jakościowych jest przydatnym odniesieniem do interpretowania współczynników regresji dla współzmiennych jakościowych, zwłaszcza zmiennych dychotomicznych. Domyślnie kategoria odniesienia jest „ostatnią” kategorią. Dlatego też, nawet jeśli klienci ze statusem *Married* mają wartość zmiennej wynoszącą 1 w pliku danych, wartości te są kodowane jako 0 do celów regresji.

Wybór zmiennych

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	3392.536	162.303	1	.000	133.828	1	.000	133.828	1	.000
2[b]	3087.314	249.392	2	.000	305.222	1	.000	439.050	2	.000
3[c]	3027.085	328.426	3	.000	60.229	1	.000	499.279	3	.000
4[d]	2990.790	347.197	4	.000	36.294	1	.000	535.574	4	.000
5[e]	2973.790	362.673	5	.000	17.000	1	.000	552.574	5	.000
6[f]	2958.796	376.140	6	.000	14.994	1	.000	567.568	6	.000
7[g]	2945.503	384.717	7	.000	13.293	1	.000	580.861	7	.000
8[h]	2936.993	417.341	8	.000	8.510	1	.004	589.371	8	.000
9[i]	2926.000	423.911	9	.000	10.994	1	.001	600.364	9	.000
10[j]	2917.551	428.078	10	.000	8.449	1	.004	608.813	10	.000
11[k]	2913.308	436.837	11	.000	4.243	1	.039	613.056	11	.000
12[l]	2908.078	440.158	12	.000	5.230	1	.022	618.286	12	.000

a. Variable(s) Entered at Step Number 1: callcard
b. Variable(s) Entered at Step Number 2: longmon
c. Variable(s) Entered at Step Number 3: equip
d. Variable(s) Entered at Step Number 4: employ
e. Variable(s) Entered at Step Number 5: multline
f. Variable(s) Entered at Step Number 6: voice
g. Variable(s) Entered at Step Number 7: address
h. Variable(s) Entered at Step Number 8: equipmon
i. Variable(s) Entered at Step Number 9: ebill
j. Variable(s) Entered at Step Number 10: callid
k. Variable(s) Entered at Step Number 11: internet
l. Variable(s) Entered at Step Number 12: reside
m. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 3526.364
n. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

Rysunek 351. Test typu omnibus

Proces budowania modelu wykorzystuje algorytm krokowej postępującej. Testy typu omnibus są miarą tego, jak dobrze działa model.Zmiana chi-kwadrat z poprzedniego kroku jest różnicą pomiędzy wartością −2 logarytmu wiarygodności modelu w poprzednim kroku i w bieżącym kroku. Jeśli krokiem było dodanie zmiennej, takie włączenie ma sens, jeśli istotność zmiany jest mniejsza niż 0,05. Jeśli krokiem było usunięcie zmiennej, takie wyłączenie ma sens, jeśli istotność zmiany jest większa niż 0,10. W dwunastu krokach do modelu dodawanych jest 12 zmiennych.

		B	SE	Wald	df	Sig.	Exp(B)
Step 12	address	-.035	.009	14.543	1	.000	.966
	employ	-.051	.010	25.767	1	.000	.950
	reside	-.103	.046	5.037	1	.025	.902
	equip	-1.948	.381	26.180	1	.000	.143
	callcard	.777	.151	26.451	1	.000	2.175
	longmon	-.233	.022	115.619	1	.000	.792
	equipmon	-.042	.011	15.377	1	.000	.959
	multline	.612	.145	17.854	1	.000	1.844
	voice	-.501	.157	10.197	1	.001	.606
	internet	-.362	.160	5.114	1	.024	.697
	callid	-.464	.148	9.790	1	.002	.629
	ebill	-.399	.156	6.557	1	.010	.671

Rysunek 352. Zmienne w równaniu (tylko krok 12)

Model końcowy obejmuje zmienne *address*, *employ*, *reside*, *equip*, *callcard*, *longmon*, *equipmon*, *multline*, *voice*, *internet*, *callid* i *ebill*. Aby zrozumieć efekt każdego z predyktorów, sprawdź wartość Exp(B), którą można interpretować jako przewidywaną zmianę ryzyka dla zwiększenia jednostki w predyktorze.

- Wartość Exp(B) dla zmiennej *address* oznacza, że ryzyko odejścia jest ograniczone o 100%−(100%×0,966)=3,4% dla każdego roku, w którym klient mieszkał pod tym samym adresem. Ryzyko odejścia dla klienta, który mieszkał pod tym samym adresem przez pięć lat, jest zmniejszone o $100\%-(100\%\times 0{,}966^{5})=15{,}88\%$.

- Wartość Exp(B) dla zmiennej *callcard* oznacza, że ryzyko odejścia dla klienta, który nie jest subskrybentem usługi karty telefonicznej, jest 2,175 razy większe niż dla klienta z usługą. Należy pamiętać z kodowania zmiennych jakościowych, że *Nie* = 1 dla regresji.
- Wartość Exp(B) dla zmiennej *internet* oznacza, że ryzyko odejścia dla klienta, który nie jest subskrybentem usługi Internetu, jest 0,697 razy większe niż z usługą. Daje to powody do zmartwienia, ponieważ sugeruje, że klienci z usługą opuszczają firmę szybciej niż klienci bez usługi.

		Score	df	Sig.
Step 12	age	.122	1	.726
	marital	.648	1	.421
	income	1.476	1	.224
	ed	6.328	4	.176
	ed(1)	.007	1	.934
	ed(2)	.203	1	.652
	ed(3)	.835	1	.361
	ed(4)	5.773	1	.016
	retire	.013	1	.908
	gender	.214	1	.644
	tollfree	3.243	1	.072
	wireless	.668	1	.414
	tollmon	.000	1	.987
	cardmon	3.163	1	.075
	wiremon	1.084	1	.298
	pager	1.808	1	.179
	callwait	.266	1	.606
	forward	2.201	1	.138
	confer	2.568	1	.109
	custcat	.864	3	.834
	custcat(1)	.466	1	.495
	custcat(2)	.450	1	.502
	custcat(3)	.019	1	.889

Rysunek 353. Zmienne poza modelem (tylko krok 12)

Zmienne pozostawione poza modelem wszystkie mają statystyki ocen z wartościami istotności wyższymi niż 0,05. Wartości istotności zmiennych *tollfree* i *cardmon* nie są mniejsze niż 0,05, ale są bardzo blisko. Interesujące może być zbadanie tych faktów w dalszych badaniach.

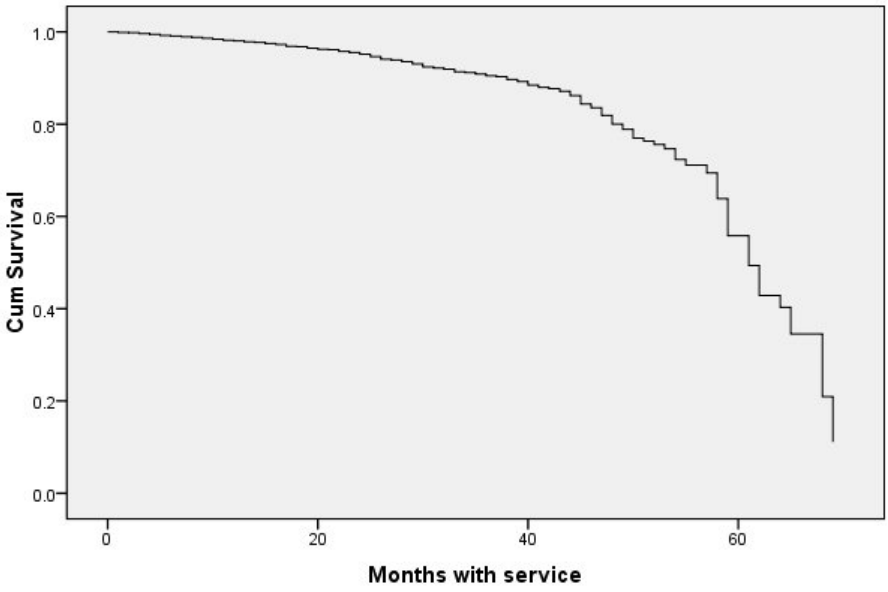
Średnie współzmiennych

	Mean
age	41.684
marital	.505
address	11.551
income	77.535
ed(1)	.204
ed(2)	.287
ed(3)	.209
ed(4)	.234
employ	10.987
retire	.953
gender	.483
reside	2.331
tollfree	.526
equip	.614
callcard	.322
wireless	.704
longmon	11.723
tollmon	13.274
equipmon	14.220
cardmon	13.781
wiremon	11.584
multline	.525
voice	.696
pager	.739
internet	.632
callid	.519
callwait	.515
forward	.507
confer	.498
ebill	.629
custcat(1)	.266
custcat(2)	.217
custcat(3)	.281

Rysunek 354. Średnie współzmiennych

Ta tabela wyświetla średnią wartość dla każdej zmiennej predykcyjnej. Zawiera przydatne informacje podczas przeglądania wykresów przeżycia, które są generowane dla wartości średnich. Należy jednak zauważyć, że „przeciętny” klient w rzeczywistości nie istnieje, gdy analizowane są średnie zmiennych wskaźnikowych dla predyktorów jakościowych. Nawet przy wszystkich predyktorach ilościowych mało prawdopodobne jest znalezienie klienta, którego wartości współzmiennej są wszystkie zbliżone do średniej. Jeśli chcesz zobaczyć krzywą przeżycia dla określonej obserwacji, możesz zmienić wartości współzmiennej, przy których rysowana jest krzywa przeżycia w oknie dialogowym Wykresy. Jeśli chcesz zobaczyć krzywą przeżycia dla określonej obserwacji, możesz zmienić wartości współzmiennej, przy których rysowana jest krzywa przeżycia w grupie Wykresy okna dialogowego Zaawansowane dane wyjściowe.

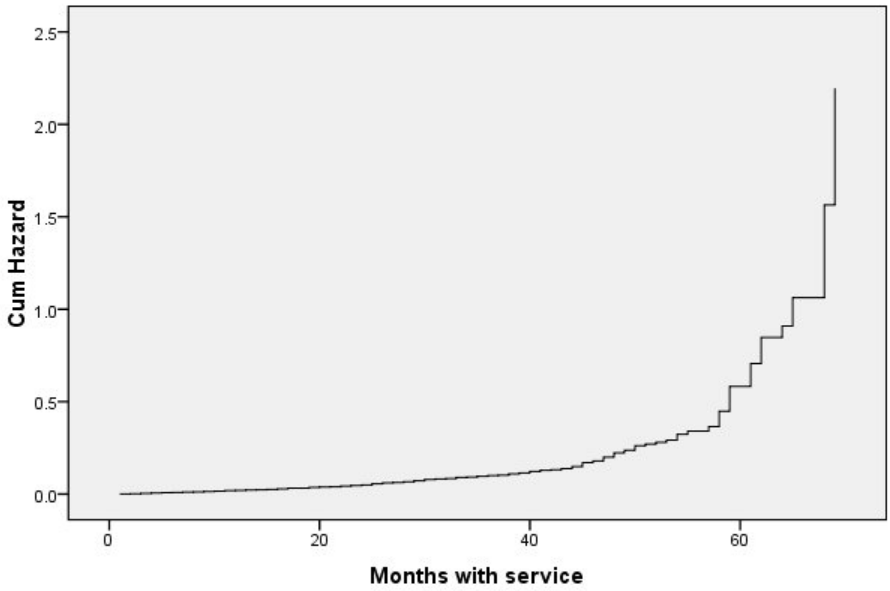
Krzywa przeżycia



Rysunek 355. Krzywa przeżycia dla „przeciętnego" klienta

Podstawowa krzywa przeżycia jest wizualną prezentacją przewidzianego przez model czasu do odejścia dla „przeciętnego" klienta. Oś pozioma przedstawia czas do zdarzenia. Oś pionowa przedstawia prawdopodobieństwo przeżycia. W ten sposób dowolny punkt na krzywej przeżycia pokazuje prawdopodobieństwo, że „przeciętny" klient pozostanie klientem po tym czasie. Po upływie 55 miesięcy krzywa przeżycia staje się mniej gładka. Istnieje mniejsza liczba klientów, którzy byli z firmą tak długo, więc dostępnych jest mniej informacji i krzywa jest kanciasta.

Krzywa hazardu

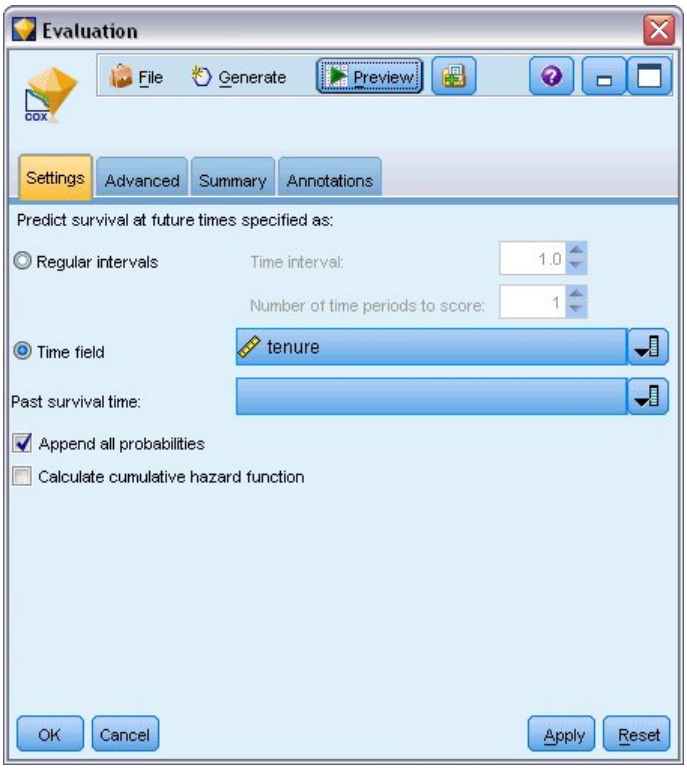


Rysunek 356. Krzywa hazardu dla „przeciętnego" klienta

Podstawowa krzywa hazardu jest wizualną prezentacją przewidzianego przez model skumulowanego potencjału odejścia dla „przeciętnego” klienta. Oś pozioma przedstawia czas do zdarzenia. Oś pionowa przedstawia skumulowane ryzyko, równe negatywnemu logarytmowi prawdopodobieństwa przeżycia. Po upływie 55 miesięcy krzywa hazardu, tak jak krzywa przeżycia, staje się mniej gładka z tego samego powodu.

Ocena

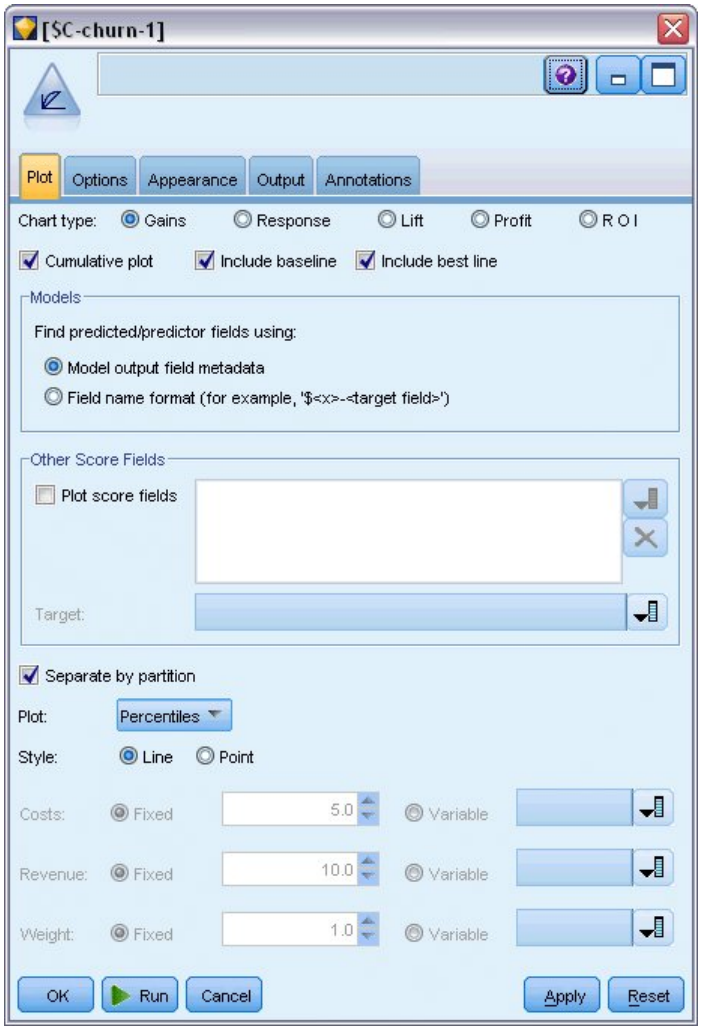
Krokowe metody wyboru gwarantują, że model będzie zawierał tylko statystycznie istotne predyktory, ale nie gwarantują, że model jest rzeczywiście skuteczny w przewidywaniu zmiennej przewidywanej. W tym celu należy przeanalizować ocenione rekordy:



Rysunek 357. Model użytkowy Coxa: karta Ustawienia

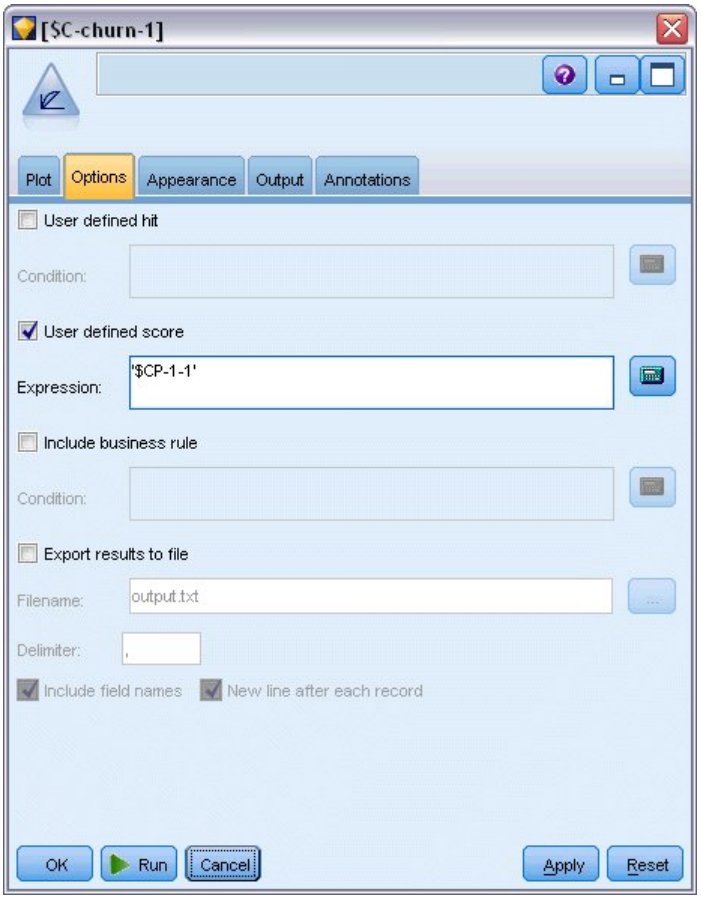
1. Umieść model użytkowy w obszarze roboczym i załącz go do węzła źródłowego, otwórz model użytkowy i kliknij kartę Ustawienia.
2. Zaznacz opcję **Zmienna czasu** i określ zmienną *tenure*. Każdy rekord będzie oceniany dla swojej długości zmiennej tenure.
3. Zaznacz opcję **Dołącz wszystkie prawdopodobieństwa**.

   Powoduje to utworzenie ocen używających wartości 0,5 jako punktu odcięcia dla tego, czy klient odchodzi. Jeśli skłonność do odejścia jest większa niż 0,5, klient jest oceniany jako klient odchodzący. Nie ma nic specjalnego w tej liczbie i inny punkt odcięcia może dać bardziej pożądane wyniki. Jednym ze sposobów wyboru punktu odcięcia jest użycie węzła ewaluacji.



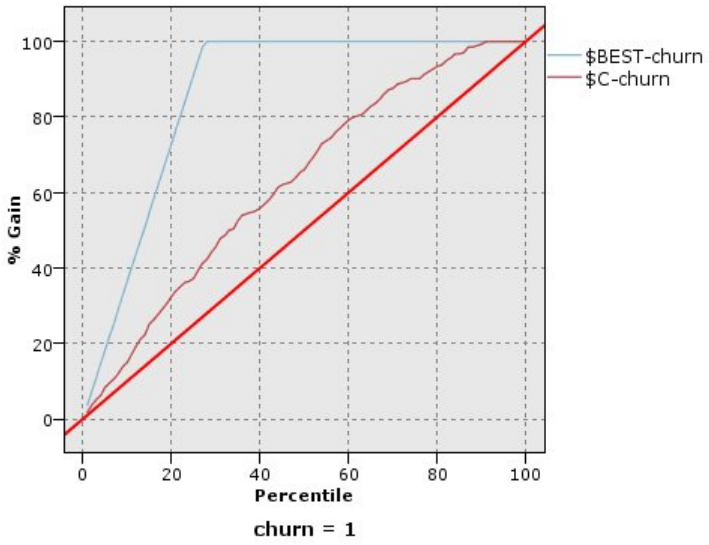
Rysunek 358. Węzeł ewaluacji: karta Wykres

4. Dołącz węzeł ewaluacji do modelu użytkowego. Na karcie Wykres zaznacz opcję **Pokaż linię najlepszych wartości**.
5. Kliknij kartę **Opcje**.



Rysunek 359. Węzeł ewaluacji: karta Opcje

6. Zaznacz opcję **Ocena definiowana przez użytkownika** i wpisz '$CP-1-1' jako wyrażenie. Jest to zmienna wygenerowana przez model odpowiadająca skłonności do odejścia.
7. Kliknij przycisk **Uruchom**.

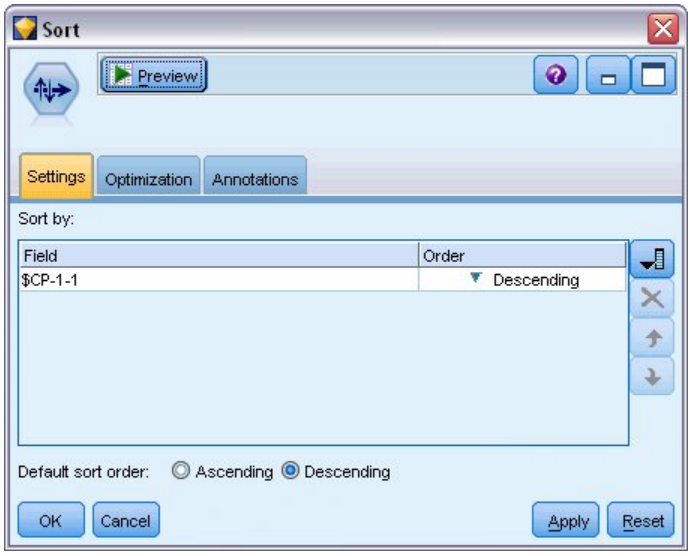


Rysunek 360. Wykres korzyści

Wykres skumulowanej korzyści przedstawia wartość procentową ogólnej liczby obserwacji w danej kategorii „uzyskaną” przez nakierowanie na procentowy udział całkowitej liczby obserwacji. Na przykład jeden z punktów na krzywej znajduje się w miejscu (10%, 15%), co oznacza, że po ocenie zbioru danych z modelem i posortowaniu wszystkich obserwacji według przewidywanej skłonności do odejścia, oczekuje się, że górne 10% zawiera w przybliżeniu 15% wszystkich obserwacji tworzących kategorię *1* (klienci odchodzący). W podobny sposób górne 60% zawiera około 79,2% klientów odchodzących. Po wybraniu 100% ocenionego zbioru danych uzyskiwani są wszyscy klienci odchodzący w zbiorze danych.

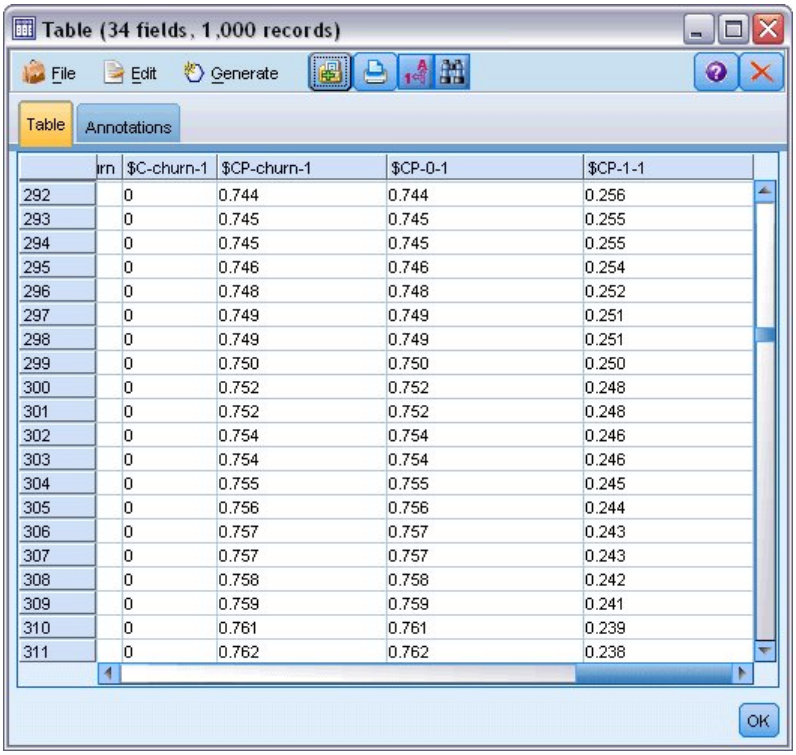
Przekątna jest krzywą „podstawy”. Po losowym wybraniu 20% rekordów z ocenionego zbioru danych można oczekiwać uzyskania mniej więcej 20% wszystkich rekordów z kategorii *1*. Im wyżej nad podstawą znajduje się krzywa, tym większe korzyści. Linia „best” przedstawia krzywą dla idealnego modelu, który przypisuje wyższą ocenę skłonności do odejścia dla każdego klienta odchodzącego niż dla każdego klienta nieodchodzącego. Można użyć wykresu skumulowanej korzyści, aby pomóc w klasyfikacji punktu odcięcia, wybierając wartość procentową, która odpowiada pożądanym korzyściom, a następnie odwzorowując tę wartość procentową na odpowiedniej wartości punktu odcięcia.

To, co stanowi „pożądane” korzyści, zależy od kosztu błędów Typu I i Typu II. Czyli jaki jest koszt zaklasyfikowania klienta odchodzącego jako klienta nieodchodzącego (Typ I)? Jaki jest koszt zaklasyfikowania klienta nieodchodzącego jako klienta odchodzącego (Typ II)? Jeśli najważniejsze jest utrzymanie klienta, to należy obniżyć błąd Typu I. Na wykresie skumulowanej korzyści może to odpowiadać zwiększonej obsłudze klienta w górnych 60% przewidywanej skłonności *1*, co obejmuje 79,2% potencjalnych klientów odchodzących, ale kosztem są czas i zasoby, które można poświęcić na zdobywanie nowych klientów. Jeśli priorytetem jest obniżenie kosztu utrzymania bieżącej bazy klientów, należy obniżyć błąd Typu II. Na tym wykresie może to odpowiadać zwiększonej obsłudze klienta z górnych 20%, co obejmuje 32,5% klientów odchodzących. Zazwyczaj obie kwestie są istotne, należy więc wybrać regułę decyzyjną do klasyfikacji klientów, która zapewnia najlepszą mieszankę czułości i swoistości.



Rysunek 361. Węzeł sortowania: karta Ustawienia

8. Załóżmy, że użytkownik zadecydował, że 45,6% to pożądane korzyści, co odpowiada górnym 30% rekordów. Aby znaleźć odpowiedni punkt odcięcia klasyfikacji, załącz węzeł sortowania do modelu użytkowego.
9. Na karcie Ustawienia wybierz sortowanie według zmiennej *$CP-1-1* w kolejności malejącej i kliknij przycisk **OK**.



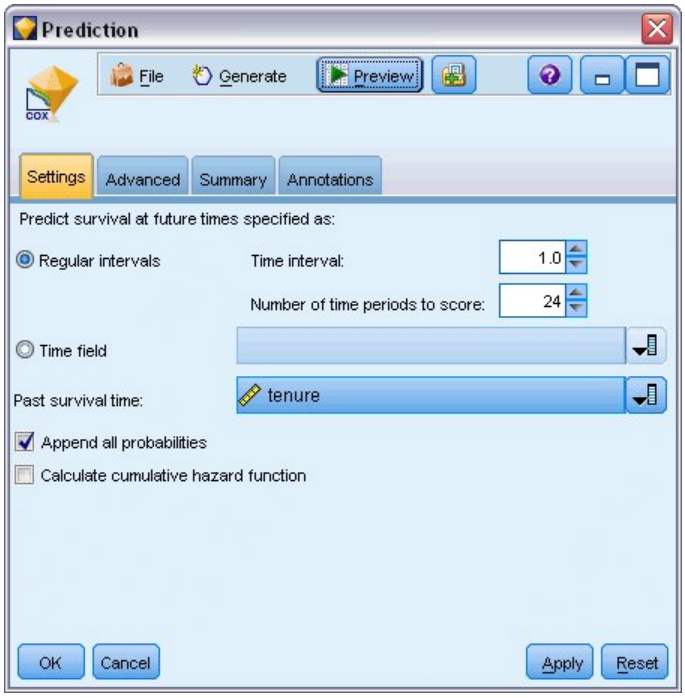
Rysunek 362. Tabela

10. Dołącz węzeł tabeli do węzła sortowania.
11. Otwórz węzeł Tabela i kliknij przycisk **Uruchom**.

Przewijając w dół wyniki, można zobaczyć, że wartość zmiennej *$CP-1-1* to 0,248 dla 300. rekordu. Użycie 0,248 jako punktu odcięcia klasyfikacji powinno spowodować ocenienie około 30% klientów jako klientów odchodzących, przy rzeczywistej liczbie klientów odchodzących na poziomie 45%.

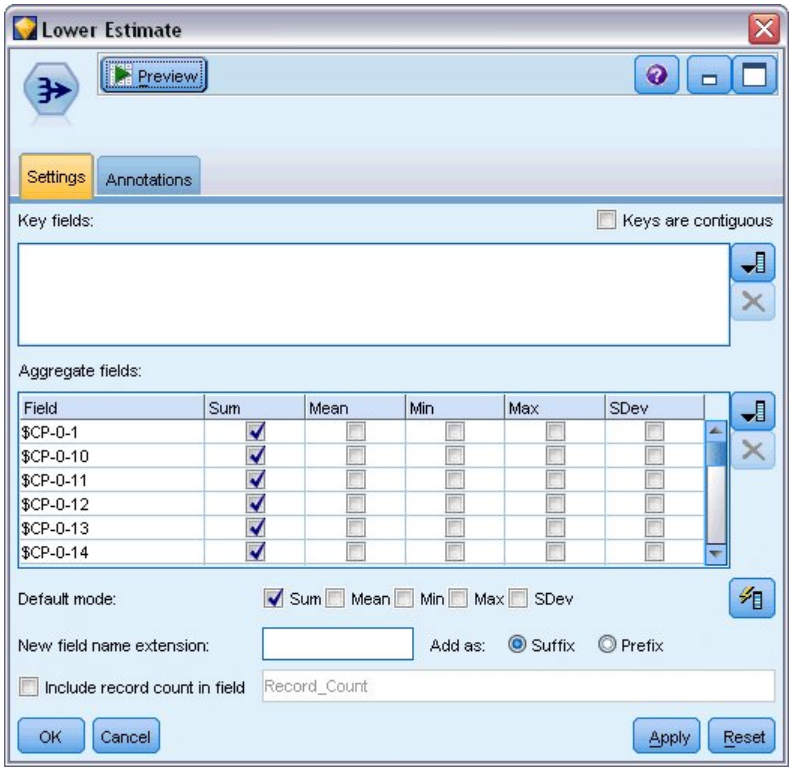
Śledzenie oczekiwanej liczby utrzymanych klientów

Po otrzymaniu satysfakcjonującego modelu można śledzić oczekiwaną liczbę klientów w zbiorze danych, którzy zostali utrzymani przez następne dwa lata. Interesujące wyzwanie stanowią wartości null oznaczające klientów, których całkowity okres korzystania z usługi (przyszły czas + zmienna *tenure*) przypada poza zakresem czasu przeżycia w danych używanych do uczenia modelu. Jednym sposobem na rozwiązanie tego problemu jest utworzenie dwóch zestawów predykcji. Jednego, w którym zakłada się, że wartości null odeszły, i drugiego, w którym zakłada się, że klienci zostali utrzymani. W ten sposób można określić górną i dolną granicę oczekiwanej liczby utrzymanych klientów.



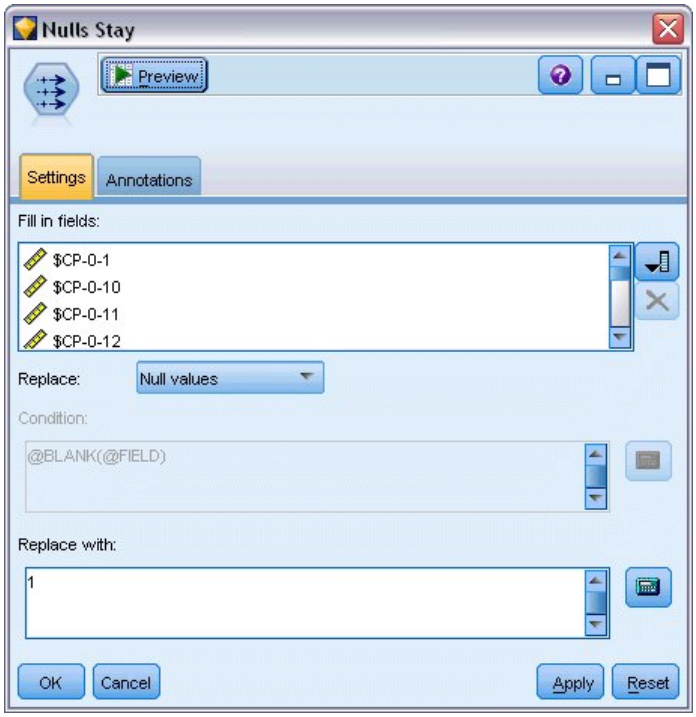
Rysunek 363. Model użytkowy Coxa: karta Ustawienia

1. Dwukrotnie kliknij model użytkowy na palecie modeli (lub skopiuj i wklej model użytkowy w obszarze roboczym strumienia) i dołącz nowy model użytkowy do węzła źródłowego.
2. Otwórz kartę Ustawienia modelu użytkowego.
3. Upewnij się, że zaznaczona jest opcja **Regularne przedziały** i określ wartość 1,0 jako przedział czasowy oraz 24 jako liczbę okresów do oceny. Określa to, że każdy rekord będzie oceniany w każdym z następujących 24 miesięcy.
4. Wybierz *tenure* jako zmienną określającą poprzedni czas przeżycia. Algorytm oceniania uwzględni czas, przez jaki każdy klient jest klientem firmy.
5. Zaznacz opcję **Dołącz wszystkie prawdopodobieństwa**.



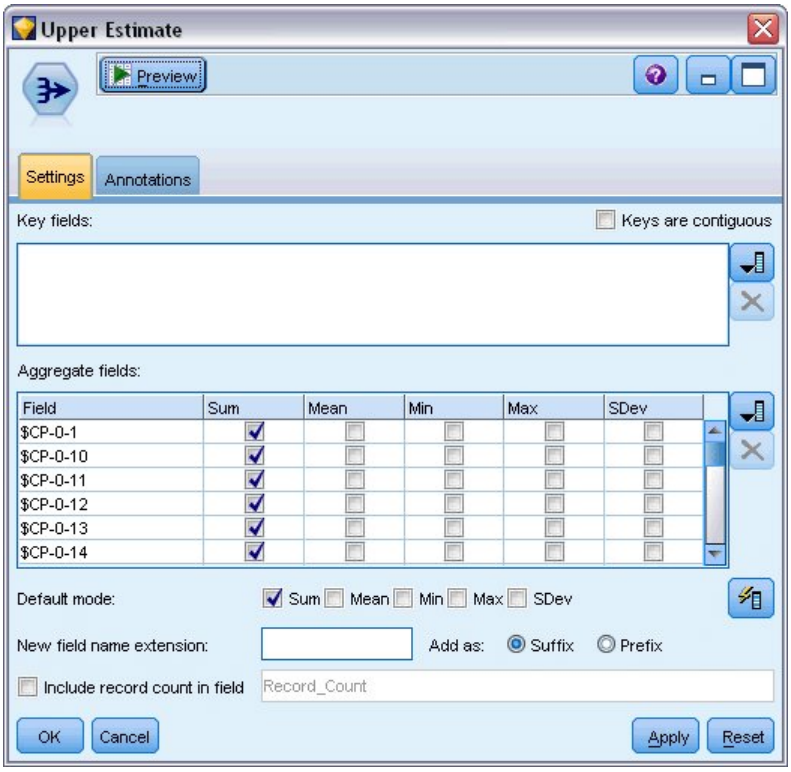
Rysunek 364. Węzeł Agregacja: karta Ustawienia

6. Dołącz węzeł Agregacja do modelu użytkowego. Na karcie Ustawienia usuń zaznaczenie pozycji **Średnia** jako trybu domyślnego.
7. Zaznacz zmienne od *$CP-0-1* do *$CP-0-24*, zmienne w formie *$CP-0-n* jako zmienne do zagregowania. Jest to najłatwiejsze, jeśli w oknie dialogowym Wybierz zmienne posortujesz zmienne według nazwy (tzn. w kolejności alfabetycznej).
8. Usuń zaznaczenie opcji **Dołącz liczebność rekordów w zmiennej**.
9. Kliknij przycisk **OK**. Ten węzeł tworzy predykcje dolnej granicy.



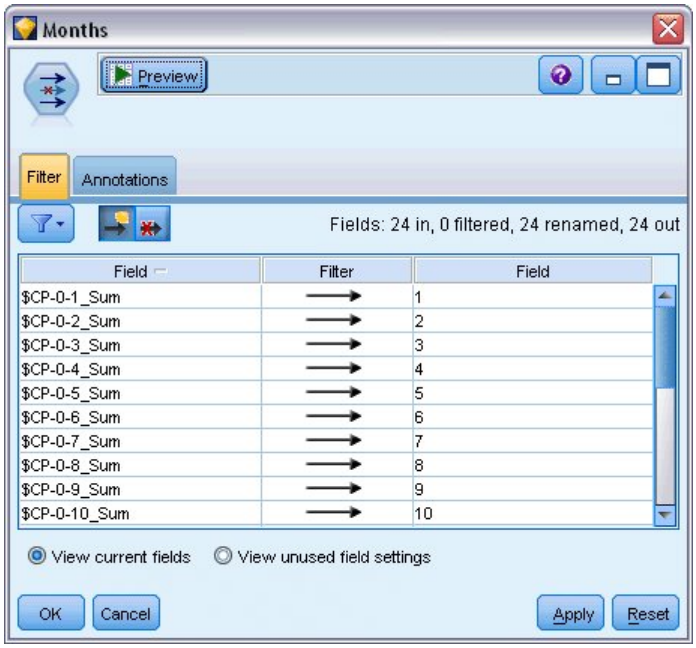
Rysunek 365. Węzeł Wypełnianie: karta Ustawienia

10. Załącz węzeł filtrowania do modelu użytkowego Coxreg, do którego właśnie dołączyliśmy węzeł Agregacja. Na karcie Ustawienia zaznacz zmienne od *$CP-0-1* do *$CP-0-24*, zmienne w formie *$CP-0-n* jako zmienne do wypełnienia. Jest to najłatwiejsze, jeśli w oknie dialogowym Wybierz zmienne posortujesz zmienne według nazwy (tzn. w kolejności alfabetycznej).
11. Wybierz zastąpienie **Wartości null** wartością 1.
12. Kliknij przycisk **OK**.



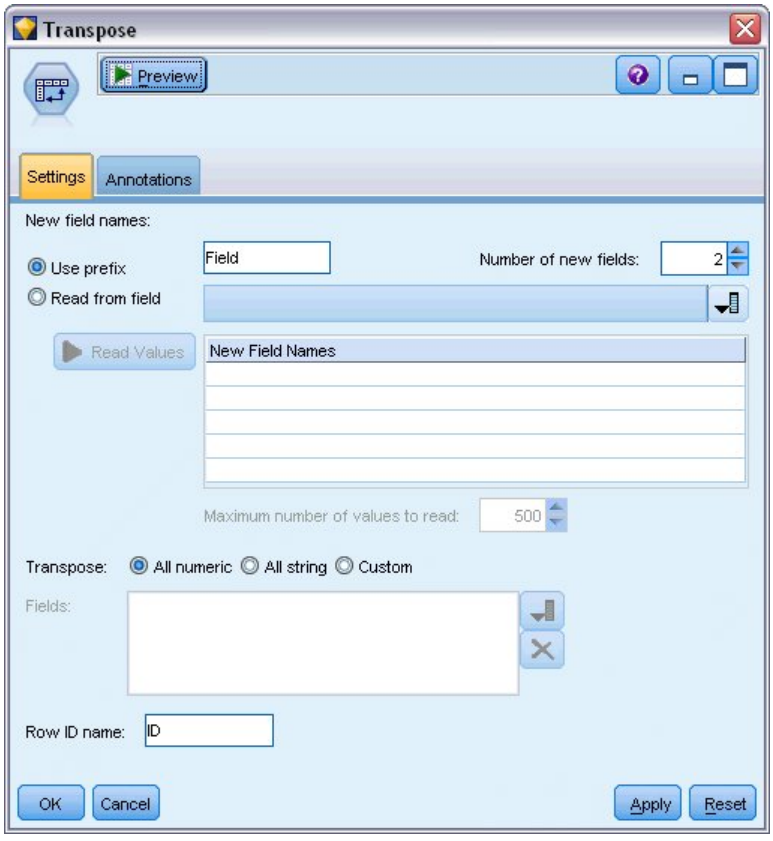
Rysunek 366. Węzeł Agregacja: karta Ustawienia

13. Dołącz węzeł Agregacja do węzła wypełniania. Na karcie Ustawienia usuń zaznaczenie pozycji **Średnia** jako trybu domyślnego.
14. Zaznacz zmienne od *$CP-0-1* do *$CP-0-24*, zmienne w formie *$CP-0-n* jako zmienne do zagregowania. Jest to najłatwiejsze, jeśli w oknie dialogowym Wybierz zmienne posortujesz zmienne według nazwy (tzn. w kolejności alfabetycznej).
15. Usuń zaznaczenie opcji **Dołącz liczebność rekordów w zmiennej**.
16. Kliknij przycisk **OK**. Ten węzeł tworzy predykcje górnej granicy.



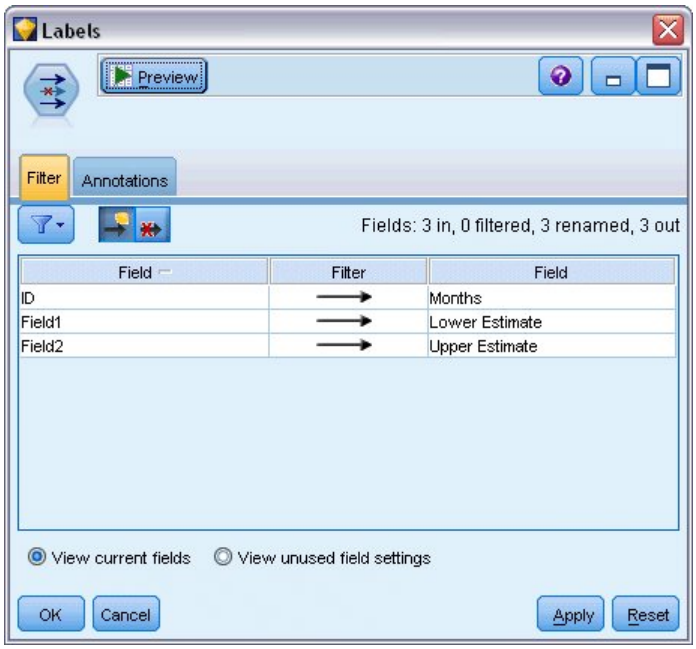
Rysunek 367. Węzeł filtrowania: karta Ustawienia

17. Załącz węzeł załączania do dwóch węzłów Agregacja, a następnie załącz węzeł filtrowania do węzła Dołączanie.
18. Na karcie Ustawienia węzła Filtr zmień nazwy zmiennych na nazwy od *1* do *24*. Za pomocą węzła Transpozycja te nazwy zmiennych staną się wartościami dla osi *x* na wykresach w dalszej części strumienia.



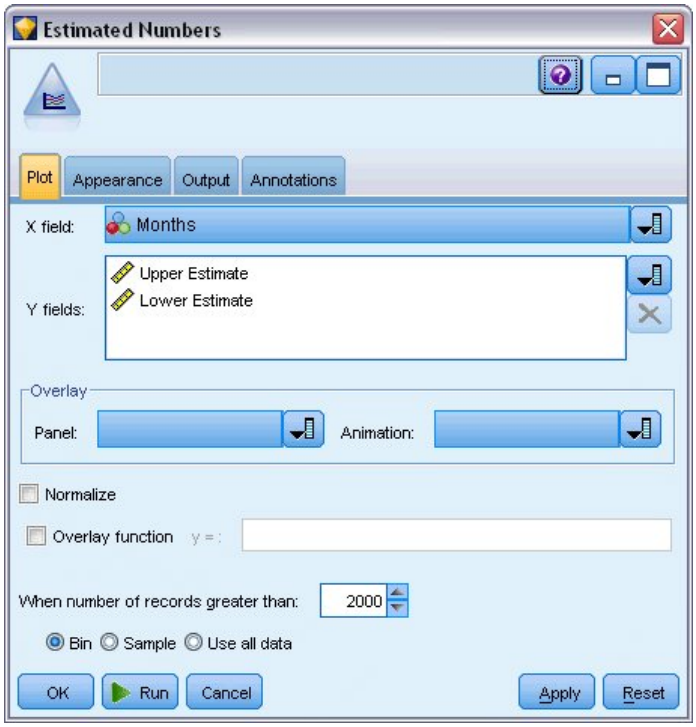
Rysunek 368. Węzeł Transpozycja: karta Ustawienia

19. Załącz węzeł Transpozycja do węzła filtrowania.
20. Wpisz 2 jako liczbę nowych zmiennych.



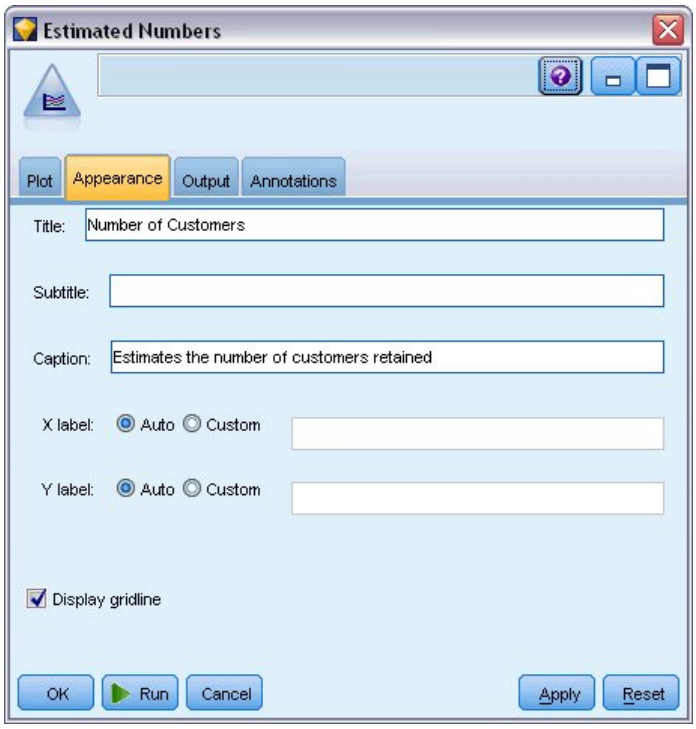
Rysunek 369. Węzeł filtrowania: karta Filtrowanie

21. Załącz węzeł filtrowania do węzła Transpozycja.
22. Na karcie Ustawienia węzła filtrowania zmień nazwę zmiennej *ID* na *Months*, *Field1* na *Lower Estimate* i *Field2* na *Upper Estimate*.



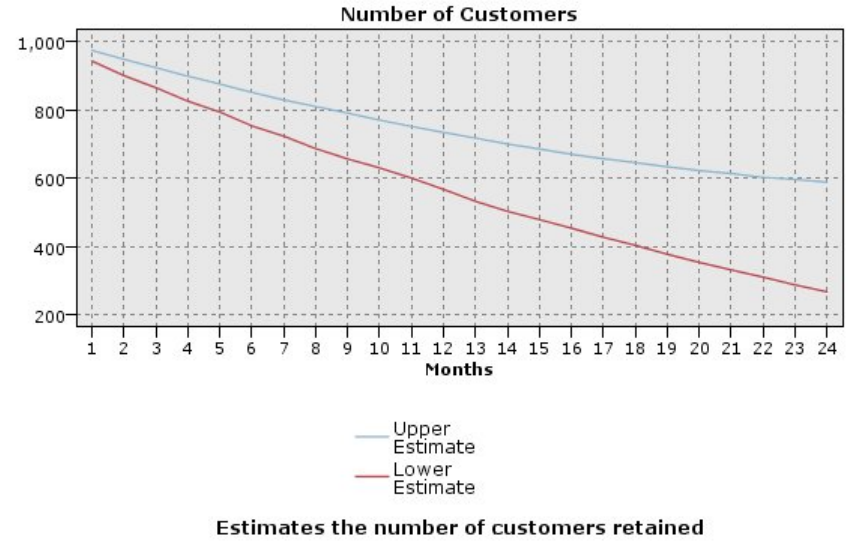
Rysunek 370. Węzeł Liniowy: karta Wykres

23. Załącz węzeł Liniowy do węzła filtrowania.
24. Na karcie Wykres ustaw zmienną *Months* w obszarze Zmienna X oraz *Lower Estimate* i *Upper Estimate* w obszarze Zmienne Y.



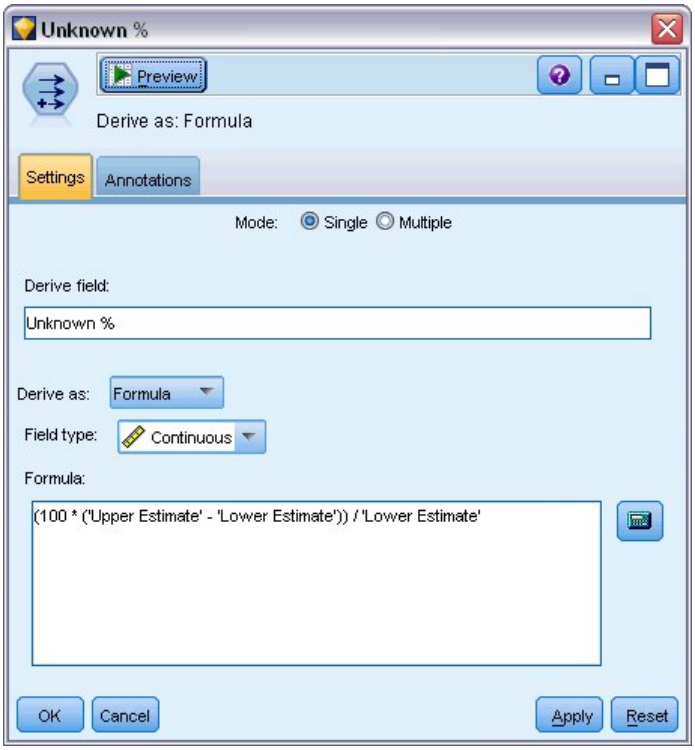
Rysunek 371. Węzeł Liniowy: karta Wygląd

25. Kliknij kartę Wygląd.
26. Wpisz Number of Customers jako tytuł.
27. Wpisz Estimates the number of customers retained jako podpis.
28. Kliknij przycisk **Uruchom**.



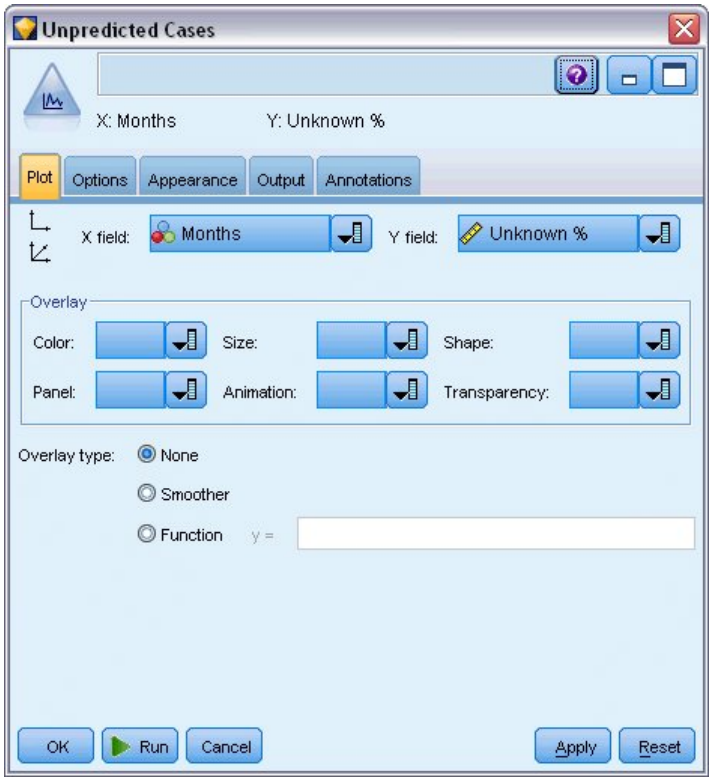
Rysunek 372. Wykres wielokrotny szacujący liczbę utrzymanych klientów

Na wykresie rysowana jest górna i dolna granica szacowanej liczby utrzymanych klientów. Różnica pomiędzy dwiema liniami to liczba klientów ocenionych jako null, czyli klientów, których status jest wysoce niepewny. W miarę upływu czasu liczba tych klientów rośnie. Po 12 miesiącach można oczekiwać utrzymania pomiędzy 601 i 735 oryginalnych klientów w zbiorze danych. Po 24 miesiącach pomiędzy 288 i 597.



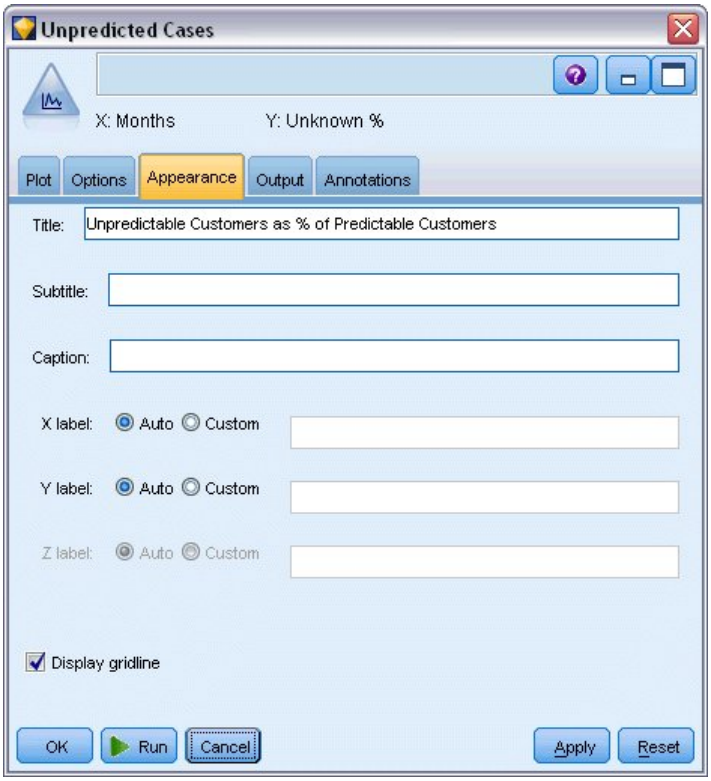
Rysunek 373. Węzeł wyliczeń: karta Ustawienia

29. Aby uzyskać kolejny wgląd w to, jak niepewne są oszacowania liczby utrzymanych klientów, załącz węzeł wyliczeń do węzła filtrowania.
30. Na karcie Ustawienia węzła wyliczeń wpisz *Unknown %* jako zmienną wyliczaną.
31. Jako typ zmiennej wybierz opcję **Ilościowa**.
32. Wpisz (100 * ('Upper Estimate' - 'Lower Estimate'))/ 'Lower Estimate' jako formułę. *Unknown %* to liczba „wątpliwych” klientów jako wartość procentowa dolnego oszacowania.
33. Kliknij przycisk **OK**.



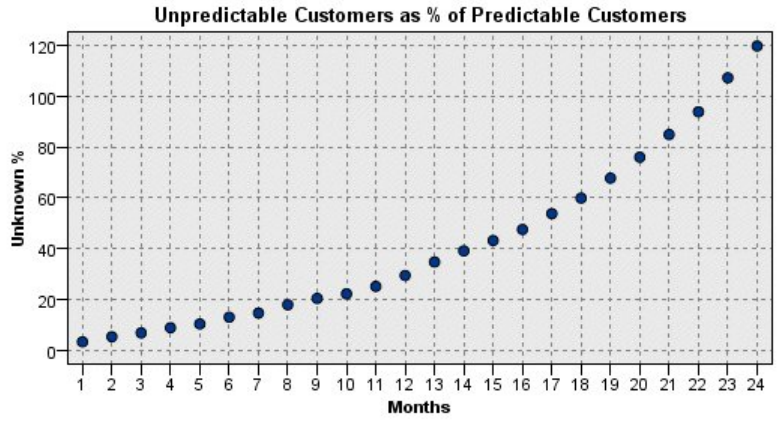
Rysunek 374. Węzeł wykresu: karta Wykres

34. Załącz węzeł wykresu do węzła wyliczeń.
35. Na karcie Wykres węzła wykresu wybierz zmienną *Months* jako zmienną X i *Unknown %* jako zmienną Y.
36. Kliknij kartę **Wygląd**.



Rysunek 375. Węzeł wykresu: karta Wygląd

37. Wpisz Unpredictable Customers as % of Predictable Customers jako tytuł.
38. Wykonaj węzeł.

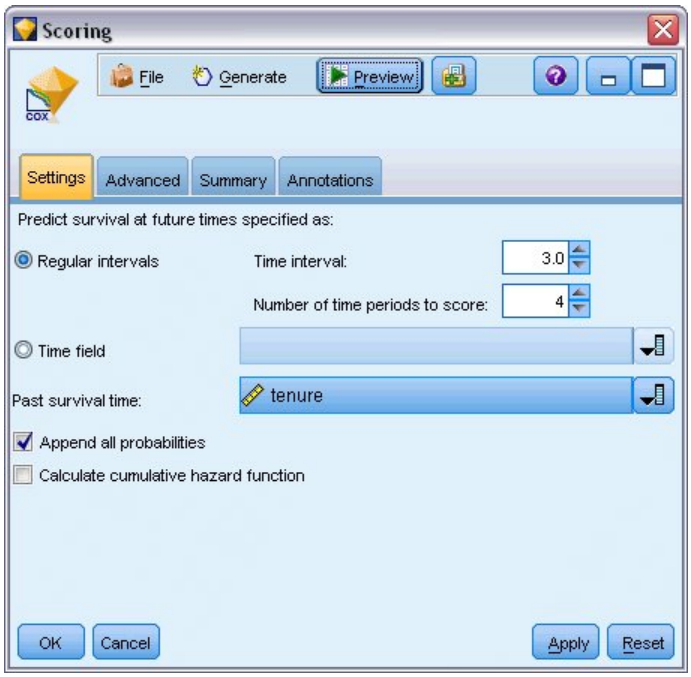


Rysunek 376. Wykres nieprzewidywalnych klientów

W pierwszym roku procentowy udział nieprzewidywalnych klientów rośnie w tempie mniej więcej liniowym, ale współczynnik wzrostu gwałtownie zwiększa się w drugim roku, gdzie w 23. miesiącu liczba klientów z wartościami null przewyższa oczekiwaną liczbę utrzymanych klientów.

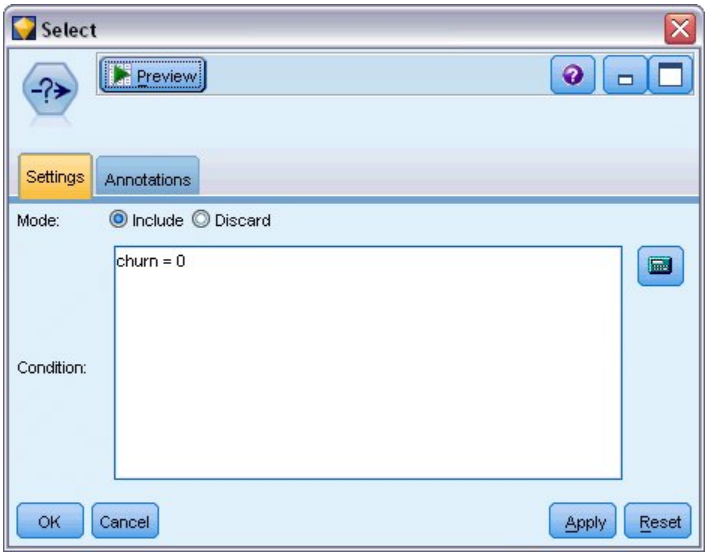
Ocenianie

Po otrzymaniu satysfakcjonującego modelu można ocenić klientów, aby zidentyfikować osoby z największym prawdopodobieństwem odejścia w następnym roku według kwartału.



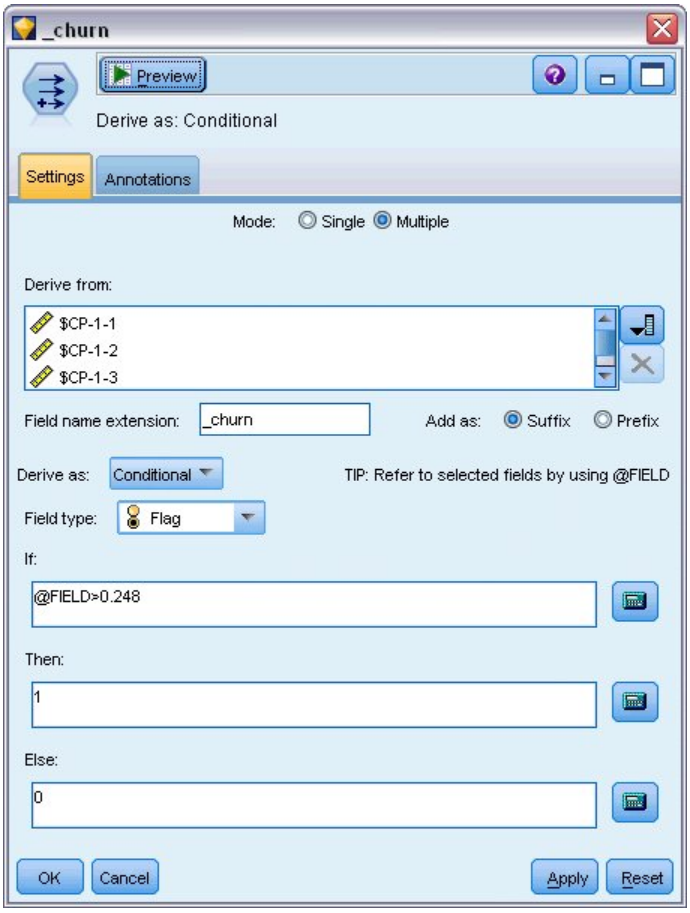
Rysunek 377. Model użytkowy Coxreg: karta Ustawienia

1. Załącz trzeci model użytkowy do węzła źródłowego i otwórz model użytkowy.
2. Upewnij się, że zaznaczona jest opcja **Regularne przedziały** i określ wartość 3,0 jako przedział czasowy oraz 4 jako liczbę okresów do oceny. Określa to, że każdy rekord będzie oceniany przez następne cztery kwartały.
3. Wybierz *tenure* jako zmienną określającą poprzedni czas przeżycia. Algorytm oceniania uwzględni czas, przez jaki każdy klient jest klientem firmy.
4. Zaznacz opcję **Dołącz wszystkie prawdopodobieństwa**. Te dodatkowe zmienne ułatwią sortowanie rekordów do wyświetlania w tabeli.



Rysunek 378. Węzeł selekcji: karta Ustawienia

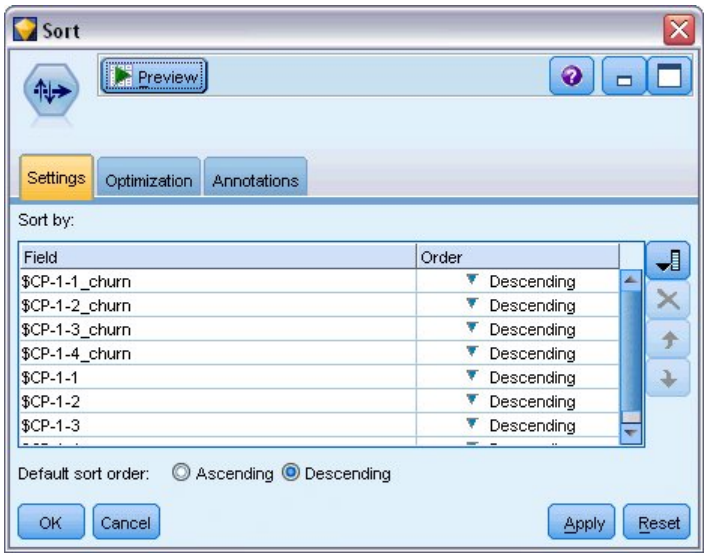
5. Załącz węzeł selekcji do modelu użytkowego. Na karcie Ustawienia wpisz churn=0 jako warunek. Powoduje to usunięcie z tabeli wyników klientów, którzy już odeszli.



Rysunek 379. Węzeł wyliczeń: karta Ustawienia

6. Załącz węzeł wyliczeń do węzła selekcji. Na karcie Ustawienia wybierz tryb **Wielokrotny**.

7. Wybierz, aby wyliczać ze zmiennych od *$CP-1-1* do *$CP-1-4*, zmiennych w formie *$CP-1-n*, a następnie wpisz _churn jako dodawany przyrostek. Jest to najłatwiejsze, jeśli w oknie dialogowym Wybierz zmienne posortujesz zmienne według nazwy (tzn. w kolejności alfabetycznej).
8. Wybierz wyliczanie zmiennej jako **Warunkowe**.
9. Jako poziom pomiaru wybierz opcję **Flaga**.
10. Wpisz @FIELD>0,248 jako warunek **Jeśli**. Przypomnijmy sobie, że to był punkt odcięcia klasyfikacji zidentyfikowany podczas Ewaluacji.
11. Wpisz 1 jako wyrażenie **To**.
12. Wpisz 0 jako wyrażenie **Inaczej**.
13. Kliknij przycisk **OK**.

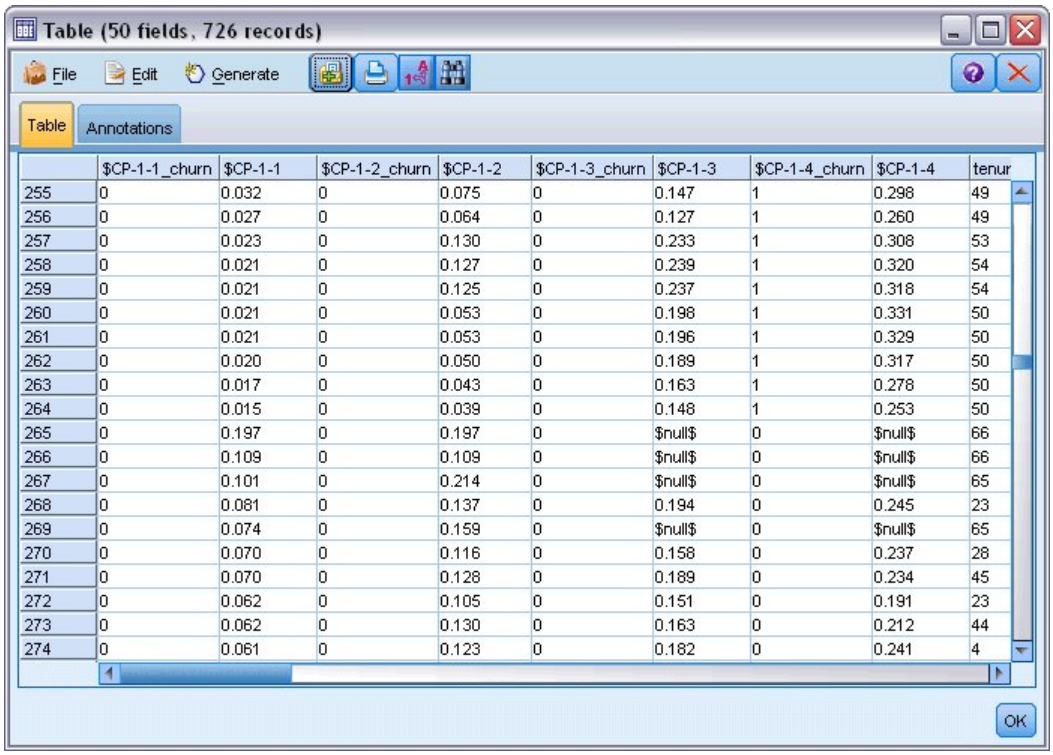


Rysunek 380. Węzeł sortowania: karta Ustawienia

14. Dołącz węzeł sortowania do węzła wyliczeń. Na karcie Ustawienia wybierz sortowanie według zmiennych od *$CP-1-1_churn* do *$CP-1-4-churn*, a następnie od *$CP-1-1* do *$CP-1-4* w kolejności malejącej. Klienci, dla których przewiduje się odejście, pojawią się na górze.

Rysunek 381. Węzeł Reorganizacja: karta Reorganizacja

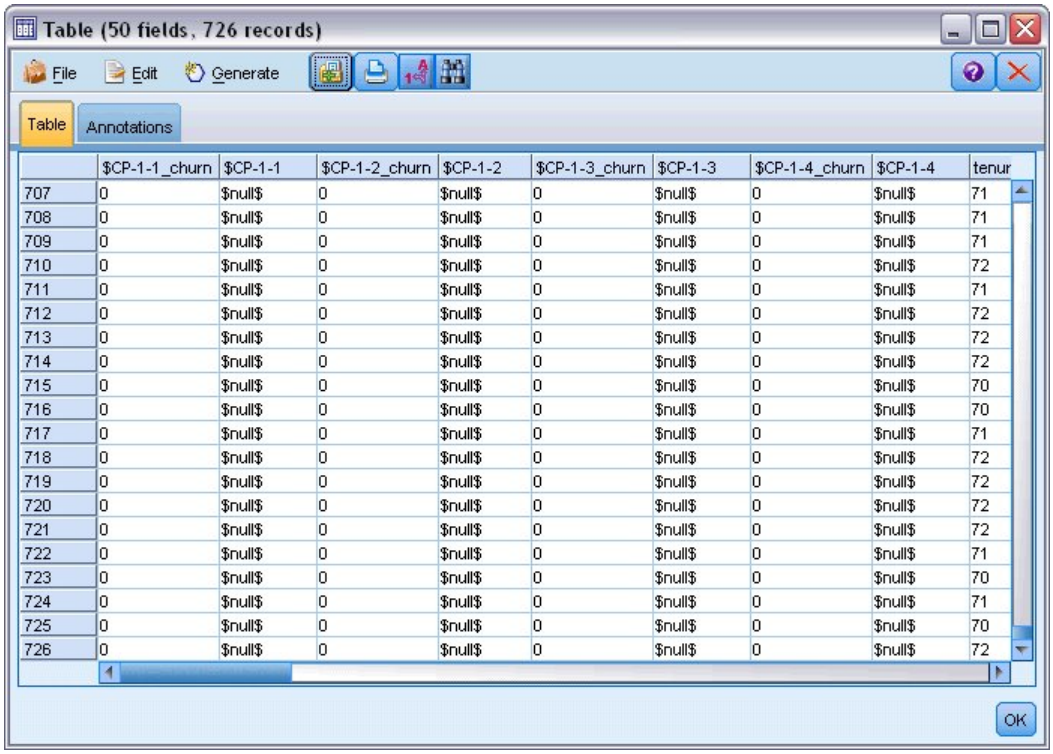
15. Załącz węzeł Reorganizacja do węzła sortowania. Na karcie Reorganizacja umieść zmienne od *$CP-1-1_churn* do *$CP-1-4* przed innymi zmiennymi. Ułatwia to tylko odczytanie wyników tabeli, więc jest to czynność opcjonalna. Należy użyć przycisków, aby przenieść zmienne na pozycję przedstawioną na rysunku.



Rysunek 382. Tabela przedstawiająca oceny klientów

16. Dołącz węzeł tabeli do węzła Reorganizacja i uruchom go.

Oczekuje się, że 264 klientów odejdzie przed końcem roku, 184 przed końcem trzeciego kwartału, 103 przed końcem drugiego kwartału i 31 w pierwszym kwartale. Należy zauważyć, biorąc pod uwagę dwóch klientów, że klient z większą skłonnością do odejścia w pierwszym kwartale nie musi mieć koniecznie większej skłonności do odejścia w późniejszych kwartałach. Zobacz na przykład rekordy 256 i 260. Dzieje się tak prawdopodobnie z powodu kształtu funkcji hazardu dla miesięcy, które następują po bieżącym okresie korzystania z usługi przez klienta. Na przykład klienci, którzy dołączyli z powodu promocji, mogą zmienić dostawcę wcześniej niż klienci, którzy dołączyli na podstawie polecenia osobistego, ale jeśli nie dokonają zmiany, mogą być w rzeczywistości bardziej lojalni przez pozostały okres. Możesz wykonać ponowne sortowanie klientów, aby uzyskać różne widoki klientów o największym prawdopodobieństwie odejścia.



	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
707	0	$null$	0	$null$	0	$null$	0	$null$	71
708	0	$null$	0	$null$	0	$null$	0	$null$	71
709	0	$null$	0	$null$	0	$null$	0	$null$	71
710	0	$null$	0	$null$	0	$null$	0	$null$	72
711	0	$null$	0	$null$	0	$null$	0	$null$	71
712	0	$null$	0	$null$	0	$null$	0	$null$	72
713	0	$null$	0	$null$	0	$null$	0	$null$	72
714	0	$null$	0	$null$	0	$null$	0	$null$	72
715	0	$null$	0	$null$	0	$null$	0	$null$	70
716	0	$null$	0	$null$	0	$null$	0	$null$	70
717	0	$null$	0	$null$	0	$null$	0	$null$	71
718	0	$null$	0	$null$	0	$null$	0	$null$	72
719	0	$null$	0	$null$	0	$null$	0	$null$	72
720	0	$null$	0	$null$	0	$null$	0	$null$	72
721	0	$null$	0	$null$	0	$null$	0	$null$	72
722	0	$null$	0	$null$	0	$null$	0	$null$	71
723	0	$null$	0	$null$	0	$null$	0	$null$	70
724	0	$null$	0	$null$	0	$null$	0	$null$	71
725	0	$null$	0	$null$	0	$null$	0	$null$	70
726	0	$null$	0	$null$	0	$null$	0	$null$	72

Rysunek 383. Tabela przedstawiająca klientów z wartościami null

W dolnej części tabeli znajdują się klienci z przewidywanymi wartościami null. Są to klienci, których całkowity okres korzystania z usługi (przyszły czas + zmienna *tenure*) przypada poza zakresem czasu przeżycia w danych używanych do uczenia modelu.

Podsumowanie

Używając regresji Coxa, znaleziono dopuszczalny model określenia czasu do odejścia, narysowano wykres oczekiwanej liczby klientów utrzymanych przez następne dwa lata i zidentyfikowano indywidualnych klientów, którzy z największym prawdopodobieństwem mogą odejść w przyszłym roku. Należy zauważyć, że choć jest to dopuszczalny model, to może nie być najlepszy. Idealnie byłoby przynajmniej porównać ten model, uzyskany metodą krokową postępującą, z modelem uzyskanym za pomocą metody krokowej wstecznej.

Wyjaśnienia podstaw matematycznych metod modelowania używanych w programie IBM SPSS Modeler przedstawiono w publikacji *IBM SPSS Modeler Algorithms Guide*.

Rozdział 27. Analiza koszyka rynkowego (Wywodzenie reguł/C5.0)

Ten przykład dotyczy danych fikcyjnych opisujących zawartość koszyków w supermarkecie (tzn. zbiorów produktów kupowanych razem) oraz powiązane dane osobiste nabywcy, które można uzyskać dzięki programowi kart lojalnościowych. Celem jest wykrycie grup klientów, którzy kupują podobne produkty i mogą być scharakteryzowani demograficznie, np. za pomocą wieku, przychodów itp.

Ten przykład ilustruje dwie fazy eksploracji danych:

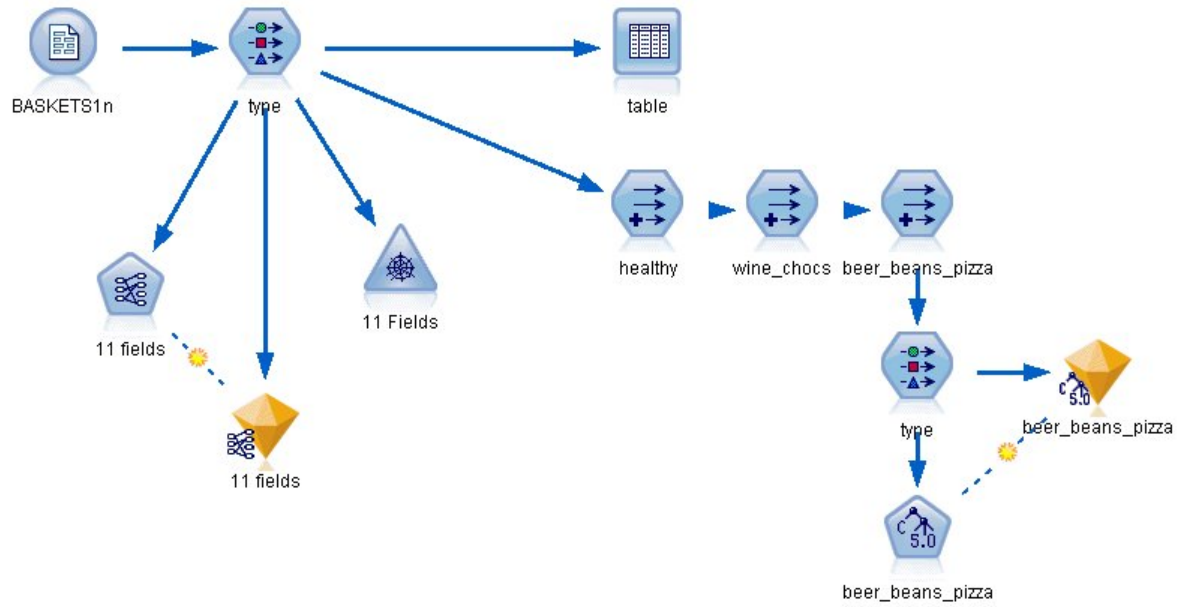
- Modelowanie reguł asocjacyjnych oraz wykres sieciowy ujawniające łącza pomiędzy zakupionymi produktami.
- Wywodzenie reguł C5.0 profilujące nabywców zidentyfikowanych grup produktów.

Uwaga: Ta aplikacja nie używa bezpośrednio modelowania predykcyjnego, więc nie udostępnia pomiarów modeli wynikowych oraz powiązanego rozróżnienia uczenie/test w procesie eksploracji danych.

W tym przykładzie zastosowano strumień o nazwie *baskrule*, który odwołuje się do pliku danych o nazwie *BASKETS1n*. Te pliki są dostępne w folderze *Demos* w katalogu instalacji programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *baskrule* znajduje się w katalogu *streams*.

Uzyskiwanie dostępu do danych

Używając węzła pliku zmiennych, połącz strumień ze zbiorem danych *BASKETS1n*, wybierając odczytanie nazw zmiennych z pliku. Podłącz węzeł typu do źródła danych, a następnie podłącz węzeł do węzła tabeli. Ustaw poziom pomiaru zmiennej *cardid* na *Nieokreślony* (ponieważ każdy identyfikator karty lojalnościowej występuje tylko raz w zbiorze danych i nie przyda się w modelowaniu). Ustaw *Nominalna* jako poziom pomiaru zmiennej *sex* (aby zapewnić, że algorytm modelowania Apriori nie będzie traktował zmiennej *sex* jako flagi).



Rysunek 384. Strumień baskrule

Teraz uruchom strumień, aby zrealizować węzeł typu i wyświetlić tabelę. Ten zbiór danych zawiera 18 zmiennych, z których każdy rekord reprezentuje koszyk.

18 zmiennych przedstawiono w następujących nagłówkach.

Podsumowanie koszyka:

- *cardid*. Identyfikator karty lojalnościowej klienta kupującego ten koszyk.
- *value*. Całkowita cena zakupu koszyka.
- *pmethod*. Metoda płatności za koszyk.

Dane osobowe posiadacza karty:

- *sex*
- *homeown*. Czy posiadacz karty jest właścicielem domu.
- *income*
- *age*

Zawartość koszyka: flagi dla obecności produktów z następujących kategorii:

- *fruitveg*
- *freshmeat*
- *dairy*
- *cannedveg*
- *cannedmeat*
- *frozenmeal*
- *beer*
- *wine*
- *softdrink*
- *fish*
- *confectionery*

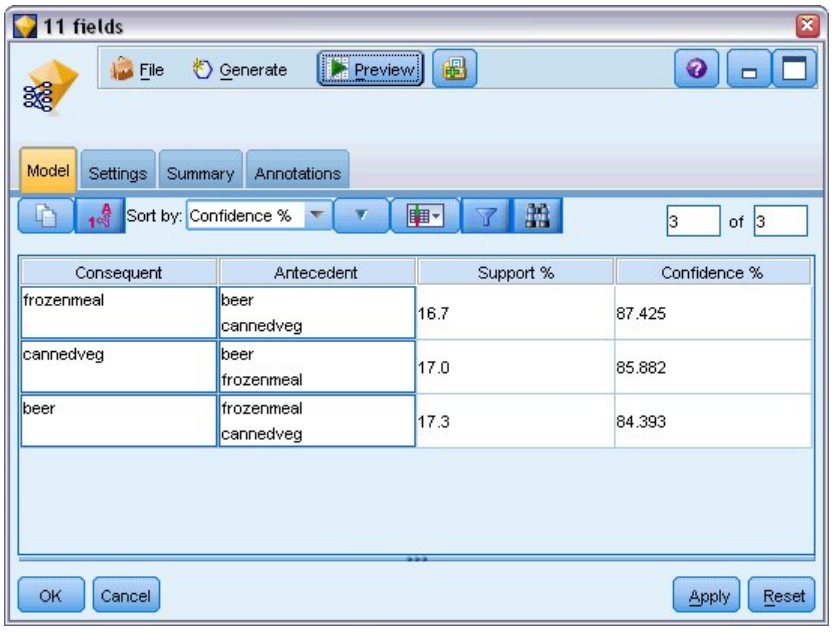
Odkrywanie podobieństw w zawartości koszyków

Najpierw należy uzyskać ogólny obraz podobieństw (powiązań) w zawartości koszyków, używając funkcji Apriori do wygenerowania reguł asocjacyjnych. Wybierz zmienne używane w procesie modelowania, edytując węzeł typu i ustawiając rolę wszystkich kategorii produktów na *Łącznie* i wszystkie pozostałe role na *Brak*. (*Łącznie* oznacza, że zmienna może być danymi wejściowymi lub wyjściowymi wynikowego modelu).

Uwaga: Można ustawić opcje dla wielu zmiennych, klikając z naciśniętym klawiszem Shift przed określeniem opcji w kolumnach.

Rysunek 385. Wybieranie zmiennych do modelowania

Po określeniu zmiennych do modelowania załącz węzeł Apriori do węzła typu, edytuj go, zaznacz opcję **Tylko wartości prawda dla flag** i kliknij przycisk Uruchom w węźle Apriori. Wynik — model na karcie Modele w prawym górnym rogu okna menedżerów — zawiera reguły asocjacyjne, które można przeglądać, używając menu kontekstowego i wybierając opcję **Przeglądaj**.



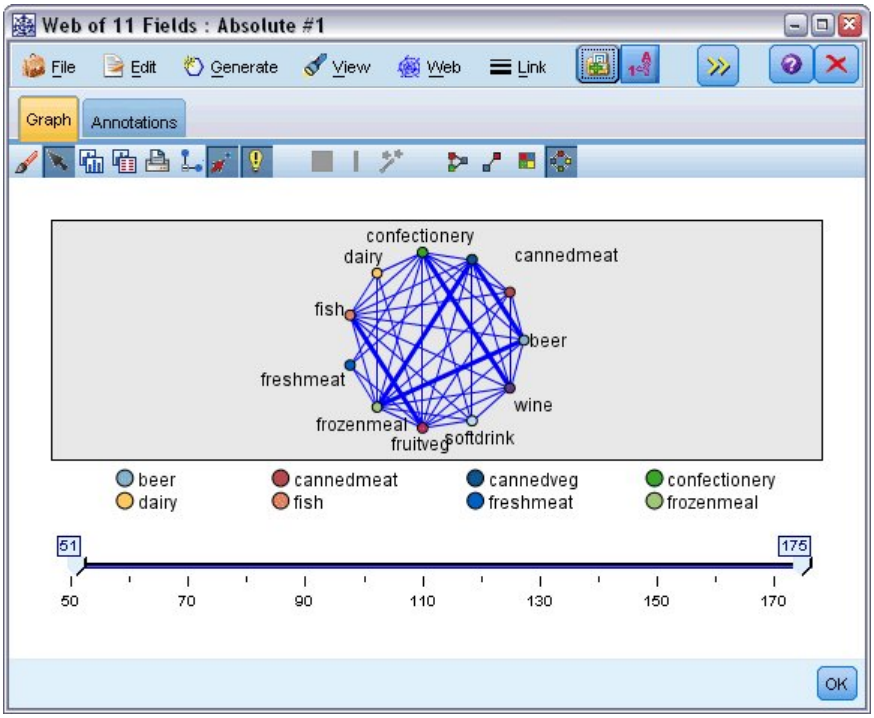
Rysunek 386. Reguły asocjacyjne

Te reguły wykazują wiele powiązań między mrożonymi posiłkami, warzywami w puszkach i piwem. Obecność dwukierunkowych reguł asocjacyjnych, takich jak:

```
frozenmeal -> beer
beer -> frozenmeal
```

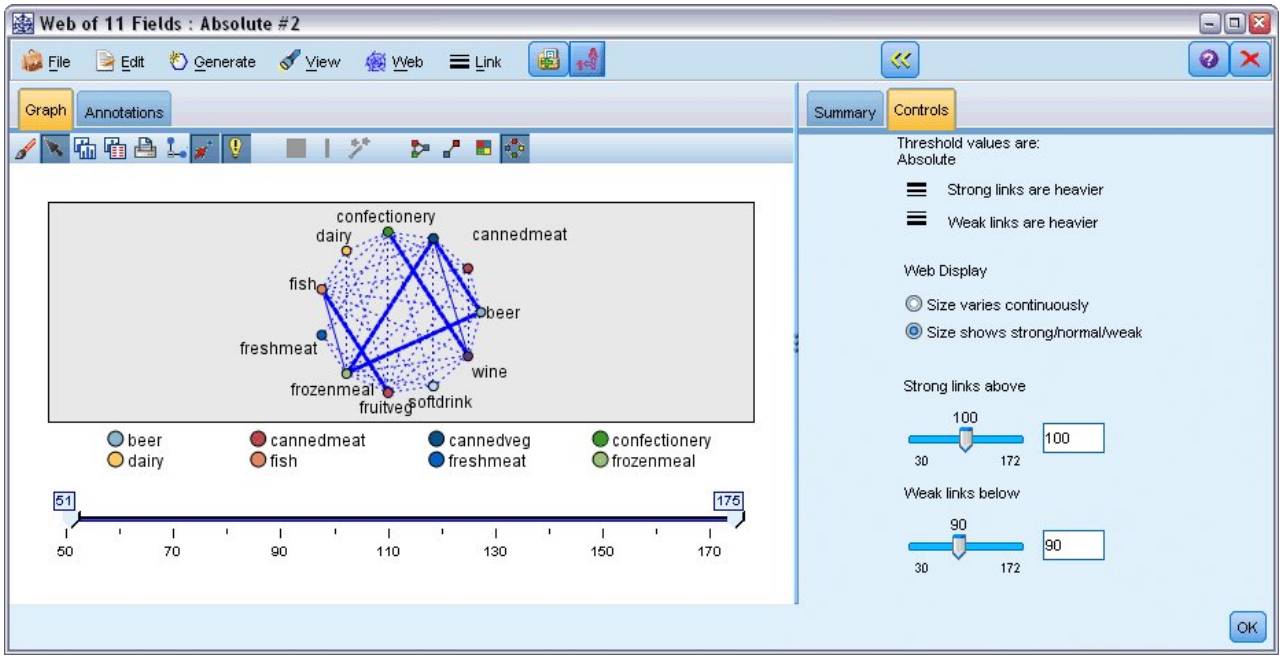
sugeruje, że wykres sieciowy (przedstawiający tylko powiązania dwukierunkowe) może podkreślić niektóre wzorce w tych danych.

Załącz węzeł sieciowy do węzła typu, edytuj węzeł sieciowy, zaznacz wszystkie pola zawartości koszyka, zaznacz opcję **Pokaż tylko flagi prawdy** i kliknij przycisk Uruchom w węźle sieciowym.



Rysunek 387. Wykres sieciowy powiązań produktów

Ponieważ większość kombinacji kategorii produktów występuje w kilku koszykach, silne łącza w tej sieci są zbyt liczne, aby wyświetlić grupy klientów sugerowane przez model.



Rysunek 388. Ograniczony węzeł sieciowy

1. Aby określić słabe i silne połączenia, kliknij żółty przycisk podwójnej strzałki na pasku narzędzi. Powoduje to rozszerzenie okna dialogowego i wyświetlenie podsumowania wykresu sieciowego i elementów sterujących.
2. Zaznacz opcję **Styl wyróżnia Silne/Normalne/Słabe łącza**.
3. Ustaw słabe łącza poniżej 90.
4. Ustaw silne łącza powyżej 100.

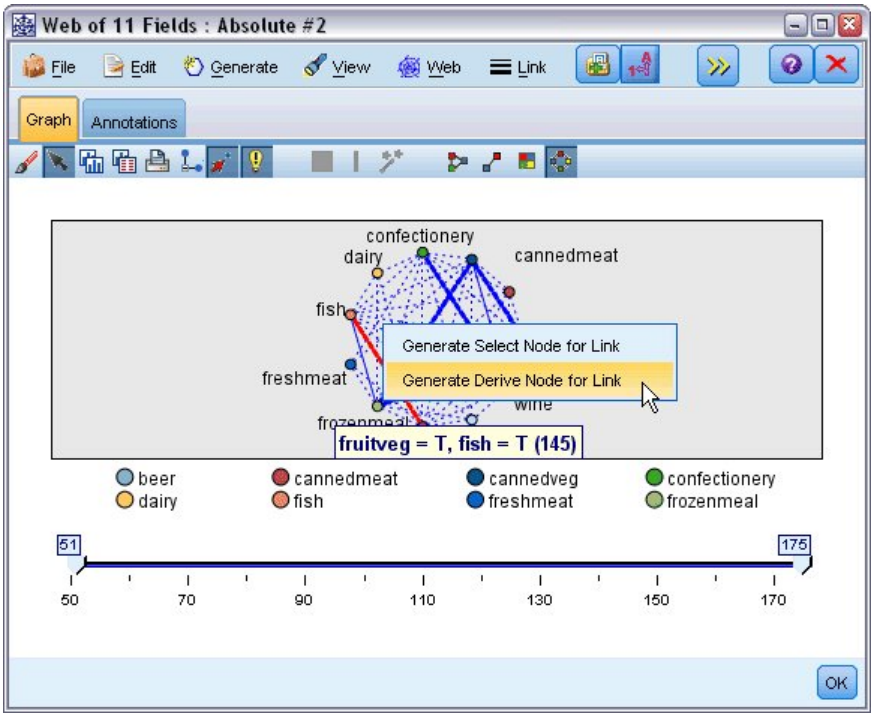
Na wynikowym wykresie wyróżniają się trzy grupy klientów:

- Osoby, które kupują rybę, owoce i warzywa, o których można powiedzieć, że stosują „zdrowe odżywianie”.
- Osoby, które kupują wino i wyroby cukiernicze.
- Osoby, które kupują piwo, mrożone potrawy i warzywa w puszce („beer, beans i pizza"”).

Profilowanie grup klientów

Zidentyfikowano trzy grupy klientów na podstawie typów produktów, które kupują, ale chcemy również wiedzieć, kim są ci klienci — tzn. poznać ich profile demograficzne. Można to osiągnąć, oznaczając każdego klienta flagą dla każdej z tych grup i używając wywodzenia reguł (C5.0) do zbudowania profili opartych na regułach dla tych flag.

Najpierw należy wyliczyć flagę dla każdej grupy. Można to automatycznie wygenerować, używając utworzonego właśnie wykresu sieciowego. Używając prawego przycisku myszy, kliknij łącze pomiędzy zmiennymi *fruitveg* i *fish*, aby podświetlić je, a następnie kliknij prawym przyciskiem myszy i wybierz opcję **Utwórz węzeł wyliczeń dla łącza.**



Rysunek 389. Wyliczanie flagi dla każdej grupy klientów

Edytuj wynikowy węzeł wyliczeń, aby zmienić nazwę zmiennej wliczania na *healthy*. Powtórz ćwiczenie dla łącza ze zmiennej *wine* do *confectionery*, nazywając wynikową zmienną wyliczania *wine_chocs*.

Dla trzeciej grupy (obejmującej trzy łącza) najpierw upewnij się, że nie są wybrane żadne łącza. Następnie zaznacz wszystkie trzy łącza zmiennych w trójkącie *cannedveg*, *beer* i *frozenmeal*, trzymając naciśnięty klawisz Shift i klikając lewy przycisk myszy. (Upewnij się, że wybrany jest tryb interaktywny, a nie tryb edycji). Następnie z menu wykresu sieciowego wybierz opcje:

Utwórz > Węzeł wyliczeń ("AND")

Zmień nazwę wynikowej zmiennej wyliczania na *beer_beans_pizza*.

Aby sprofilować te grupy klientów, połącz istniejący węzeł typu z tymi trzema węzłami wyliczania w szeregu, a następnie podłącz kolejny węzeł typu. W nowym węźle typu ustaw rolę wszystkich zmiennych na *Brak* z wyjątkiem *value*, *pmethod*, *sex*, *homeown*, *income* i *age*, które należy ustawić na *Dane wejściowe* oraz istotnej grupy klientów (na przykład *beer_beans_pizza*), którą należy ustawić na *Przewidywana*. Załącz węzeł C5.0, ustaw typ Wynik na **Zestaw reguł** i kliknij przycisk Uruchom w węźle. Wynikowy model (dla zmiennej *beer_beans_pizza*) zawiera wyraźny profil demograficzny dla tej grupy klientów:

```
Rule 1 for T:
if sex = M
and income <= 16,900
then T
```

Tą samą metodę można zastosować do innych flag grup klientów, wybierając je jako wynik w drugim węźle typu. W tym kontekście szerszy zakres alternatywnych profili można wygenerować, używając metody Apriori zamiast C5.0. Modelu Apriori można użyć do jednoczesnego profilowania wszystkich flag grup klientów, ponieważ nie jest ograniczony do zmiennej pojedynczego wyniku.

Podsumowanie

Ten przykład pokazuje, jak można użyć programu IBM SPSS Modeler do wykrywania powiązań lub łączy w bazie danych, używając modelowania (metoda Apriori) lub wizualizacji (wykres sieciowy). Takie łącza odpowiadają grupom obserwacji w danych, a takie grupy można szczegółowo zbadać i sprofilować za pomocą modelowania (używając zestawu reguł C5.0).

W sektorze handlu detalicznego takiego grupowania można by na przykład użyć do kierowania specjalnych ofert, aby poprawić współczynniki odpowiedzi na przesyłki bezpośrednie lub aby dostosować zakres produktów zamówiony przez oddział w celu dopasowania do popytu bazy demograficznej.

Rozdział 28. Ocena ofert nowych pojazdów (KNN)

Analiza najbliższego sąsiedztwa jest metodą klasyfikacji obserwacji na podstawie ich podobieństwa do innych obserwacji. Zostało to opracowane w nauczaniu maszynowym jako sposób rozpoznawania wzorców danych bez konieczności zapewnienia dokładnej zgodności z jakimikolwiek zapamiętanymi wzorcami lub obserwacjami. Podobne obserwacje znajdują się blisko siebie, a niepodobne — daleko. Zatem odległość między dwoma obserwacjami stanowi miarę ich niepodobieństwa.

Obserwacje znajdujące się blisko siebie nazywają się „sąsiedztwem". Podczas prezentacji nowej (wstrzymanej) obserwacji obliczana jest odległość od każdej obserwacji modelu. Zostaje określona klasyfikacja najbardziej podobnych obserwacji najbliższego sąsiedztwa, a nowa obserwacja zostaje umieszczona w kategorii, która zawiera największą liczbę obserwacji najbliższego sąsiedztwa.

Można określić liczbę najbliższych elementów sąsiednich do analizowania; ta wartość to *k*. Rysunki przedstawiają, jak nowa obserwacja będzie sklasyfikowana za pomocą dwóch różnych wartości *k*. Jeśli *k* = 5, nowa obserwacja jest umieszczana w kategorii *1*, ponieważ większość najbliższych elementów sąsiednich należy do kategorii *1*. Jeśli jednak *k* = 9, nowa obserwacja jest umieszczana w kategorii *0*, ponieważ większość najbliższych elementów sąsiednich należy do kategorii *0*.

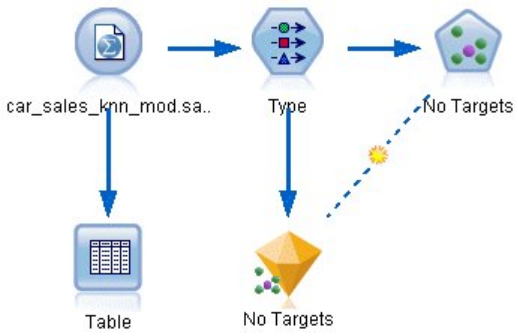
Analiza najbliższego sąsiedztwa może być również użyta do obliczania docelowych wartości ilościowych. W tej sytuacji do uzyskania przewidywanej wartości dla nowej obserwacji stosowana jest docelowa wartość średniej lub mediany najbliższych sąsiadów.

Producent motoryzacyjny opracował prototypy dwóch nowych pojazdów — samochodu osobowego i ciężarowego. Przed wprowadzeniem nowych modeli do oferty producent chce określić, które istniejące pojazdy na rynku są najbardziej podobne do prototypów, tzn. które pojazdy są „najbliższymi sąsiadami" i z którymi modelami będą konkurować.

Producent zgromadził dane o istniejących modelach w wielu kategoriach i dodał szczegółowe dane swoich prototypów. Kategorie, w których modele będą porównywane obejmują cenę w tysiącach (*price*), wielkość silnika (*engine_s*), moc (*horsepow*), rozstaw osi (*wheelbas*), szerokość (*width*), długość (*length*), ciężar własny (*curb_wgt*), pojemność zbiornika paliwa (*fuel_cap*) oraz spalanie paliwa (*mpg*).

W tym przykładzie zastosowano strumień o nazwie *car_sales_knn.str*, który jest dostępny w folderze *Demos*, podfolder *streams*. Plik danych to *car_sales_knn_mod.sav*. Więcej informacji można znaleźć w temacie "Folder Demos" na stronie 5.

Tworzenie strumienia

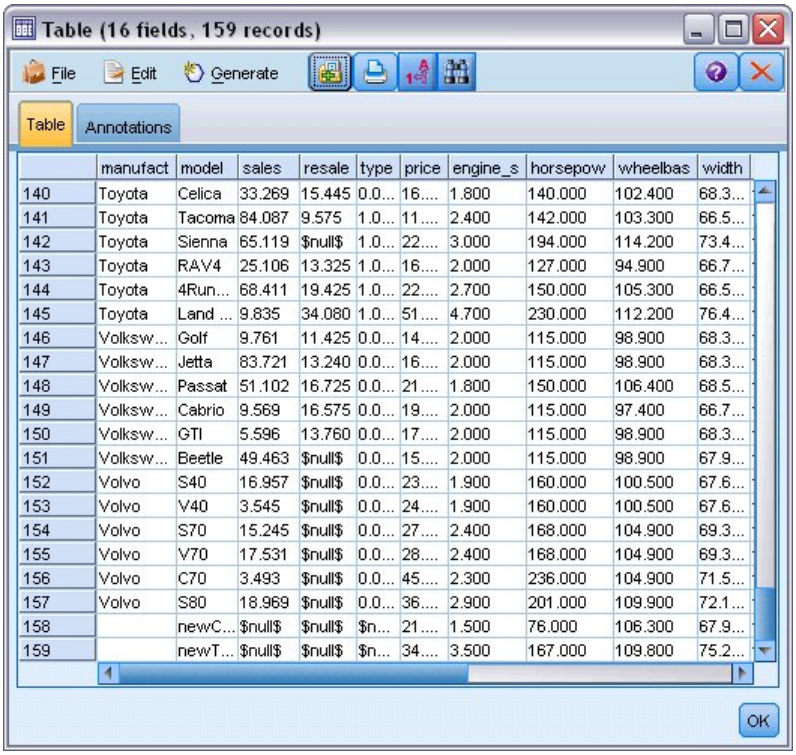


Rysunek 390. Przykładowy strumień dla modelowania KNN

Utwórz nowy strumień i dodaj węzeł źródłowy Plik Statistics wskazujący na plik *car_sales_knn_mod.sav* w folderze *Demos* instalacji programu IBM SPSS Modeler.

Najpierw zobaczmy, jakie dane zgromadził producent.

1. Załącz węzeł tabeli do węzła źródłowego Plik Statistics.
2. Otwórz węzeł Tabela i kliknij przycisk **Uruchom**.



	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...

Rysunek 391. Dane źródłowe dla samochodów osobowych i ciężarowych

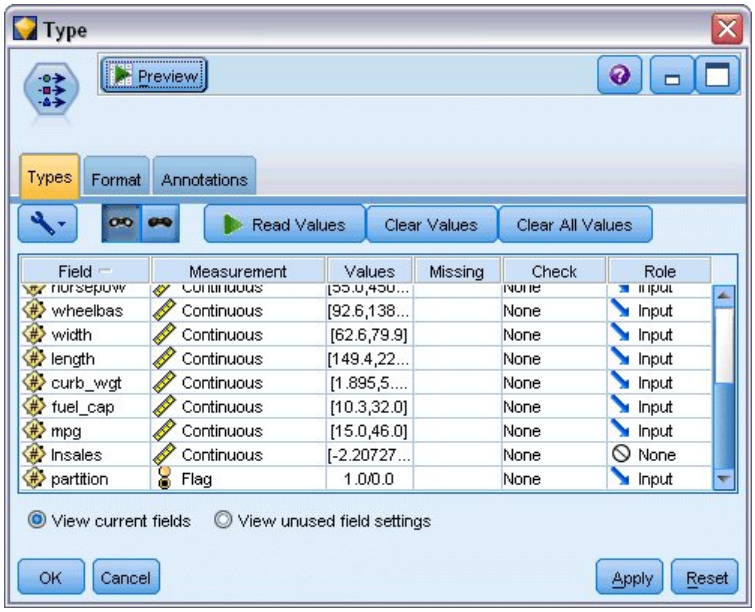
Na końcu pliku zostały dodane dane dwóch prototypów o nazwach *newCar* i *newTruck*.

W danych źródłowych widać, że producent używa klasyfikacji „samochód ciężarowy” (wartość 1 w kolumnie *type*) raczej swobodnie w znaczeniu pojazdu niebędącego samochodem osobowym.

Ostatnia kolumna *partition* jest niezbędna, aby dwa prototypy mogły być wyznaczone jako wstrzymania, gdy zidentyfikowane zostanie najbliższe sąsiedztwo. W ten sposób ich dane nie będą wpływać na obliczenia,

ponieważ chcemy przeanalizować resztę rynku. Ustawienie wartości *partition* dwóch rekordów wstrzymania na 1, podczas gdy wszystkie inne rekordy mają wartość 0 w tej zmiennej, pozwala na użycie tej zmiennej później, gdy będziemy ustawiać obserwacje centralne — rekordy, dla których chcemy obliczyć najbliższe sąsiedztwo.

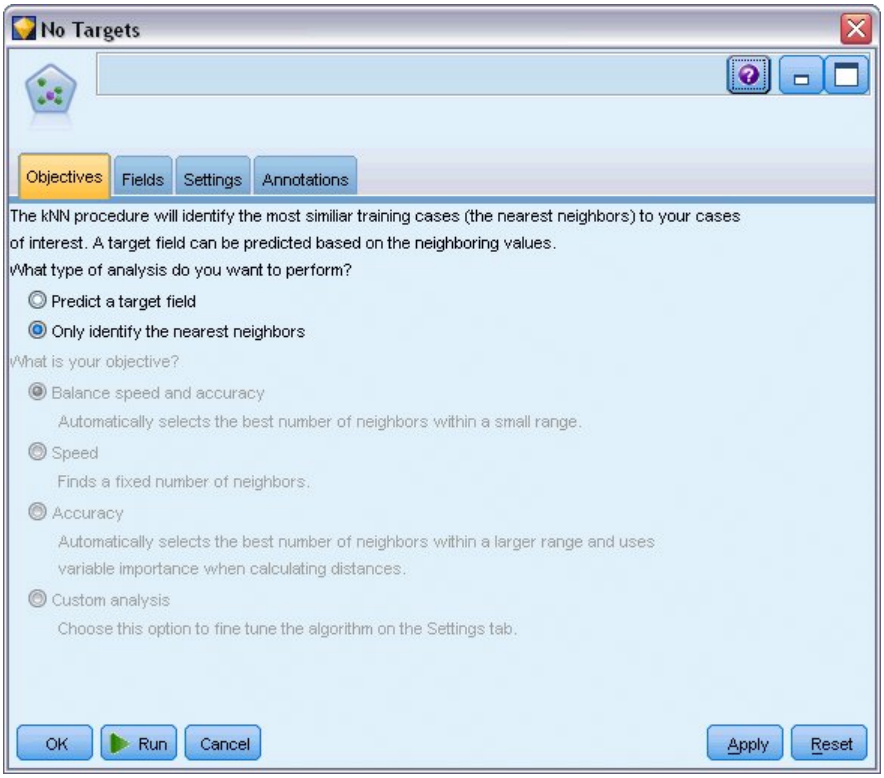
Na razie pozostaw otwarte okno wyników tabeli, ponieważ będziemy później z niego korzystać.



Rysunek 392. Ustawienia węzła typu

3. Do strumienia dodaj węzeł typu.
4. Załącz węzeł typu do węzła źródłowego Plik Statistics.
5. Otwórz węzeł typu.

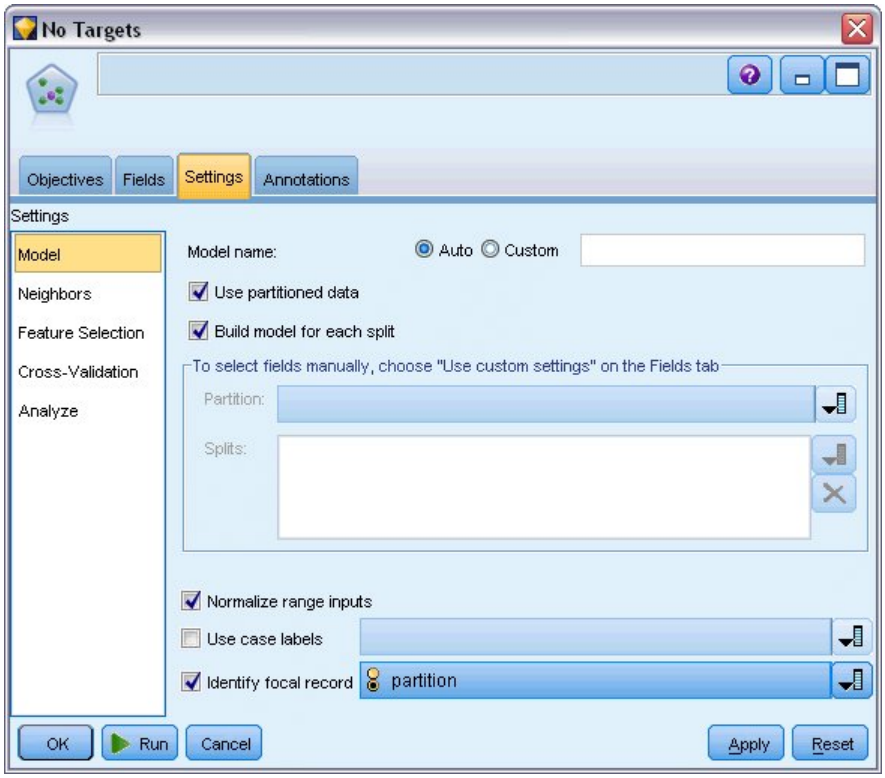
   Chcemy porównać tylko zmienne od *price* do *mpg*, pozostawimy więc role dla wszystkich tych zmiennych ustawione na **Dane wejściowe**.
6. Ustaw role dla wszystkich pozostałych zmiennych (od *manufact* do *type* oraz *lnsales*) na **Brak**.
7. Poziom pomiaru dla ostatniej zmiennej, *partition*, ustaw na **Flaga**. Upewnij się, że jej rola jest ustawiona na **Dane wejściowe**.
8. Kliknij przycisk **Odczytaj wartości**, aby wczytać wartości danych do strumienia.
9. Kliknij przycisk **OK**.



Rysunek 393. Wybór identyfikacji najbliższego sąsiedztwa

10. Załącz węzeł KNN do węzła typu.
11. Otwórz węzeł KNN.

    Tym razem nie będziemy przewidywać zmiennej przewidywanej, ponieważ chcemy tylko znaleźć najbliższe sąsiedztwo dla dwóch prototypów.
12. Na karcie **Cele** zaznacz opcję **Tylko zidentyfikuj najbliższych sąsiadów**.
13. Kliknij kartę **Ustawienia**.



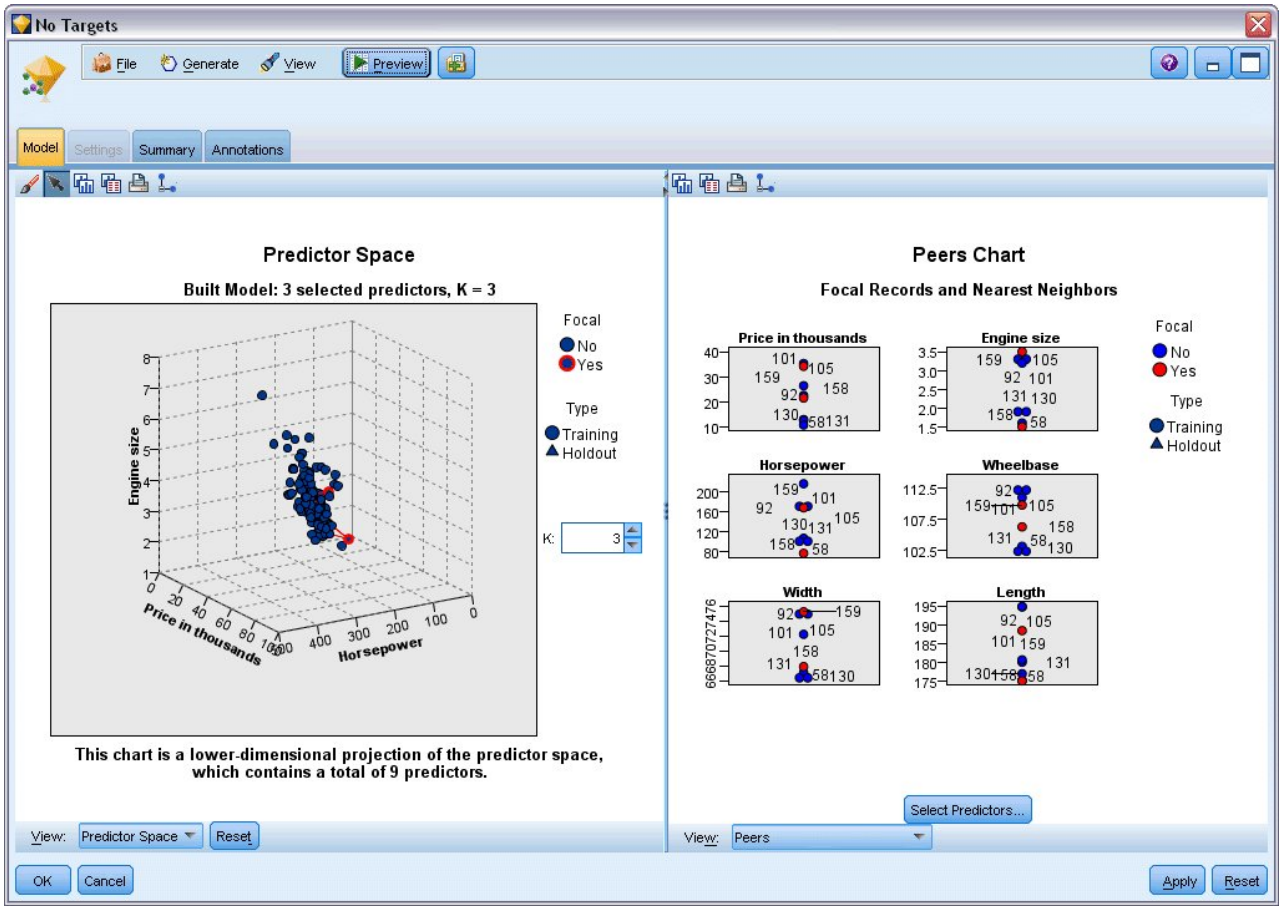
Rysunek 394. Użycie zmiennej partition do identyfikacji obserwacji centralnych

Teraz możemy użyć zmiennej *partition* do identyfikacji obserwacji centralnych — rekordów, dla których chcemy zidentyfikować najbliższe sąsiedztwo. Używając zmiennej flagi, zapewniamy, że rekordy, dla których ustawimy wartość 1 staną się naszymi obserwacjami centralnymi.

Jak widzieliśmy, jedyne rekordy, które mają wartość 1 w tej zmiennej to *newCar* i *newTruck*, więc będą to nasze obserwacje centralne.

14. Na panelu **Model** karty **Ustawienia** zaznacz pole wyboru **Wskaż obserwację centralną**.
15. Z listy rozwijanej dla tej zmiennej wybierz **partition**.
16. Kliknij przycisk **Uruchom**.

Badanie wyników



Rysunek 395. Okno Przegląd modelu

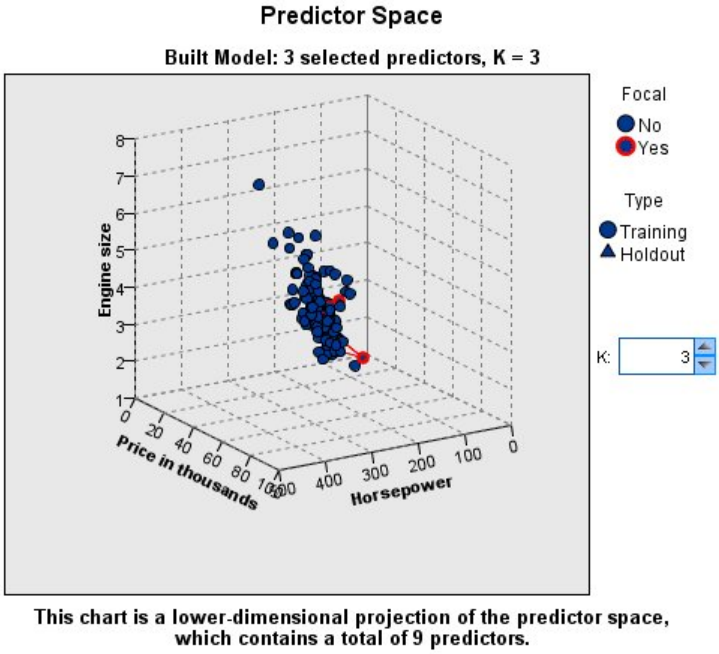
Model użytkowy został utworzony w obszarze roboczym strumienia i na palecie modeli. Otwórz jeden z modeli użytkowych, aby zobaczyć Przegląd modelu, który zawiera okno składające się z dwóch paneli:

- Pierwszy panel wyświetla przegląd modelu nazywany widokiem głównym. Główny widok dla modelu Najbliższe sąsiedztwo jest znany jako **przestrzeń predyktorów**.
- Drugi panel wyświetla jeden z dwóch rodzajów widoków:

  Pomocniczy widok modelu przedstawia więcej informacji o modelu, ale nie koncentruje się na samym modelu.

  Połączony widok jest widokiem przedstawiającym szczegółowe informacje o modelu, gdy użytkownik rozwinie część widoku głównego.

Przestrzeń predyktorów



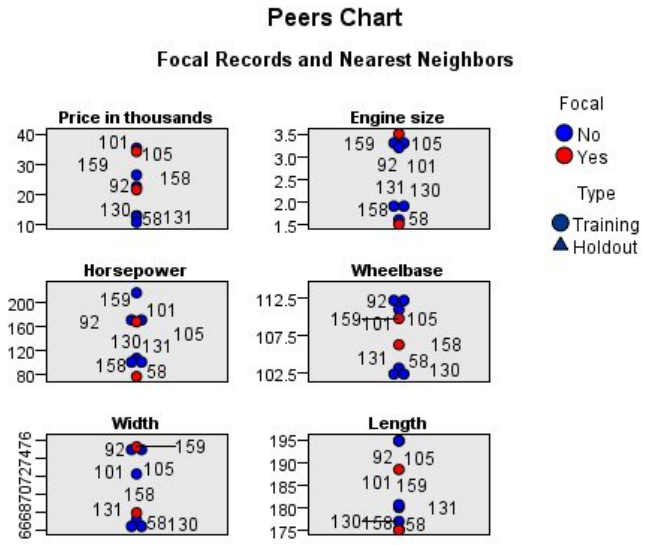
Rysunek 396. Wykres przestrzeni predyktorów

Wykres przestrzeni predyktorów jest interaktywnym wykresem 3-W, który tworzy wykres punktów danych dla trzech właściwości (trzech pierwszych zmiennych wejściowych w danych wejściowych) reprezentujących cenę, wielkość silnika i moc.

Dwie obserwacje centralne są zaznaczone na czerwono z liniami łączącymi je z *k* najbliższymi sąsiadami.

Klikając i przeciągając wykres, można go obracać, aby uzyskać lepszy widok rozkładu punktów w przestrzeni predyktorów. Kliknij przycisk **Resetuj**, aby powrócić do widoku domyślnego.

Wykres elementów równorzędnych



Rysunek 397. Wykres elementów równorzędnych

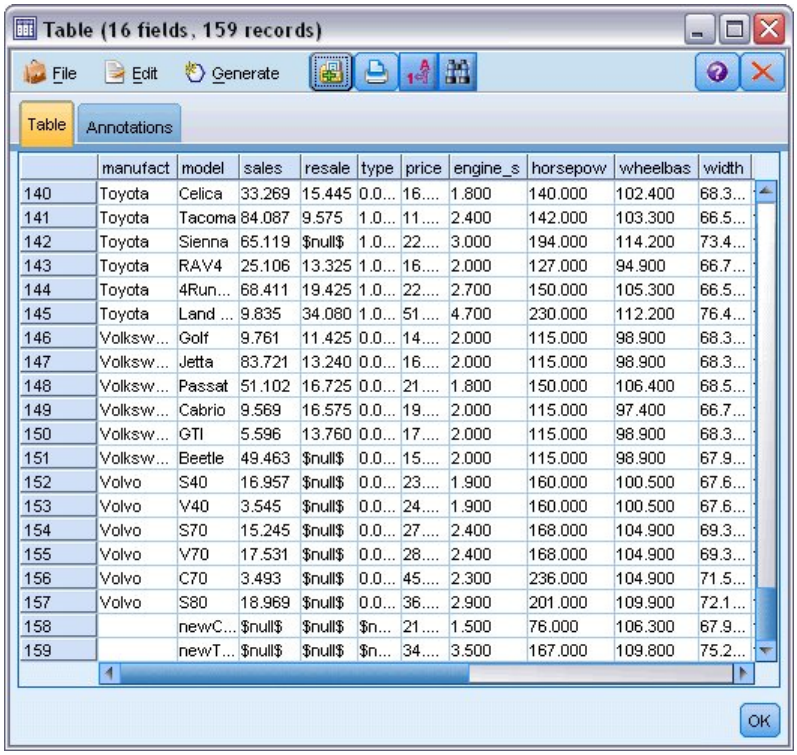
Domyślnym widokiem dodatkowym jest wykres elementów równorzędnych, który wyróżnia dwie obserwacje centralne wybrane w przestrzeni predyktorów oraz ich *k* najbliższych sąsiadów dla każdej z sześciu właściwości — pierwszych sześciu zmiennych wejściowych danych źródłowych.

Pojazdy są reprezentowane przez numery rekordów w danych źródłowych. W tym miejscu potrzebujemy wyników z węzła tabeli, aby zidentyfikować pojazdy.

Jeśli wyniki węzła tabeli są wciąż dostępne:

1. Kliknij kartę **Wyniki** okna menedżera w prawym górnym rogu głównego okna programu IBM SPSS Modeler.
2. Dwukrotnie kliknij wpis **Tabela (16 zmiennych, 159 rekordów)**.

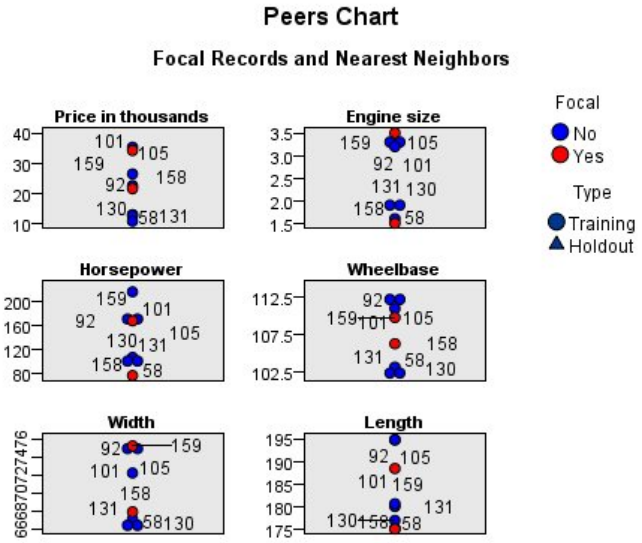
   Jeśli wyniki tabeli nie są już dostępne:
3. W głównym oknie IBM SPSS Modeler otwórz węzeł tabeli.
4. Kliknij przycisk **Uruchom**.



	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...

Rysunek 398. Identyfikowanie rekordów na podstawie numerów rekordów

Przewijając w dół do końca tabeli, możemy zobaczyć, że *newCar* i *newTruck* to dwa ostatnie rekordy danych o numerach 158 i 159.



Rysunek 399. Porównywanie właściwości na wykresie elementów równorzędnych

Na tej podstawie możemy zobaczyć na wykresie elementów równorzędnych na przykład, że samochód ciężarowy *newTruck* (159)ma większy silnik niż jakikolwiek element z najbliższego sąsiedztwa, a samochód *newCar* (158) ma mniejszy silnik niż jakikolwiek inny element *jego* najbliższego sąsiedztwa.

Dla każdej z sześciu właściwości można przemieszczać mysz nad poszczególnymi kropkami, aby zobaczyć rzeczywistą wartość każdej właściwości dla określonej obserwacji.

Ale które pojazdy są najbliższymi sąsiadami samochodów *newCar* i *newTruck*?

Wykres elementów równorzędnych jest odrobinę zatłoczony, przejdźmy więc do prostszego widoku.

5. Kliknij listę rozwijaną **Widok** w dolnej części wykresu elementów równorzędnych (wpis, który obecnie pokazuje wartość **Elementy o zbliżonych wartościach**).
6. Wybierz pozycję **Tabela sąsiadów i odległości**.

Tabela sąsiadów i odległości

k Nearest Neighbors and Distances

Displayed for Initial Focal Records

Focal Record	Nearest Neighbors			Nearest Distar	
	1	2	3	1	2
158	131	130	58	0.979	0.990
159	105	92	101	0.580	0.634

Rysunek 400. Tabela sąsiadów i odległości

Taki widok jest lepszy. Teraz widzimy trzy modele, do których każdy z dwóch prototypów ma najbliżej na rynku.

Dla samochodu *newCar* (obserwacja centralna 158) są to: Saturn SC (131), Saturn SL (130) i Honda Civic (58).

Nie ma tu wielkich niespodzianek — wszystkie trzy samochody średniej wielkości klasy sedan, więc samochód *newCar* powinien dobrze pasować, zwłaszcza biorąc pod uwagę doskonałe spalanie paliwa.

Dla samochodu ciężarowego *newTruck* (obserwacja centralna 159) najbliższymi sąsiadami są: Nissan Quest (105), Mercury Villager (92) i Mercedes M-Class (101).

Jak widzieliśmy wcześniej, nie są to koniecznie samochody ciężarowe w tradycyjnym rozumieniu, ale pojazdy, które nie są zaklasyfikowane jako samochody osobowe. Patrząc na wyniki węzła tabeli dla najbliższego sąsiedztwa, widzimy, że pojazd *newTruck* jest relatywnie drogi oraz jest jednym z najcięższych pojazdów tego typu. Spalanie paliwa jest jednak znów lepsze niż u najbliższej konkurencji, więc należy uznać to za zaletę.

Podsumowanie

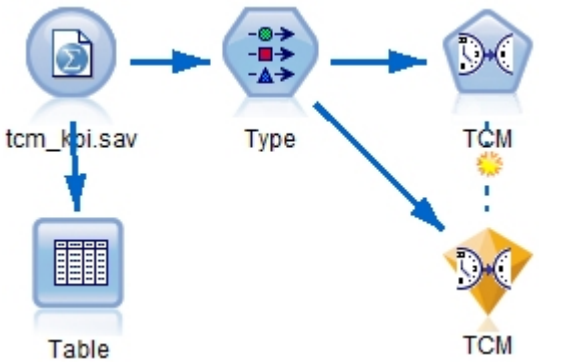
Zobaczyliśmy, jak można użyć analizy najbliższego sąsiedztwa do porównania szerokiego zestawu właściwości w obserwacjach z określonego zbioru danych. Obliczyliśmy również, dla dwóch bardzo różnych rekordów wstrzymania, obserwacje, które najbardziej przypominają te wstrzymania.

Rozdział 29. Odkrywanie relacji przyczynowych w metrykach biznesowych (TCM)

Firma śledzi różne kluczowe wskaźniki wydajności, które opisują stan finansowy spółki w czasie. Śledzi też różne metryki, które może kontrolować. Firma jest zainteresowana użyciem modeli przyczynowych szeregów czasowych do odkrycia relacji przyczynowych pomiędzy kontrolowanymi metrykami i kluczowymi wskaźnikami wydajności. Chciałaby również wiedzieć o wszelkich relacjach przyczynowych pomiędzy kluczowymi wskaźnikami wydajności.

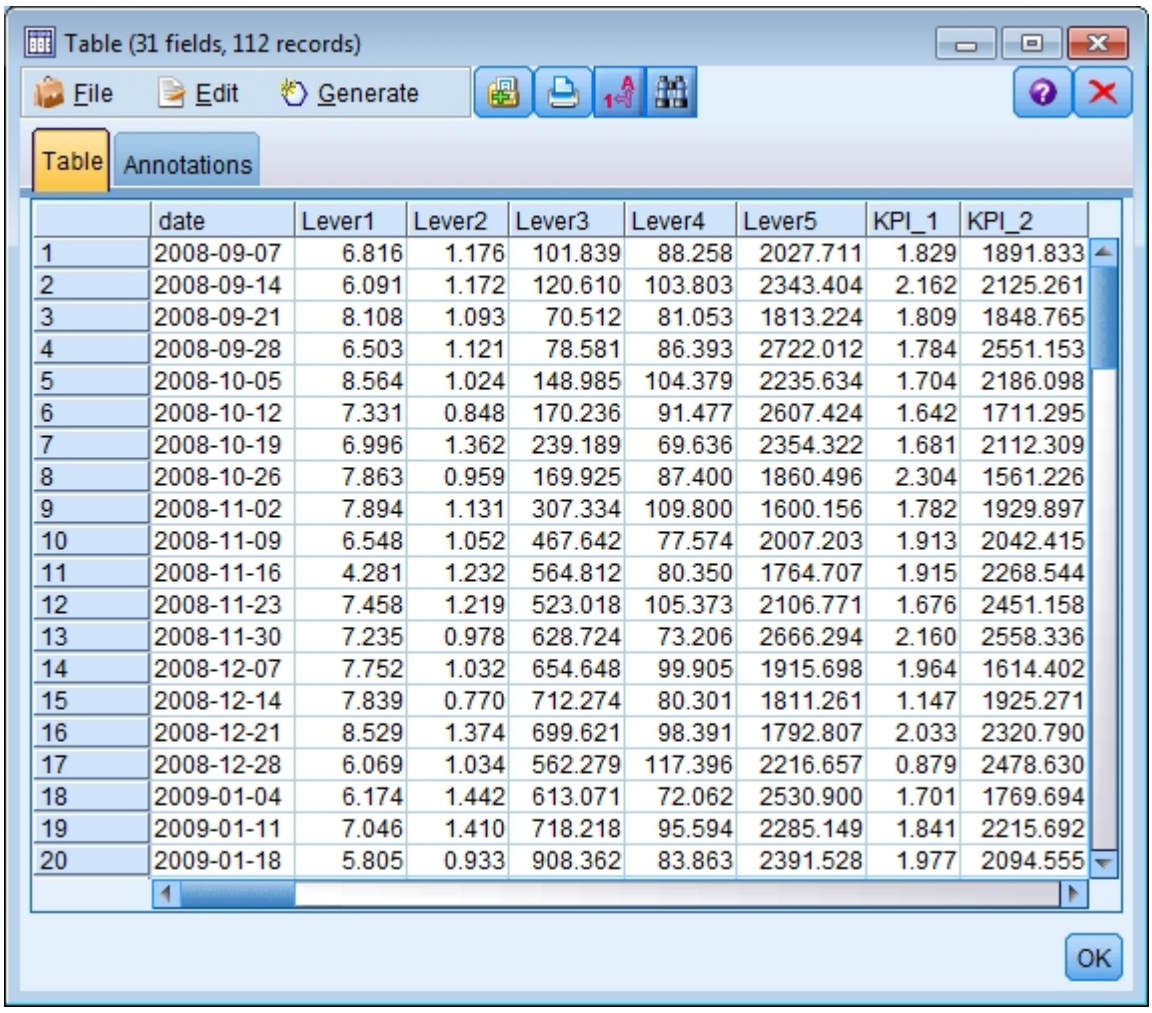
Plik danych tcm_kpi.sav zawiera tygodniowe dane dotyczące kluczowych wskaźników wydajności i kontrolowanych metryk. Dane dla kluczowych wskaźników wydajności są zapisane w zmiennych rozpoczynających się przedrostkiem *KPI*. Dane dla kontrolowanych metryk są zapisane w zmiennych rozpoczynających się przedrostkiem *Lever*.

Tworzenie strumienia



Rysunek 401. Przykładowy strumień dla modelowania TCM

1. Utwórz nowy strumień i dodaj węzeł źródłowy Plik Statistics wskazujący na plik *tcm_kpi.sav* w folderze *Demos* instalacji programu IBM SPSS Modeler .
2. Załącz węzeł tabeli do węzła źródłowego Plik Statistics.
3. Otwórz węzeł tabeli i kliknij przycisk **Uruchom**, aby przyjrzeć się danym. Są to tygodniowe dane kluczowych wskaźników wydajności i kontrolowanych metryk. Dane dla kluczowych wskaźników wydajności są zapisane w zmiennych z przedrostkiem *KPI*, a dane dla kontrolowanych metryk są zapisane w zmiennych z przedrostkiem *Lever*.



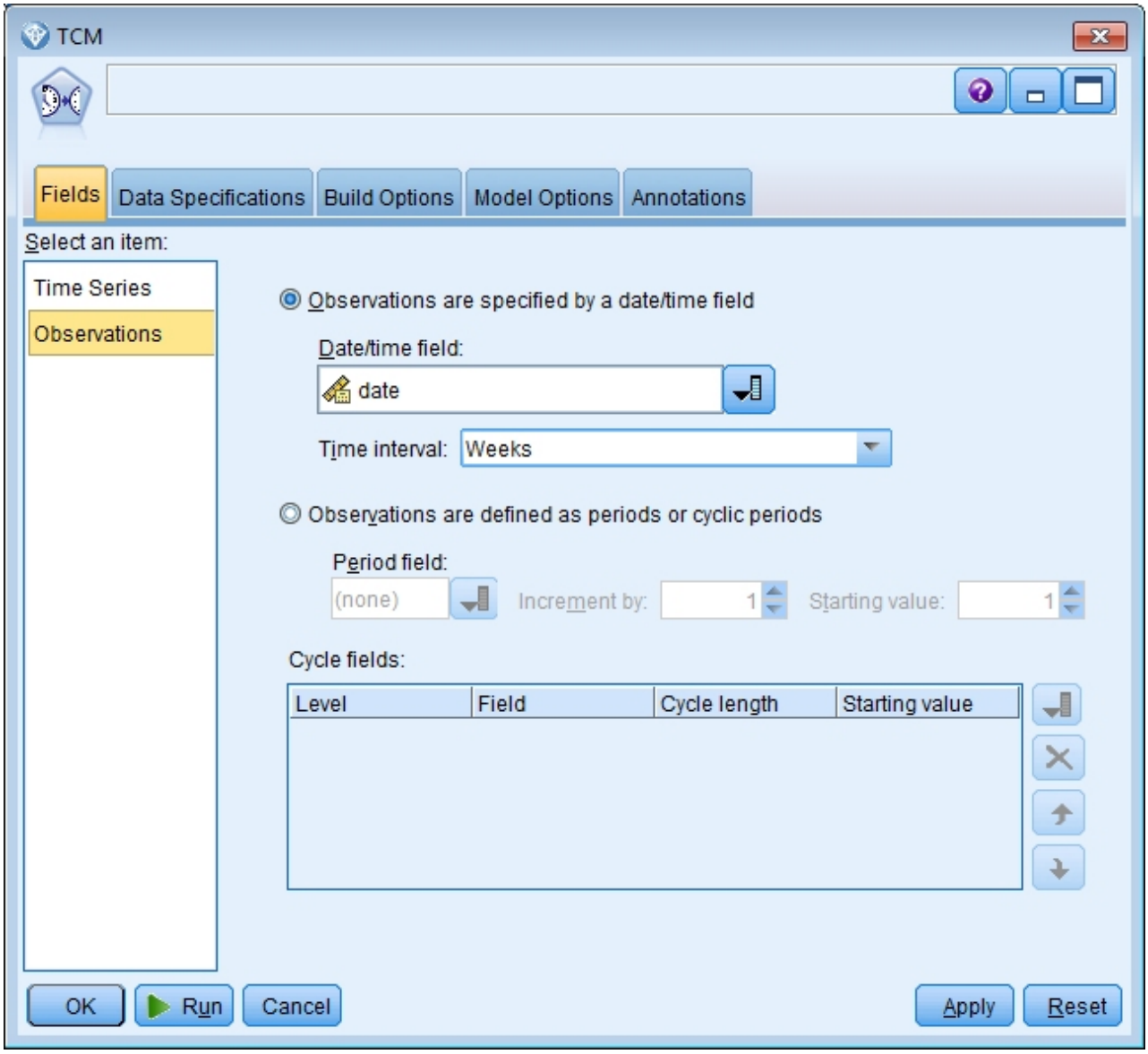
	date	Lever1	Lever2	Lever3	Lever4	Lever5	KPI_1	KPI_2
1	2008-09-07	6.816	1.176	101.839	88.258	2027.711	1.829	1891.833
2	2008-09-14	6.091	1.172	120.610	103.803	2343.404	2.162	2125.261
3	2008-09-21	8.108	1.093	70.512	81.053	1813.224	1.809	1848.765
4	2008-09-28	6.503	1.121	78.581	86.393	2722.012	1.784	2551.153
5	2008-10-05	8.564	1.024	148.985	104.379	2235.634	1.704	2186.098
6	2008-10-12	7.331	0.848	170.236	91.477	2607.424	1.642	1711.295
7	2008-10-19	6.996	1.362	239.189	69.636	2354.322	1.681	2112.309
8	2008-10-26	7.863	0.959	169.925	87.400	1860.496	2.304	1561.226
9	2008-11-02	7.894	1.131	307.334	109.800	1600.156	1.782	1929.897
10	2008-11-09	6.548	1.052	467.642	77.574	2007.203	1.913	2042.415
11	2008-11-16	4.281	1.232	564.812	80.350	1764.707	1.915	2268.544
12	2008-11-23	7.458	1.219	523.018	105.373	2106.771	1.676	2451.158
13	2008-11-30	7.235	0.978	628.724	73.206	2666.294	2.160	2558.336
14	2008-12-07	7.752	1.032	654.648	99.905	1915.698	1.964	1614.402
15	2008-12-14	7.839	0.770	712.274	80.301	1811.261	1.147	1925.271
16	2008-12-21	8.529	1.374	699.621	98.391	1792.807	2.033	2320.790
17	2008-12-28	6.069	1.034	562.279	117.396	2216.657	0.879	2478.630
18	2009-01-04	6.174	1.442	613.071	72.062	2530.900	1.701	1769.694
19	2009-01-11	7.046	1.410	718.218	95.594	2285.149	1.841	2215.692
20	2009-01-18	5.805	0.933	908.362	83.863	2391.528	1.977	2094.555

Rysunek 402. Dane źródłowe dla kluczowych wskaźników wydajności i kontrolowanych metryk

4. Do strumienia dodaj węzeł typu.
5. Załącz węzeł typu do węzła źródłowego Plik Statistics.

Uruchamianie analizy

1. Załącz węzeł TCM do węzła typu, a następnie otwórz węzeł TCM i przejdź do sekcji **Obserwacje** na karcie **Zmienne**.



Rysunek 403. Modelowanie przyczynowe szeregów czasowych, obserwacje

2. Wybierz zmienną *date* w polu Zmienna typu data/czas i wybierz *Tygodnie* w polu Przedział czasowy.
3. Kliknij pozycję **Szereg czasowy** i zaznacz opcję **Użyj wstępnie zdefiniowanych ról**.

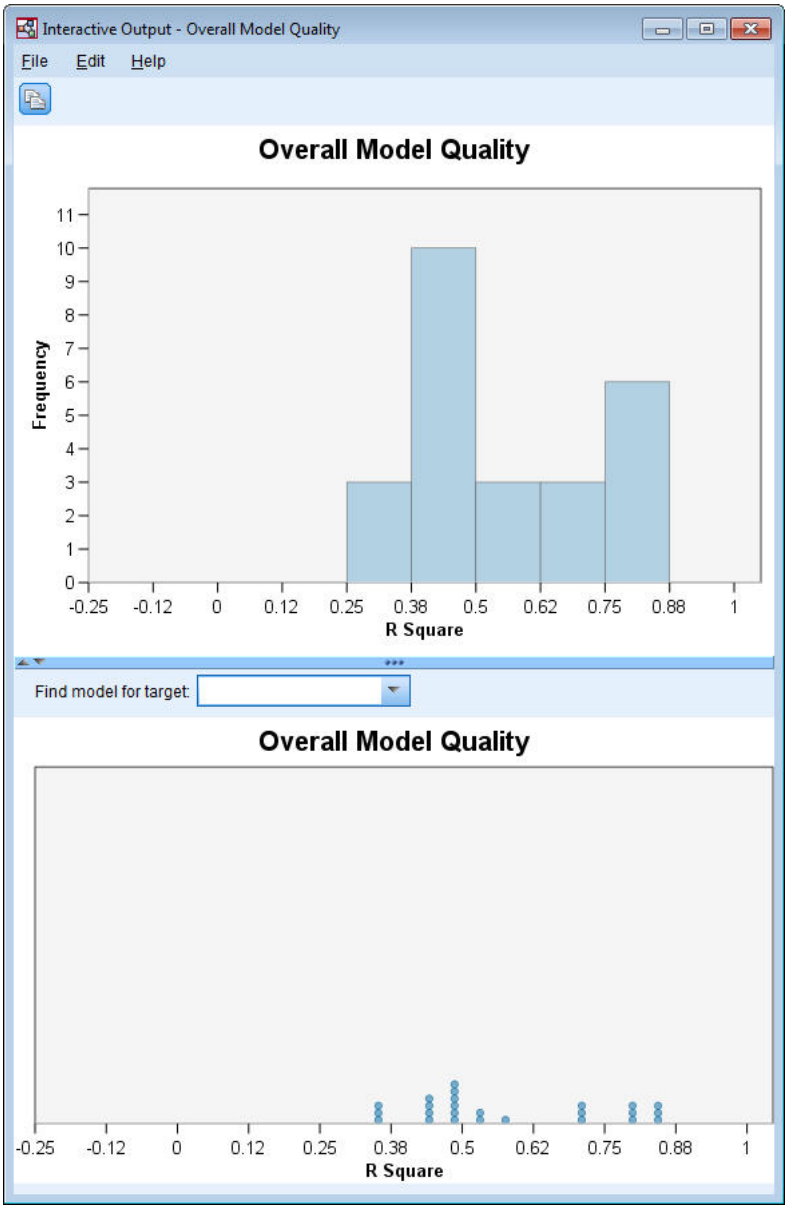
   W przykładowym zbiorze danych tcm_kpi.sav zmienne od *Lever1* do *Lever5* mają role Dane wejściowe, a zmienne od *KPI_1* do *KPI_25* mają role Łącznie. Kiedy zaznaczona jest opcja **Użyj wstępnie zdefiniowanych ról**, zmienne z rolą Dane wejściowe są traktowane jako potencjalne zmienne wejściowe, a zmienne z rolą Łącznie są traktowane jako potencjalne zmienne wejściowe i zmienne przewidywane dla modeli przyczynowych szeregów czasowych.

   Procedura modeli przyczynowych szeregów czasowych określa najlepsze dane wejściowe dla każdej zmiennej przewidywanej na podstawie zbioru potencjalnych zmiennych wejściowych. W tym przykładzie potencjalne zmienne wejściowe to zmienne od *Lever1* do *Lever5* i zmienne od *KPI_1* do *KPI_25*.
4. Kliknij przycisk **Uruchom**.

Wykres Ogólna jakość modelu

Pozycja wyników Ogólna jakość modelu, która jest generowana domyślnie, wyświetla wykres słupkowy oraz powiązany wykres punktowy dopasowania modelu dla wszystkich modeli. Istnieje osobny model dla każdego szeregu przewidywanego. Dopasowanie modelu jest mierzone według wybranej statystyki dopasowania. Ten przykład używa domyślnej statystyki dopasowania, którą jest R-kwadrat.

Pozycja Ogólna jakość modelu zawiera interaktywne funkcje. Aby włączyć te funkcje, aktywuj pozycję, klikając dwukrotnie wykres Ogólna jakość modelu w przeglądarce.



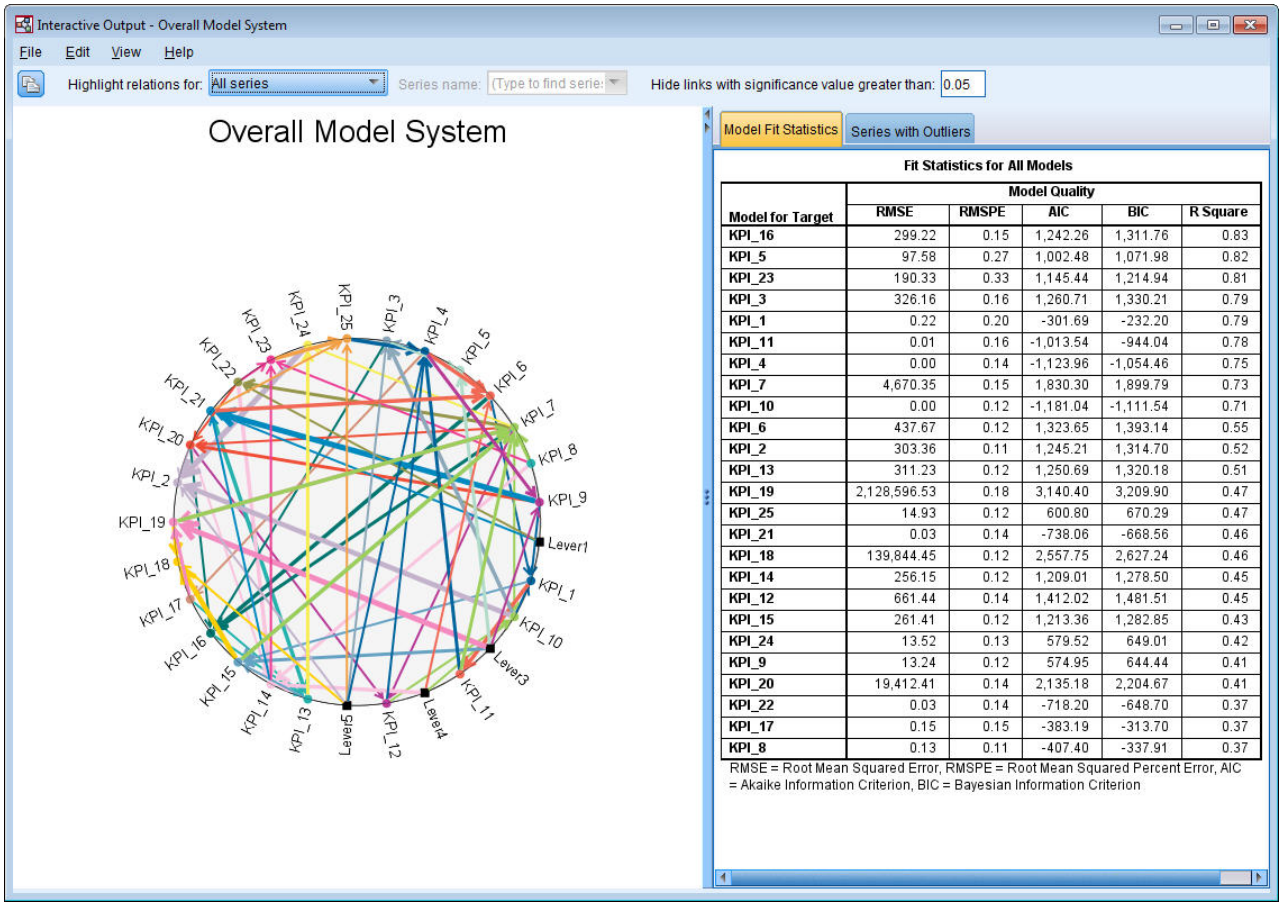
Rysunek 404. Ogólna jakość modelu

Kliknięcie słupka na wykresie słupkowym filtruje wykres punktowy tak, aby wyświetlane były wyłącznie modele powiązane z wybranym słupkiem. Umieszczenie kursora nad punktem na wykresie punktowym wyświetla podpowiedź zawierającą nazwę powiązanego szeregu oraz wartość statystyki dopasowania. Model dla określonego szeregu docelowego na wykresie punktowym można znaleźć, określając nazwę szeregu w polu **Znajdź model dla zmiennej przewidywanej**.

Ogólny system modelu

Pozycja wyników Ogólny system modelu, która jest generowana domyślnie, wyświetla graficzną reprezentację relacji przyczynowych pomiędzy szeregami w systemie modelu. Domyślnie wyświetlane są relacje dla 10 najlepszych modeli, określonych na podstawie statystyki dopasowania R-kwadrat. Liczba najlepszych modeli (określanych również jako najlepiej dopasowane modele) oraz statystyka dopasowania są określane w ustawieniach Prezentacja szeregów (na karcie Opcje budowania) okna dialogowego Modelowanie przyczynowe szeregów czasowych.

Pozycja Ogólny system modelu zawiera interaktywne funkcje. Aby włączyć te funkcje, aktywuj pozycję, klikając dwukrotnie wykres Ogólny system modelu w przeglądarce. W tym przykładzie najważniejsze jest dostrzeżenie relacji pomiędzy wszystkimi szeregami w systemie. W wynikach interaktywnych wybierz pozycję **Wszystkie szeregi** z listy rozwijanej **Podświetl relacje dla**.



Model for Target	Model Quality				
	RMSE	RMSPE	AIC	BIC	R Square
KPI_16	299.22	0.15	1,242.26	1,311.76	0.83
KPI_5	97.58	0.27	1,002.48	1,071.98	0.82
KPI_23	190.33	0.33	1,145.44	1,214.94	0.81
KPI_3	326.16	0.16	1,260.71	1,330.21	0.79
KPI_1	0.22	0.20	-301.69	-232.20	0.79
KPI_11	0.01	0.16	-1,013.54	-944.04	0.78
KPI_4	0.00	0.14	-1,123.96	-1,054.46	0.75
KPI_7	4,670.35	0.15	1,830.30	1,899.79	0.73
KPI_10	0.00	0.12	-1,181.04	-1,111.54	0.71
KPI_6	437.67	0.12	1,323.65	1,393.14	0.55
KPI_2	303.36	0.11	1,245.21	1,314.70	0.52
KPI_13	311.23	0.12	1,250.69	1,320.18	0.51
KPI_19	2,128,596.53	0.18	3,140.40	3,209.90	0.47
KPI_25	14.93	0.12	600.80	670.29	0.47
KPI_21	0.03	0.14	-738.06	-668.56	0.46
KPI_18	139,844.45	0.12	2,557.75	2,627.24	0.46
KPI_14	256.15	0.12	1,209.01	1,278.50	0.45
KPI_12	661.44	0.14	1,412.02	1,481.51	0.45
KPI_15	261.41	0.12	1,213.36	1,282.85	0.43
KPI_24	13.52	0.13	579.52	649.01	0.42
KPI_9	13.24	0.12	574.95	644.44	0.41
KPI_20	19,412.41	0.14	2,135.18	2,204.67	0.41
KPI_22	0.03	0.14	-718.20	-648.70	0.37
KPI_17	0.15	0.15	-383.19	-313.70	0.37
KPI_8	0.13	0.11	-407.40	-337.91	0.37

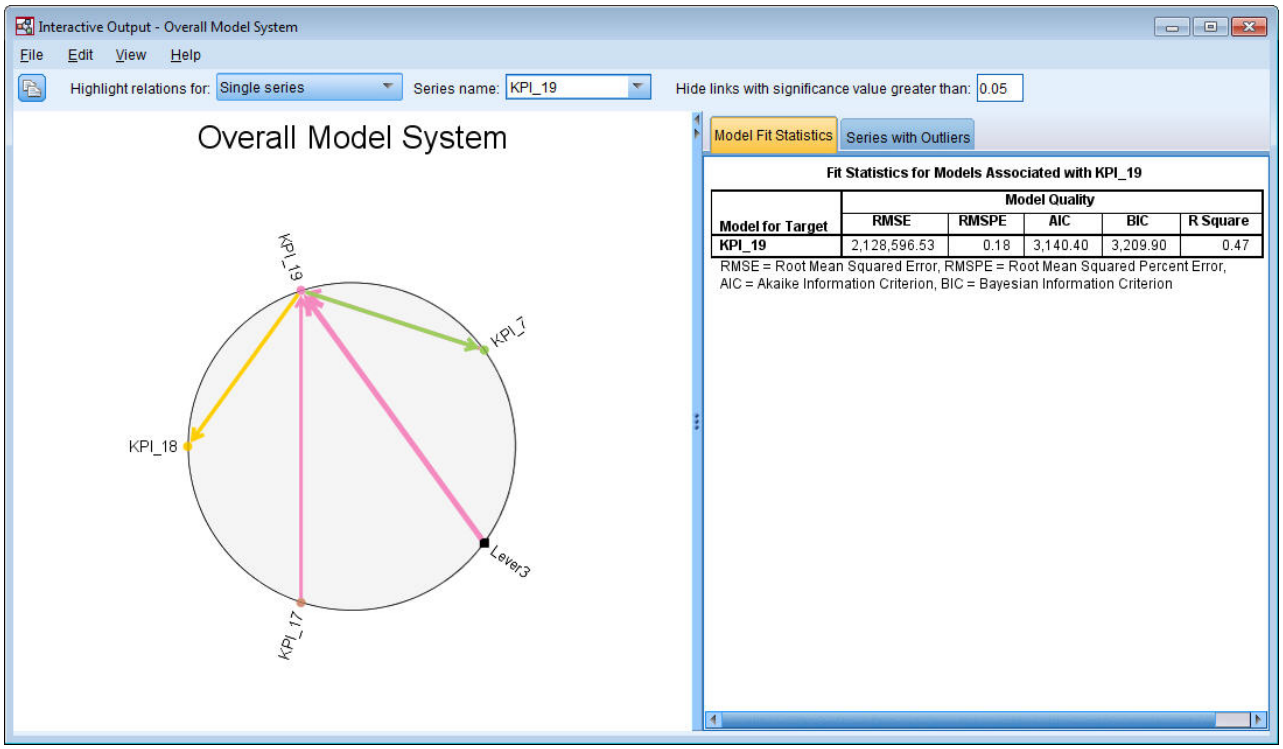
Rysunek 405. Ogólny system modelu, widok dla wszystkich szeregów

Wszystkie linie łączące konkretną zmienną przewidywaną z jej danymi wejściowymi mają ten sam kolor, a strzałka na każdej linii jest skierowana ze strony danych wejściowych w stronę ich zmiennej przewidywanej. Na przykład zmienna *Lever3* jest danymi wejściowymi zmiennej *KPI_19*.

Grubość każdej linii oznacza istotność związku przyczynowego, przy czym grubsze linie oznaczają bardziej istotny związek. Domyślnie związki przyczynowe z wartością istotności większą niż 0,05 są ukryte. Przy poziomie 0,05 tylko zmienne *Lever1*, *Lever3*, *Lever4* i *Lever5* mają istotne związki przyczynowe ze zmiennymi kluczowych wskaźników wydajności. Można zmienić wartość graniczną poziomu istotności, wprowadzając wartość w polu, które jest oznaczone **Ukryj połączenia o istotności większej niż**.

Oprócz odkrycia relacji przyczynowych pomiędzy zmiennymi *Lever* i zmiennymi kluczowych wskaźników wydajności, analiza odkryła również relacje pomiędzy zmiennymi kluczowych wskaźników wydajności. Na przykład zmienna *KPI_10* została wybrana jako dane wejściowe w modelu dla zmiennej *KPI_2*.

Można filtrować widok, aby wyświetlić tylko relacje dla pojedynczego szeregu. Na przykład, aby wyświetlić tylko relacje dla zmiennej *KPI_19*, kliknij etykietę zmiennej *KPI_19*, kliknij prawym przyciskiem myszy i wybierz opcję **Podświetl relacje dla szeregu**.



Rysunek 406. Ogólny system modelu, widok dla pojedynczego szeregu

Ten widok przedstawia dane wejściowe dla zmiennej *KPI_19*, które mają wartość istotności niższą lub równą 0,05. Wykres pokazuje również, że przy poziomie istotności 0,05 zmienna *KPI_19* została wybrana jako dane wejściowe dla dwóch zmiennych: *KPI_18* i *KPI_7*.

Oprócz wyświetlania relacji dla wybranych szeregów, pozycja wyników zawiera również informacje o wartościach odstających, które wykryto dla szeregu. Kliknij kartę **Szereg z wartościami odstającymi**.

Series with Outliers for KPI_19

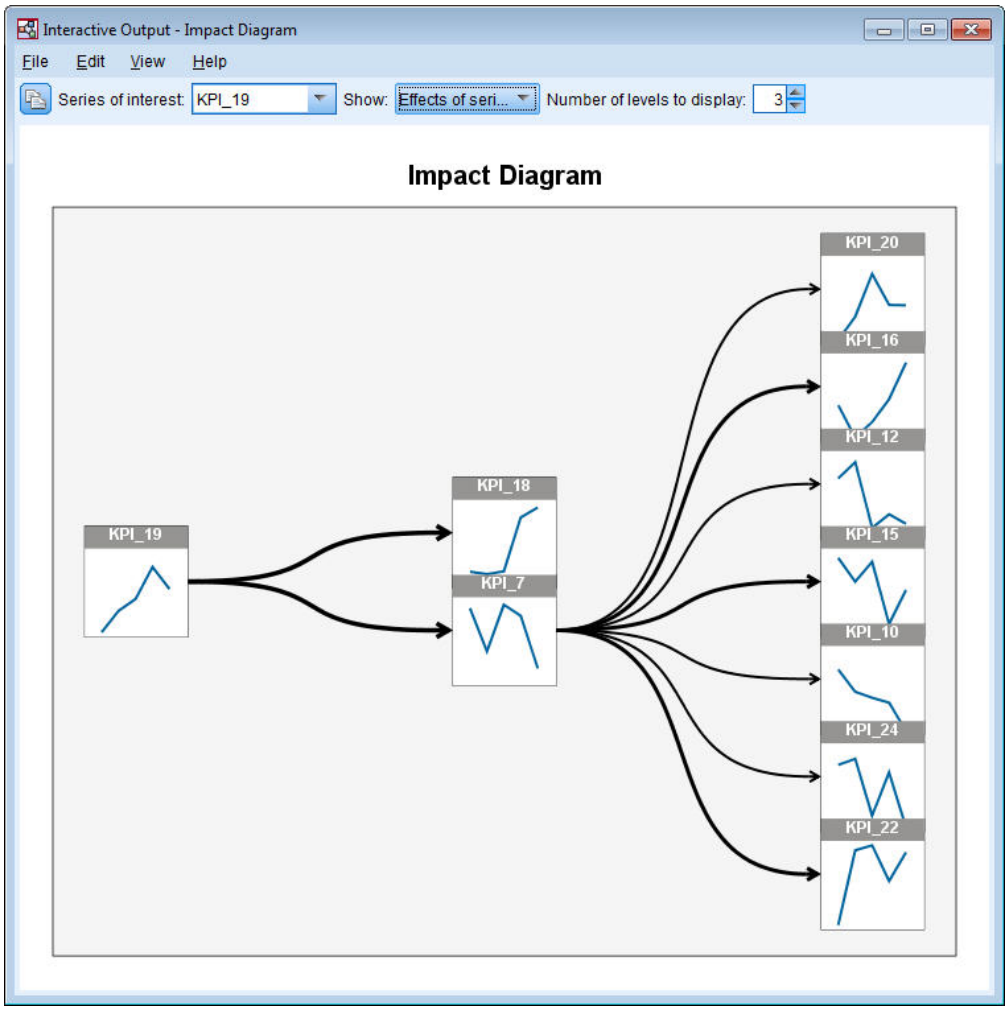
Series	Time	Observed Value
KPI_19	2008-10-12	7,358,201.68
	2009-04-05	2.10E+007
	2010-09-19	6,492,157.97

Rysunek 407. Wartości odstające dla zmiennej KPI_19

Wykryto trzy wartości odstające dla zmiennej *KPI_19*. Biorąc pod uwagę system modelu, który zawiera wszystkie odkryte połączenia, można wyjść poza wykrywanie wartości odstających i określić szereg, który prawdopodobnie powoduje powstanie konkretnej wartości odstającej. Taki typ analizy to Analiza podstawowych przyczyn wartości odstających, która została omówiona w dalszym temacie tego studium przypadku.

Diagramy wpływu

Można uzyskać pełny wgląd we wszystkie relacje, które są powiązane z określonym szeregiem, generując diagram wpływu. Kliknij etykietę zmiennej *KPI_19* na wykresie Ogólny system modelu, kliknij prawym przyciskiem myszy i wybierz opcję **Utwórz Diagram wpływu**.



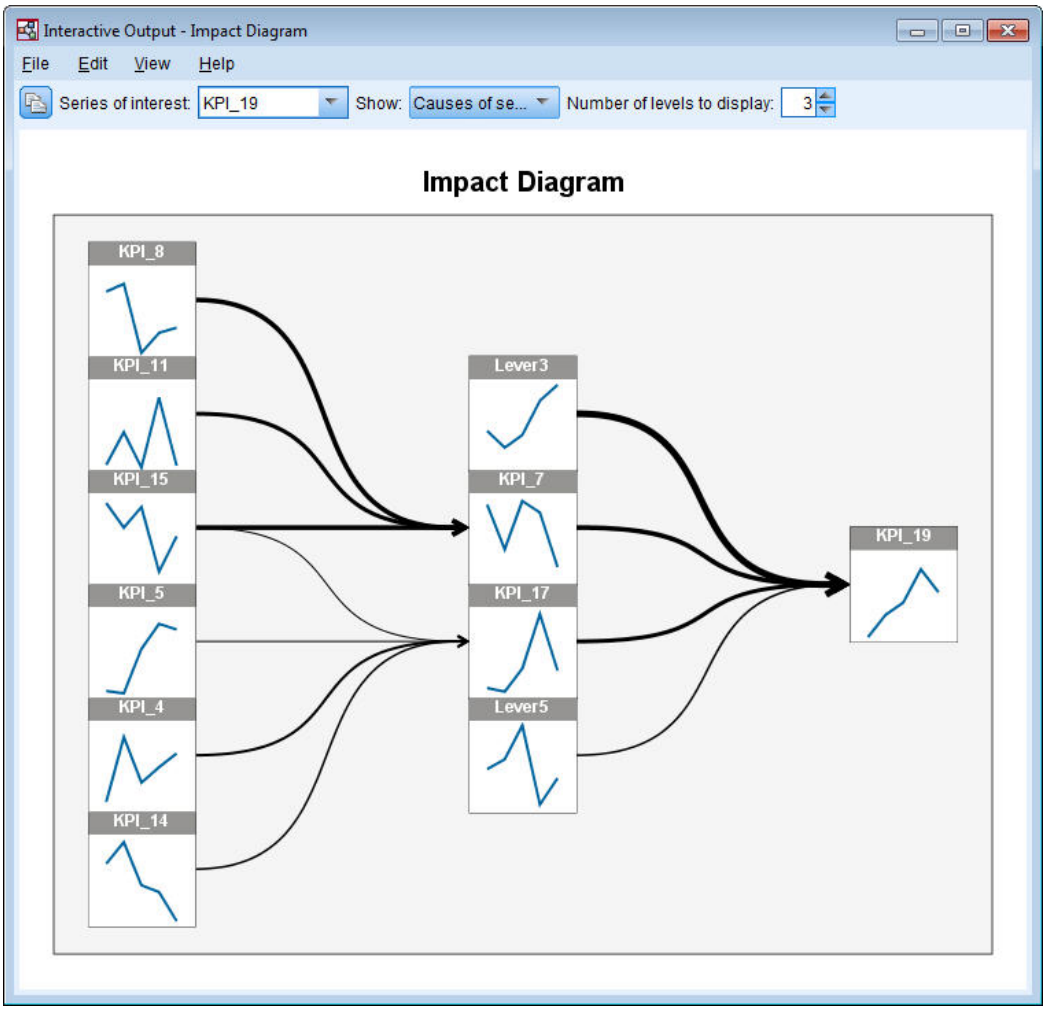
Rysunek 408. Diagram wpływu efektów

Kiedy diagram wpływu jest tworzony z Ogólnego systemu modelu, tak jak w tym przypadku, diagram wstępnie przedstawia szeregi, na które wpływa wybrany szereg. Domyślnie diagramy wpływu pokazują trzy poziomy wyników, gdzie pierwszy poziom jest tylko badanym szeregiem. Każdy dodatkowy poziom przedstawia bardziej pośrednie wyniki szeregu będącego przedmiotem zainteresowania. Można zmienić wartość **Liczba poziomów do wyświetlenia**, aby wyświetlić więcej lub mniej poziomów efektów. Diagram wpływu dla tego przykładu pokazuje, że zmienna *KPI_19* stanowi bezpośrednie dane wejściowe zmiennych *KPI_18* i *KPI_7*, ale pośrednio wpływa na więcej szeregów przez wpływ na szereg *KPI_7*. Tak jak w przypadku ogólnego systemu modelu grubość linii wskazuje istotność związków przyczynowych.

Wykres, który jest wyświetlany w każdym węźle diagramu wpływu, pokazuje ostatnie L+1 wartości powiązanego szeregu na koniec okresu szacowania i wartości prognozy, gdzie L jest liczbą składników opóźnień, które zostały uwzględnione w każdym modelu. Można uzyskać szczegółowy wykres sekwencji tych wartości, klikając powiązany węzeł.

Podwójne kliknięcie węzła ustawia powiązany szereg jako szereg badany i odświeża diagram wpływu na podstawie tego szeregu. Można również określić nazwę szeregu w polu **Interesujący szereg**, aby wybrać inny szereg do badania.

Diagramy wpływu mogą również przedstawiać szeregi, które wpływają na badany szereg. Takie szeregi są określane jako *przyczyny*. Aby zobaczyć szereg, który wpływa na zmienną *KPI_19*, wybierz pozycje **Przyczyny szeregu** z listy rozwijanej **Przedstaw**.



Rysunek 409. Diagram wpływu przyczyn

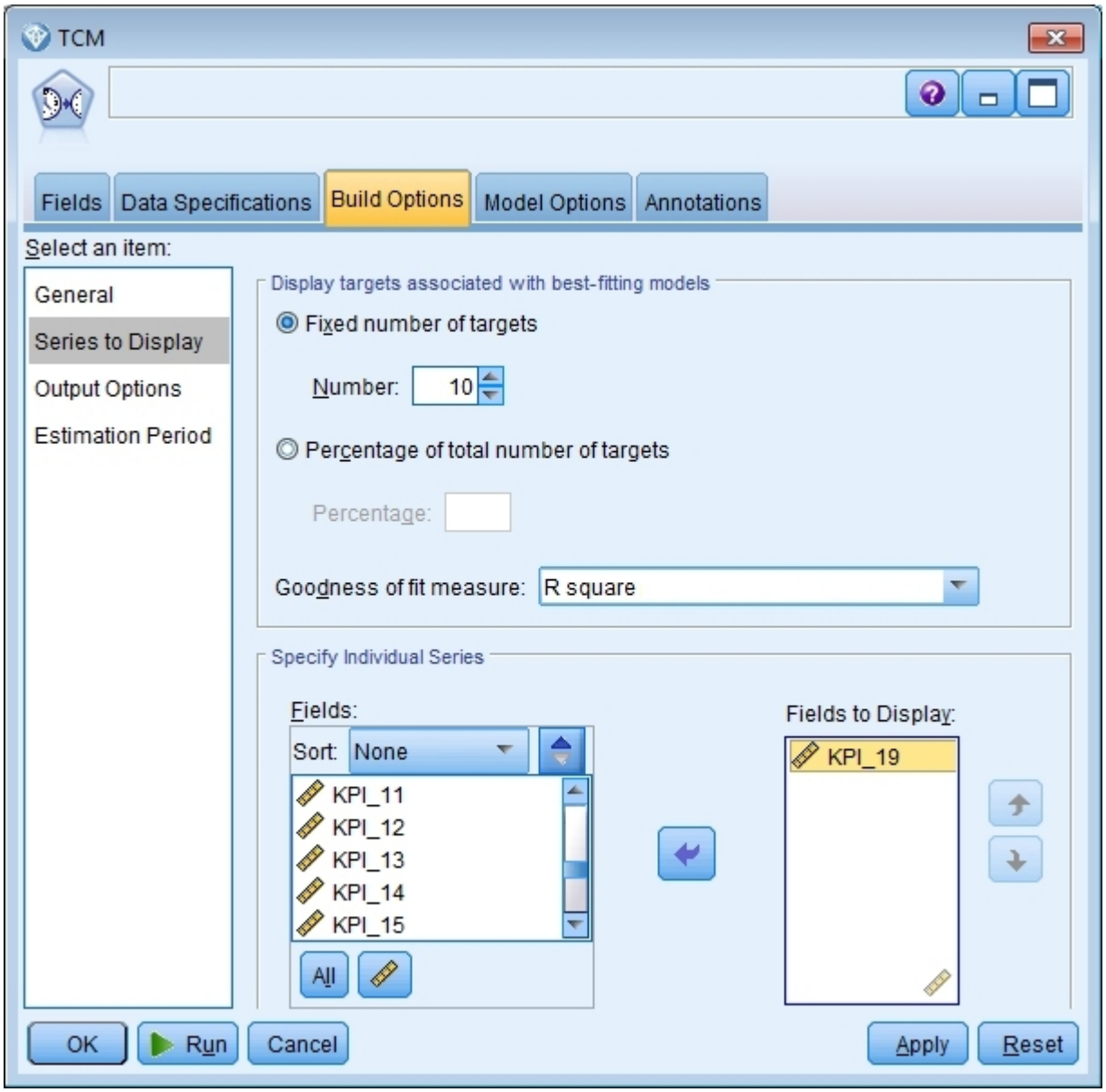
Ten widok pokazuje, że model dla zmiennej *KPI_19* ma cztery zmienne wejściowe i zmienna *Lever3* ma najbardziej istotny związek przyczynowy ze zmienną *KPI_19*. Diagram przedstawia również szeregi, które pośrednio wpływają na zmienną *KPI_19* przez ich wpływ na zmienne *KPI_7* i *KPI_17*. Ta sama koncepcja poziomów, którą omówiono w przypadku wpływu, dotyczy również przyczyn. Podobnie można zmienić wartość **Liczba poziomów do wyświetlenia**, aby wyświetlić więcej lub mniej poziomów przyczyn.

Określanie przyczyn podstawowych wartości odstających

W systemie modelu przyczynowego szeregów czasowych można wykroczyć poza wykrywanie wartości odstających i określić szeregi, które prawdopodobnie powodują konkretną wartość odstającą. Ten proces to Analiza podstawowych przyczyn wartości odstających i musi być uruchamiany osobno dla każdego szeregu. Analiza wymaga systemu modelu przyczynowego szeregów czasowych oraz danych, które były używane do zbudowania systemu. W tym przykładzie aktywnym zbiorem danych są dane, które były używane do zbudowania systemu modelu.

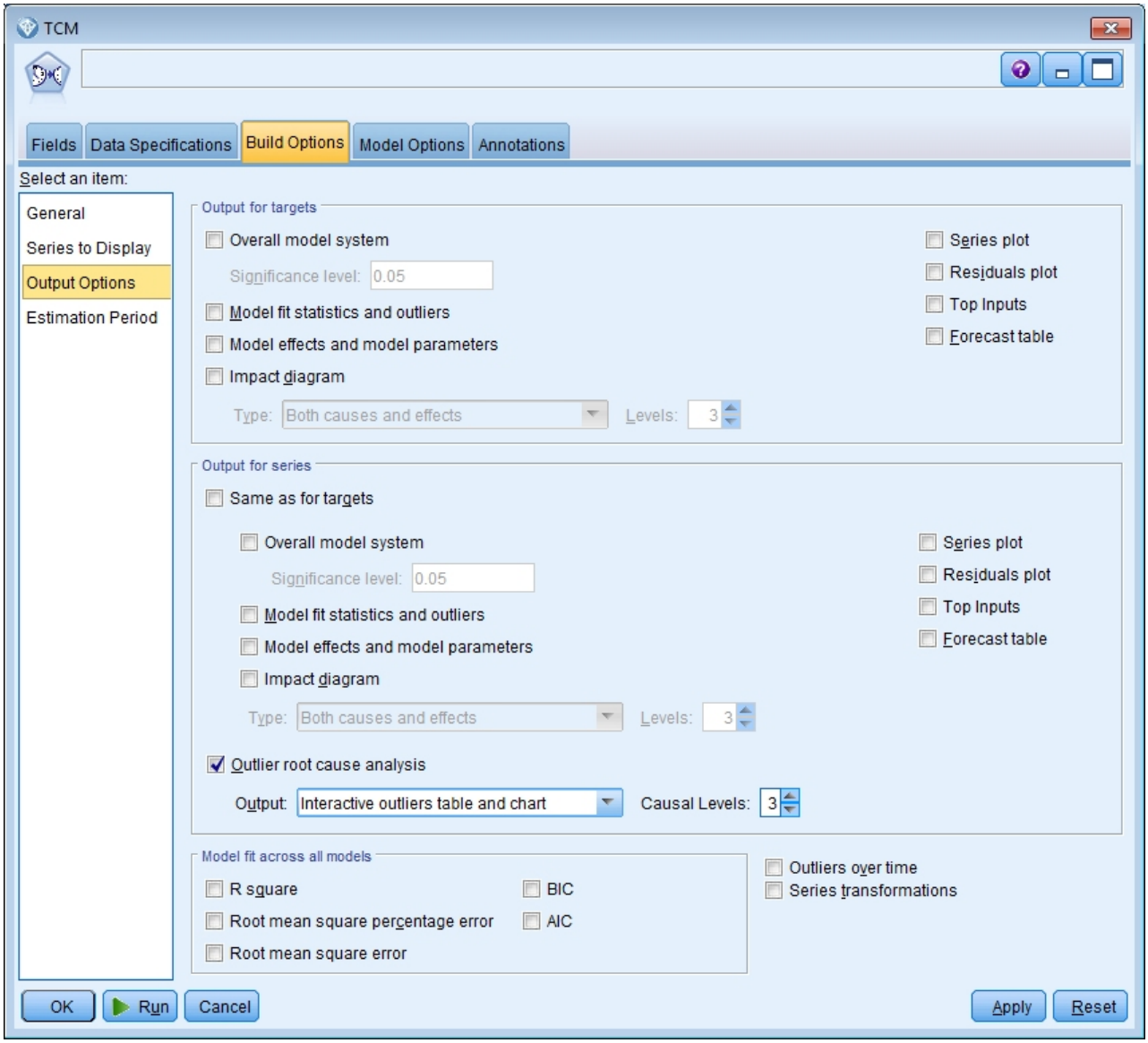
Aby uruchomić analizę podstawowych przyczyn wartości odstających:

1. W oknie dialogowym TCM przejdź na kartę **Opcje budowania**, a następnie kliknij opcję **Prezentacja szeregów** na liście **Wybierz element** .



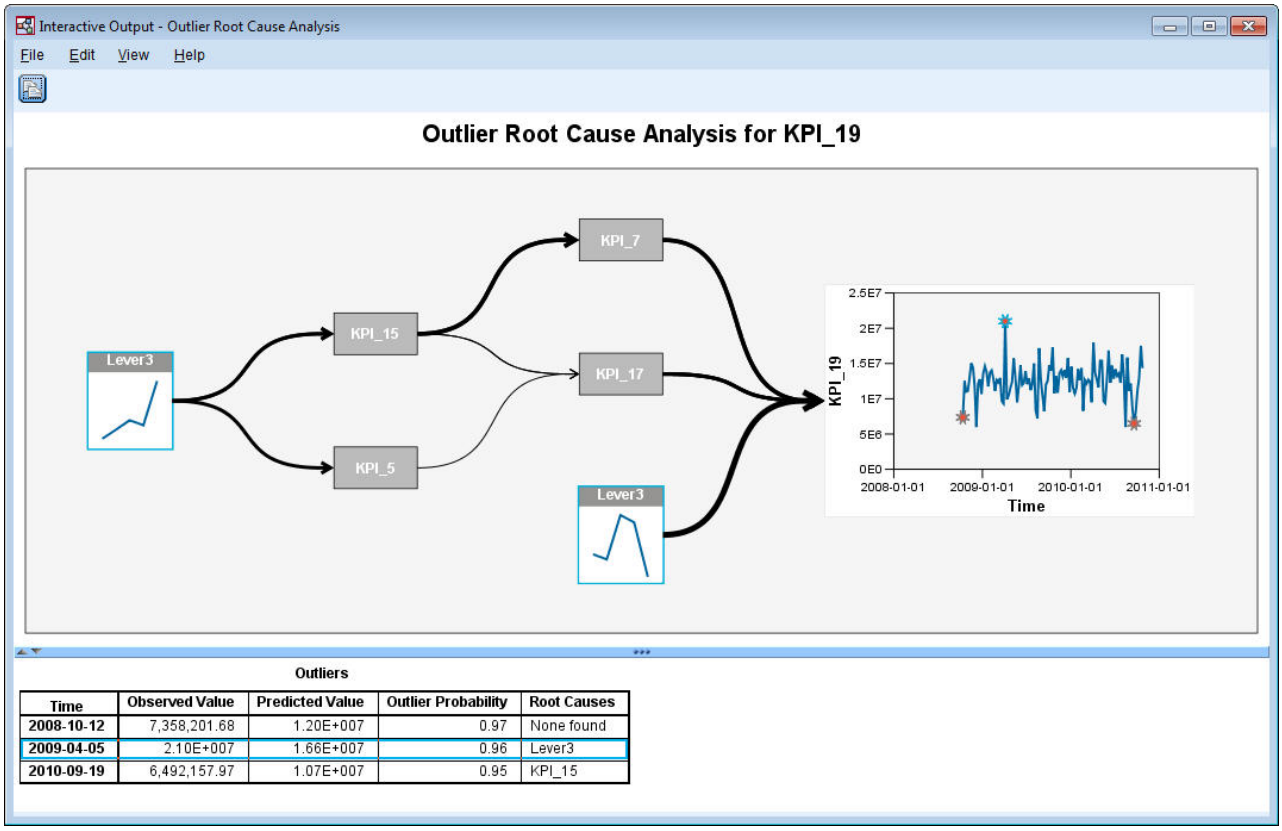
Rysunek 410. Szereg modelu przyczynowego szeregów czasowych do wyświetlenia

2. Przenieś zmienną *KPI_19* na listę **Prezentowane zmienne**.
3. Kliknij pozycję **Opcje wyników** na liście **Wybierz element** karty Opcje.



Rysunek 411. Opcje wyników modelu przyczynowego szeregów czasowych

4. Usuń zaznaczenia opcji **Ogólny system modelu**, **Wyniki tego samego typu jak dla zmiennych przewidywanych**, **R-kwadrat** i **Transformacje szeregów**.
5. Zaznacz opcję **Analiza podstawowych przyczyn wartości odstających** i zachowaj istniejące ustawienia pozycji **Wynik** i **Poziomy przyczynowe**.
6. Kliknij przycisk **Uruchom**.
7. Dwukrotnie kliknij wykres Analiza podstawowych przyczyn wartości odstających dla zmiennej *KPI_19* w przeglądarce, aby go aktywować.



Rysunek 412. Analiza podstawowych przyczyn wartości odstających dla zmiennej KPI_19

Wyniki analizy są podsumowane w tabeli Wartości odstające. Tabela pokazuje, że wykryto przyczyny podstawowe dla wartości odstających 2009-04-05 i 2010-09-19, ale nie wykryto przyczyny podstawowej dla wartości odstającej z dnia 2008-10-12. Kliknięcie wiersza w tabeli Wartości odstające podświetla ścieżkę do szeregu przyczyny podstawowej, jak pokazano na przykładzie dla wartości odstającej z dnia 2009-04-05. Ta czynność podświetla również wybraną wartość odstającą na wykresie sekwencji. Można również kliknąć ikonę dla wartości odstającej bezpośrednio na wykresie sekwencji, aby podświetlić ścieżkę do szeregu przyczyny podstawowej dla tej wartości odstającej.

Dla wartości odstającej z dnia 2009-04-05 przyczyną podstawową jest zmienna *Lever3*. Diagram pokazuje, że zmienna *Lever3* jest bezpośrednią zmienną wejściową dla zmiennej *KPI_19*, ale również pośrednio wpływa na zmienną *KPI_19* przez swój wpływ na szereg, który wpływa na zmienną *KPI_19*. Jednym z konfigurowalnych parametrów analizy podstawowych przyczyn wartości odstających jest liczba poziomów przyczynowych do wyszukiwania przyczyn podstawowych. Domyślnie wyszukiwane są trzy poziomy. Wystąpienia szeregów przyczyn podstawowych są wyświetlane do określonej liczby poziomów przyczynowych. W tym przykładzie zmienna *Lever3* występuje zarówno na pierwszym poziomie przyczynowym, jak i na trzecim poziomie przyczynowym.

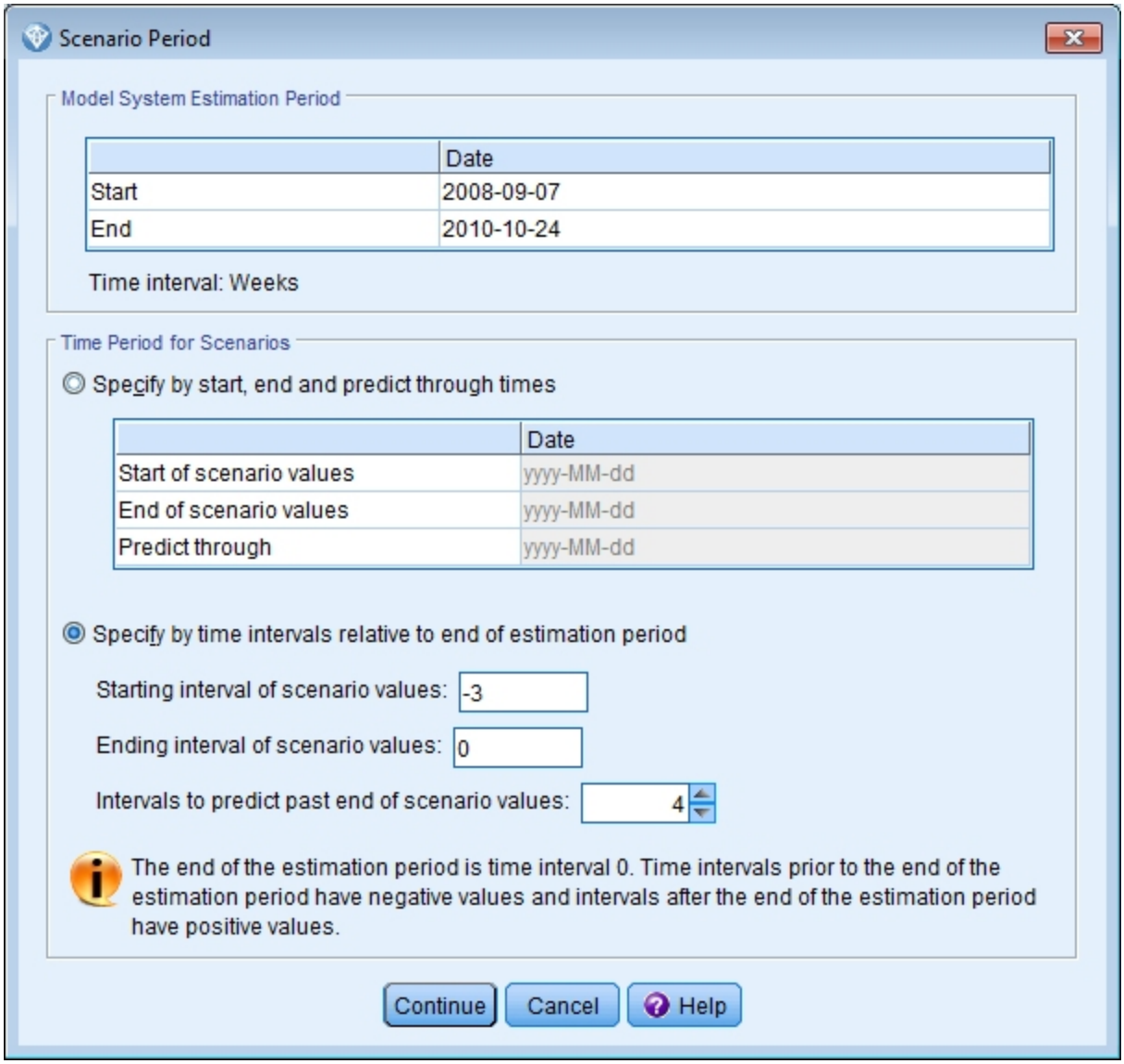
Każdy węzeł na podświetlonej ścieżce dla obserwacji odstającej zawiera wykres, którego zakres czasu zależy od poziomu, na którym występuje węzeł. Dla węzłów na pierwszym poziomie przyczynowym zakres to od T-1 do T-L, gdzie T to czas, w którym występuje obserwacja odstająca, a L to liczba składników opóźnienia, które są uwzględnione w każdym modelu. Dla węzłów na drugim poziomie przyczynowym zakres to od T-2 do T-L-1, a dla trzeciego poziomu zakres to od T-3 do T-L-2. Można uzyskać szczegółowy wykres sekwencji tych wartości, klikając powiązany węzeł.

Uruchamianie scenariuszy

Używając systemu modelowania przyczynowego szeregów czasowych, można uruchamiać scenariusze zdefiniowane przez użytkownika. Zmienna *scenario* jest definiowana przez szereg czasowy będący szeregiem źródłowym (*root series*) oraz przez zestaw zdefiniowanych przez użytkownika wartości dla tego szeregu w podanym zakresie czasu. Podane wartości są następnie używane do generowania predykcji dla szeregów czasowych, na które wpływa szereg źródłowy. Analiza wymaga systemu modelu przyczynowego szeregów czasowych oraz danych, które były używane do zbudowania systemu. W tym przykładzie aktywnym zbiorem danych są dane, które były używane do zbudowania systemu modelu.

Aby uruchomić scenariusze:

1. W oknie dialogowym wyników TCM kliknij przycisk **Analiza scenariusza**.
2. W oknie dialogowym Scenariusze modeli przyczynowych szeregów czasowych kliknij opcję **Zdefiniuj okres scenariusza**.



Rysunek 413. Okres scenariusza

3. Zaznacz opcję **Wyznaczony przez przedziały zależne od końca okresu estymacji**.

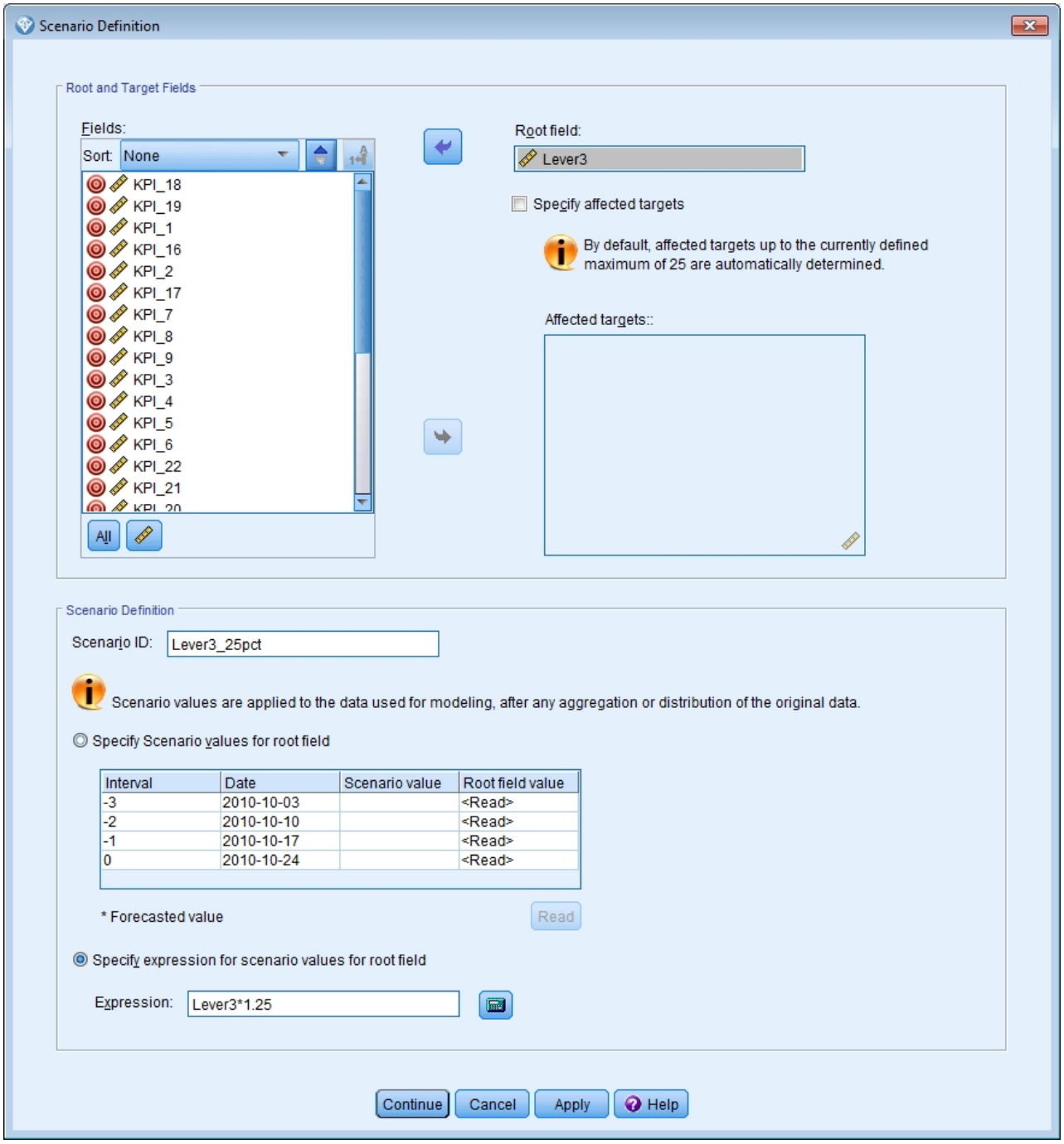
4. Wprowadź -3 dla początkowego przedziału i wprowadź 0 dla przedziału końcowego.

   Te ustawienia określają, że każdy scenariusz opiera się na wartościach, które są określone dla ostatnich czterech przedziałów czasowych w okresie estymacji. W tym przykładzie ostatnie cztery przedziały czasowe to ostatnie cztery tygodnie. Zakres czasu, w którym określane są wartości scenariusza, jest określany jako *okres scenariusza*.

5. Wprowadź 4 dla przedziałów, aby przewidywać poza końcem wartości scenariuszy.

   To ustawienie określa, że predykcje są generowane przez cztery przedziały czasu poza końcem okresu scenariusza.

6. Kliknij przycisk **Kontynuuj**.
7. Kliknij opcję **Dodaj scenariusz** na karcie Scenariusze.

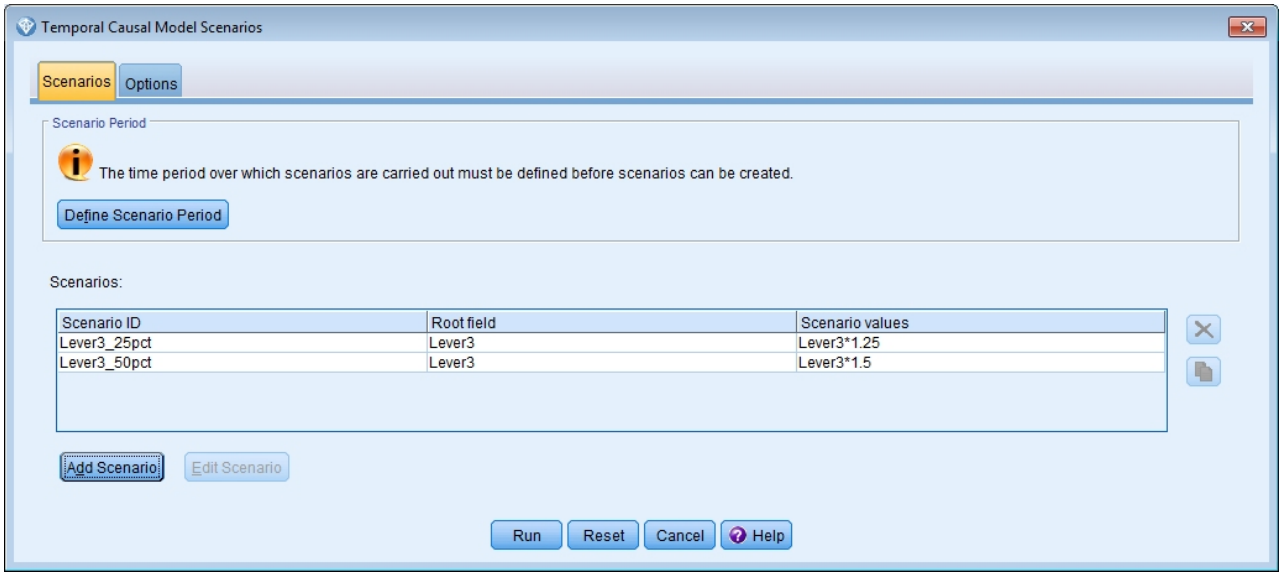


Rysunek 414. Definicja scenariusza

8. Przenieś zmienną *Lever3* do pola **Zmienna źródłowa**, aby zbadać, jak określone wartości zmiennej *Lever3* w okresie scenariusza wpływają na predykcje innych szeregów, które są przyczynowo powiązane ze zmienną *Lever3*.
9. Wprowadź Lever3_25pct w polu Identyfikator scenariusza.
10. Zaznacz pole **Określ wyrażenie dla wartości scenariusza dla zmiennej źródłowej** i wprowadź Lever3*1,25 jako wyrażenie.

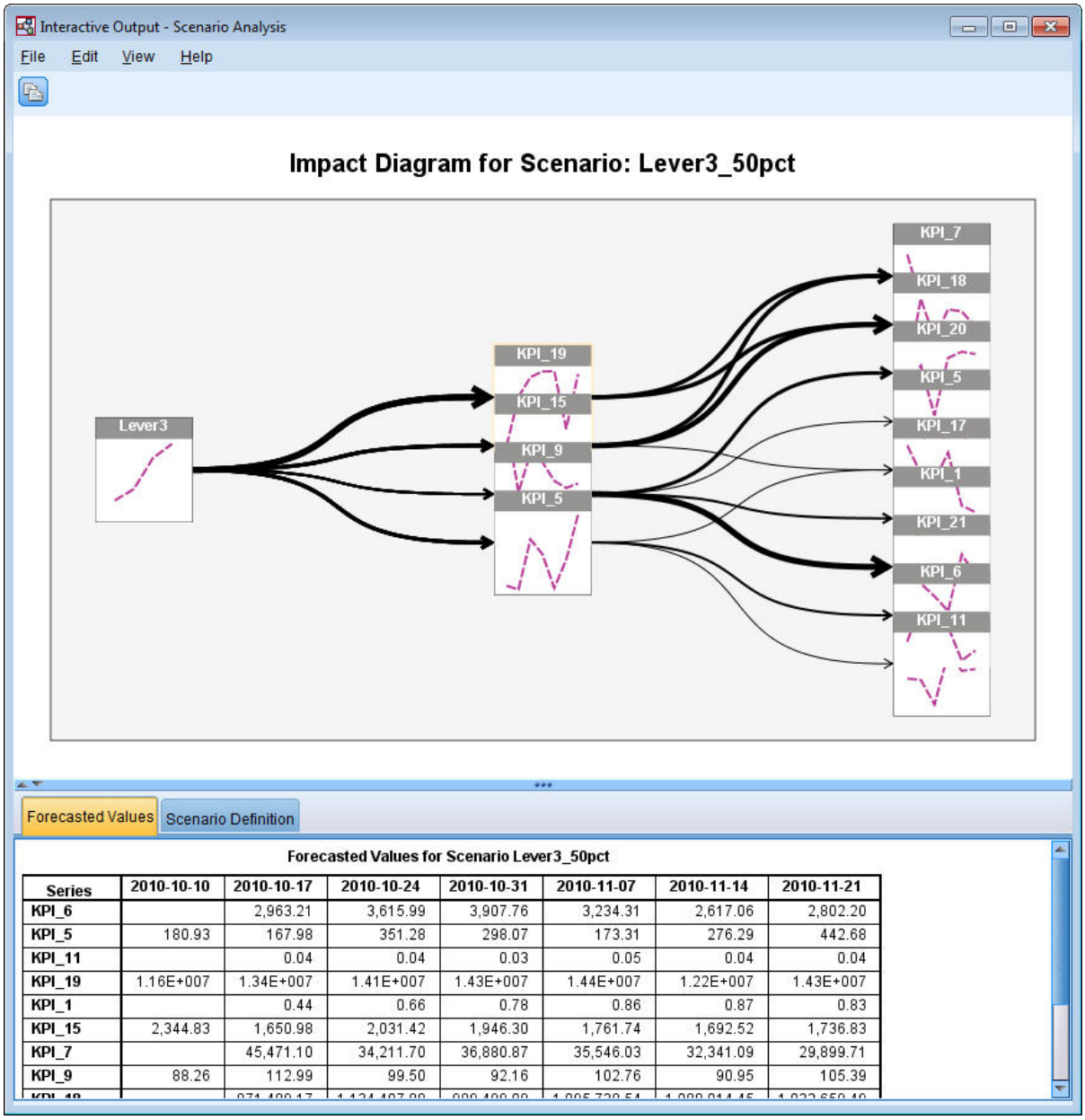
    To ustawienie określa, że wartości zmiennej *Lever3* w okresie scenariusza są 25% większe niż obserwowane wartości. Dla bardziej skomplikowanych wyrażeń można użyć konstruktora wyrażeń, klikając ikonę kalkulatora.
11. Kliknij przycisk **Kontynuuj**.

12. Powtórz kroki 10–14, aby zdefiniować scenariusz, który ma zmienną *Lever3* jako zmienną źródłową, Lever3_50pct jako Identyfikator scenariusza i Lever3*1,5 jako wyrażenie.



Rysunek 415. Scenariusze

13. Kliknij kartę **Opcje** i wprowadź *2* dla maksymalnego poziomu docelowych zmiennych przewidywanych.
14. Kliknij przycisk **Uruchom**.
15. Dwukrotnie kliknij wykres Diagram wpływu dla scenariusza *Lever3_50pct* w przeglądarce, aby go aktywować.



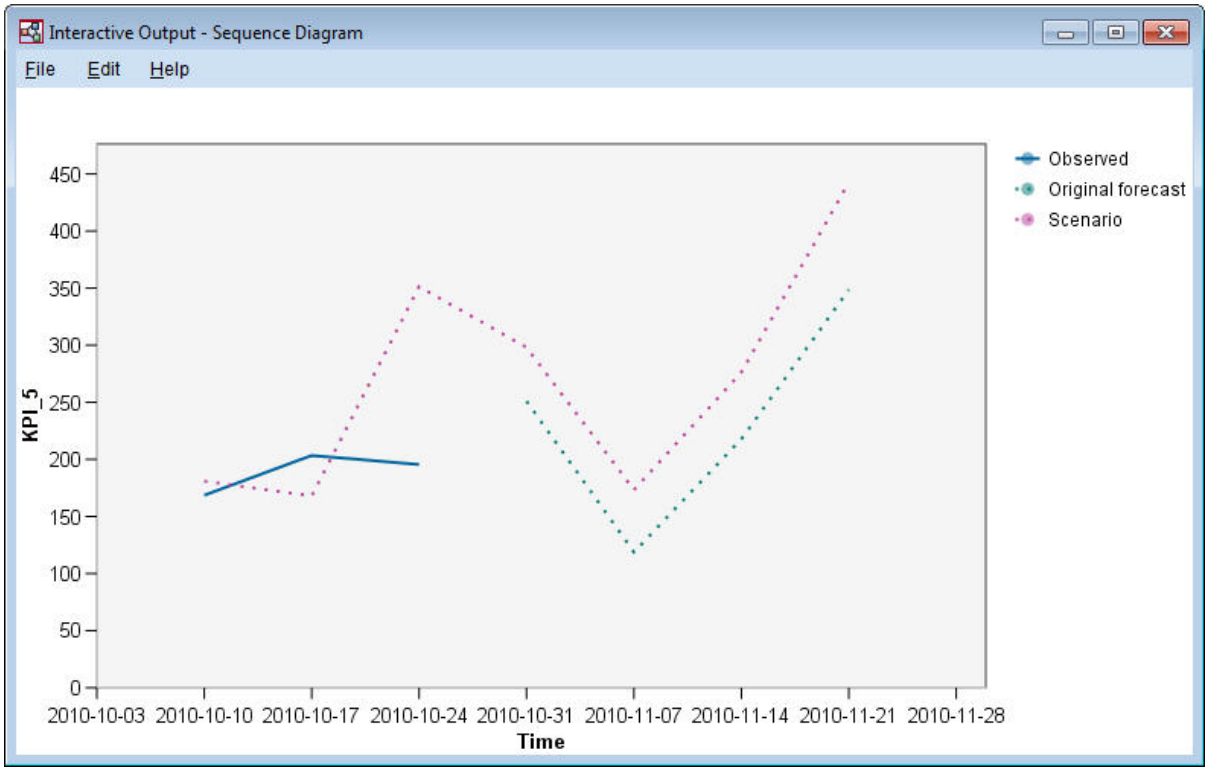
Series	2010-10-10	2010-10-17	2010-10-24	2010-10-31	2010-11-07	2010-11-14	2010-11-21
KPI_6		2,963.21	3,615.99	3,907.76	3,234.31	2,617.06	2,802.20
KPI_5	180.93	167.98	351.28	298.07	173.31	276.29	442.68
KPI_11		0.04	0.04	0.03	0.05	0.04	0.04
KPI_19	1.16E+007	1.34E+007	1.41E+007	1.43E+007	1.44E+007	1.22E+007	1.43E+007
KPI_1		0.44	0.66	0.78	0.86	0.87	0.83
KPI_15	2,344.83	1,650.98	2,031.42	1,946.30	1,761.74	1,692.52	1,736.83
KPI_7		45,471.10	34,211.70	36,880.87	35,546.03	32,341.09	29,899.71
KPI_9	88.26	112.99	99.50	92.16	102.76	90.95	105.39

Rysunek 416. Diagram wpływu dla scenariusza Lever3_50pct

Diagram wpływu przedstawia szereg, na który wpływa szereg początkowy *Lever3*. Pokazane są dwa poziomy efektów, ponieważ określono *2* dla maksymalnego poziomu docelowych zmiennych przewidywanych.

Tabela Prognozowane wartości obejmuje predykcje dla wszystkich szeregów, na które wpływa zmienna *Lever3*, do drugiego poziomu efektów. Predykcje dla szeregu przewidywanego na pierwszym poziomie efektów rozpoczynają się od pierwszego okresu po rozpoczęciu okresu scenariusza. W tym przykładzie predykcje dla szeregu przewidywanego na pierwszym poziomie rozpoczynają się dnia 2010-10-10. Predykcje dla szeregu przewidywanego na drugim poziomie efektów rozpoczynają się od drugiego okresu po rozpoczęciu okresu scenariusza. W tym przykładzie predykcje dla szeregu przewidywanego na drugim poziomie rozpoczynają się dnia 2010-10-17. Przemienny charakter predykcji odzwierciedla fakt, że modele szeregów czasowych opierają się na opóźnionych wartościach danych wejściowych.

16. Kliknij węzeł dla zmiennej *KPI_5*, aby wygenerować szczegółowy diagram sekwencji.



Rysunek 417. Diagram sekwencji dla zmiennej KPI_5

Wykres sekwencji przedstawia przewidywane wartości ze scenariusza i pokazuje również wartości szeregu przy braku scenariusza. Kiedy okres scenariusza zawiera czasy w okresie szacowania, wyświetlane są obserwowane wartości szeregu. Dla czasów poza końcem okresu szacowania wyświetlane są oryginalne prognozy.

Uwagi

Niniejsza publikacja została przygotowana z myślą o produktach i usługach oferowanych w Stanach Zjednoczonych. Materiał ten jest również dostępny w IBM w innych językach. Jednakże w celu uzyskania dostępu do takiego materiału istnieje konieczność posiadania egzemplarza produktu w takim języku.

Produktów, usług lub opcji opisywanych w tym dokumencie IBM nie musi oferować we wszystkich krajach. Informacje o produktach i usługach dostępnych w danym kraju można uzyskać od lokalnego przedstawiciela IBM. Odwołanie do produktu, programu lub usługi IBM nie oznacza, że można użyć wyłącznie tego produktu, programu lub usługi IBM. Zamiast nich można zastosować ich odpowiednik funkcjonalny pod warunkiem, że nie narusza to praw własności intelektualnej IBM. Jednakże cała odpowiedzialność za ocenę przydatności i sprawdzenie działania produktu, programu lub usługi pochodzących od producenta innego niż IBM spoczywa na użytkowniku.

IBM może posiadać patenty lub złożone wnioski patentowe na towary i usługi, o których mowa w niniejszej publikacji. Przedstawienie niniejszej publikacji nie daje żadnych uprawnień licencyjnych do tychże patentów. Pisemne zapytania w sprawie licencji można przesyłać na adres:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

Zapytania dotyczące zestawów znaków dwubajtowych (DBCS) należy kierować do lokalnych działów własności intelektualnej IBM (IBM Intellectual Property Department) lub wysłać je na piśmie na adres:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan, Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokio 103-8510, Japonia

INTERNATIONAL BUSINESS MACHINES CORPORATION DOSTARCZA TĘ PUBLIKACJĘ W STANIE, W JAKIM SIĘ ZNAJDUJE ("AS IS") BEZ UDZIELANIA JAKICHKOLWIEK GWARANCJI (RĘKOJMIĘ RÓWNIEŻ WYŁĄCZA SIĘ), WYRAŹNYCH LUB DOMNIEMANYCH, A W SZCZEGÓLNOŚCI DOMNIEMANYCH GWARANCJI PRZYDATNOŚCI HANDLOWEJ, PRZYDATNOŚCI DO OKREŚLONEGO CELU ORAZ GWARANCJI, ŻE PUBLIKACJA TA NIE NARUSZA PRAW OSÓB TRZECICH. Ustawodawstwa niektórych krajów nie dopuszczają zastrzeżeń dotyczących gwarancji wyraźnych lub domniemanych w odniesieniu do pewnych transakcji; w takiej sytuacji powyższe zdanie nie ma zastosowania.

Informacje zawarte w niniejszej publikacji mogą zawierać nieścisłości techniczne lub błędy drukarskie. Informacje te są okresowo aktualizowane, a zmiany te zostaną uwzględnione w kolejnych wydaniach tej publikacji. IBM zastrzega sobie prawo do wprowadzania ulepszeń i/lub zmian w produktach i/lub programach opisanych w tej publikacji w dowolnym czasie, bez wcześniejszego powiadomienia.

Wszelkie wzmianki w tej publikacji na temat stron internetowych innych podmiotów zostały wprowadzone wyłącznie dla wygody użytkownika i w żadnym wypadku nie stanowią zachęty do ich odwiedzania. Materiały dostępne na tych stronach nie są częścią materiałów opracowanych dla tego produktu IBM, a użytkownik korzysta z nich na własną odpowiedzialność.

IBM ma prawo do używania i rozpowszechniania informacji przysłanych przez użytkownika w dowolny sposób, jaki uzna za właściwy, bez żadnych zobowiązań wobec ich autora.

Licencjobiorcy tego programu, którzy chcieliby uzyskać informacje na temat programu w celu: (i) wdrożenia wymiany informacji między niezależnie utworzonymi programami i innymi programami (łącznie z tym opisywanym) oraz (ii) wspólnego wykorzystywania wymienianych informacji, powinni skontaktować się z:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

Informacje takie mogą być udostępnione, o ile spełnione zostaną odpowiednie warunki, w tym, w niektórych przypadkach, zostanie uiszczona stosowna opłata.

Licencjonowany program opisany w niniejszej publikacji oraz wszystkie inne licencjonowane materiały dostępne dla tego programu są dostarczane przez IBM na warunkach określonych w Umowie IBM z Klientem, Międzynarodowej Umowie Licencyjnej IBM na Program lub w innych podobnych umowach zawartych między IBM i użytkownikami.

Dane dotyczące wydajności i cytowane przykłady zostały przedstawione jedynie w celu zobrazowania sytuacji. Faktyczne wyniki dotyczące wydajności mogą się różnić w zależności do konkretnych warunków konfiguracyjnych i operacyjnych.

Informacje dotyczące produktów innych podmiotów niż IBM zostały uzyskane od dostawców tych produktów, z ich publicznych ogłoszeń lub innych dostępnych publicznie źródeł. IBM nie testował tych produktów i nie może potwierdzić dokładności pomiarów wydajności, kompatybilności ani żadnych innych danych związanych z tymi produktami. Pytania dotyczące możliwości produktów innych podmiotów należy kierować do dostawców tych produktów.

Wszelkie stwierdzenia dotyczące przyszłych kierunków rozwoju i zamierzeń IBM mogą zostać zmienione lub wycofane bez powiadomienia.

Publikacja ta zawiera przykładowe dane i raporty używane w codziennej pracy. W celu kompleksowego ich zilustrowania, podane przykłady zawierają nazwiska osób prywatnych, nazwy przedsiębiorstw oraz nazwy produktów. Wszystkie te nazwy/nazwiska są fikcyjne i jakiekolwiek podobieństwo do istniejących nazw/nazwisk jest całkowicie przypadkowe.

Znaki towarowe

IBM, logo IBM i ibm.com są znakami towarowymi lub zastrzeżonymi znakami towarowymi International Business Machines Corp. zarejestrowanymi w wielu systemach prawnych na całym świecie. Pozostałe nazwy produktów i usług mogą być znakami towarowymi IBM lub innych przedsiębiorstw. Aktualna lista znaków towarowych IBM dostępna jest w serwisie WWW IBM, w sekcji "Copyright and trademark information" Informacje o prawach autorskich i znakach towarowych), pod adresem www.ibm.com/legal/copytrade.shtml.

Adobe, logo Adobe, PostScript oraz logo PostScript są znakami towarowymi lub zastrzeżonymi znakami towarowymi Adobe Systems Incorporated w Stanach Zjednoczonych i/lub w innych krajach.

Intel, logo Intel, Intel Inside, logo Intel Inside, Intel Centrino, logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium i Pentium są znakami towarowymi lub zastrzeżonymi znakami towarowymi Intel Corporation lub przedsiębiorstw podporządkowanych Intel Corporation w Stanach Zjednoczonych i w innych krajach.

Linux jest zastrzeżonym znakiem towarowym Linusa Torvaldsa w Stanach Zjednoczonych i/lub w innych krajach.

Microsoft, Windows, Windows NT oraz logo Windows są znakami towarowymi Microsoft Corporation w Stanach Zjednoczonych i/lub w innych krajach.

UNIX jest zastrzeżonym znakiem towarowym The Open Group w Stanach Zjednoczonych i/lub w innych krajach.

Java oraz wszystkie znaki towarowe i logo dotyczące języka Java są znakami towarowymi lub zastrzeżonymi znakami towarowymi Oracle i/lub przedsiębiorstw afiliowanych.

Warunki dotyczące dokumentacji produktu

Zezwolenie na korzystanie z tych publikacji jest przyznawane na poniższych warunkach.

Zakres stosowania

Niniejsze warunki stanowią uzupełnienie warunków używania serwisu WWW IBM.

Użytek osobisty

Użytkownik ma prawo kopiować te publikacje do własnego, niekomercyjnego użytku pod warunkiem zachowania wszelkich uwag dotyczących praw własności. Użytkownik nie ma prawa dystrybuować ani wyświetlać tych publikacji czy ich części, ani też wykonywać na ich podstawie prac pochodnych bez wyraźnej zgody IBM.

Użytek służbowy

Użytkownik ma prawo kopiować te publikacje, dystrybuować je i wyświetlać wyłącznie w ramach przedsiębiorstwa Użytkownika pod warunkiem zachowania wszelkich uwag dotyczących praw własności. Użytkownik nie ma prawa wykonywać na podstawie tych publikacji ani ich fragmentów prac pochodnych, kopiować ich, dystrybuować ani wyświetlać poza przedsiębiorstwem Użytkownika bez wyraźnej zgody IBM.

Prawa

Z wyjątkiem zezwoleń wyraźnie udzielonych w niniejszym dokumencie, nie udziela się jakichkolwiek innych zezwoleń, licencji ani praw, wyraźnych czy domniemanych, odnoszących się do tych publikacji czy jakichkolwiek informacji, danych, oprogramowania lub innej własności intelektualnej, o których mowa w niniejszym dokumencie.

IBM zastrzega sobie prawo do anulowania zezwolenia przyznanego w niniejszym dokumencie w każdej sytuacji, gdy, według uznania IBM, korzystanie z tych publikacji jest szkodliwe dla IBM lub jeśli IBM uzna, że warunki niniejszego dokumentu nie są przestrzegane.

Użytkownik ma prawo pobierać, eksportować lub reeksportować niniejsze informacje pod warunkiem zachowania bezwzględnej i pełnej zgodności z obowiązującym prawem i przepisami, w tym ze wszelkimi prawami i przepisami eksportowymi Stanów Zjednoczonych.

IBM NIE UDZIELA JAKICHKOLWIEK GWARANCJI, W TYM TAKŻE RĘKOJMI, DOTYCZĄCYCH TREŚCI TYCH PUBLIKACJI. PUBLIKACJE TE SĄ DOSTARCZANE W STANIE, W JAKIM SIĘ ZNAJDUJĄ ("AS-IS") BEZ UDZIELANIA JAKICHKOLWIEK GWARANCJI (RĘKOJMIĘ RÓWNIEŻ WYŁĄCZA SIĘ), WYRAŹNYCH CZY DOMNIEMANYCH, A W SZCZEGÓLNOŚCI DOMNIEMANYCH GWARANCJI PRZYDATNOŚCI HANDLOWEJ, PRZYDATNOŚCI DO OKREŚLONEGO CELU CZY NIENARUSZANIA PRAW OSÓB TRZECICH.

Indeks

R

S

Ś

T

U

W

Z

Drukowane w USA