IBM SPSS Data Preparation 29

Uwaga

Przed użyciem tych informacji i produktu, którego one dotyczą, przeczytaj informacje znajdujące się w sekcji “Uwagi” na stronie 27.

Informacje o produkcie

Niniejsze wydanie dotyczy wersji 29, wydania 0, modyfikacji 1 produktu IBM® SPSS Statistics oraz wszystkich kolejnych wydań i modyfikacji, dopóki nie zostanie to określone inaczej w nowych wydaniach.

Spis treści

Rozdział 1. Przygotowanie danych..... 1
Wprowadzenie do przygotowania danych..... 1
Korzystanie z procedur przygotowywania danych..... 1
Reguły walidacyjne..... 1
Ładowanie predefiniowanych reguł walidacyjnych..... 2
Definiowanie reguł walidacyjnych..... 2
Sprawdzenie poprawności danych..... 4
Walidacja danych: Sprawdzenia podstawowe..... 4
Walidacja danych: Reguły pojedynczej zmiennej..... 5
Walidacja danych: Reguły wielu zmiennych..... 6
Walidacja danych: Raport..... 6
Walidacja danych: Zapisz..... 6
Automatyczne przygotowywanie danych..... 7
Otrzymywanie automatycznego przygotowania danych..... 8
Otrzymywanie interaktywnego przygotowania danych..... 8
Karta Zmienne..... 8
Karta Ustawienia..... 9
Karta Analiza..... 13
Przywrócone oceny..... 19
Zidentyfikuj obserwacje nietypowe..... 20
Identyfikowanie obserwacji nietypowych: Wyniki..... 21
Identyfikowanie obserwacji nietypowych: Zapisz..... 21
Identyfikowanie obserwacji nietypowych: Braki danych..... 22
Identyfikowanie obserwacji nietypowych: Opcje..... 22
Dodatkowe właściwości komendy DETECTANOMALY..... 23
Kategoryzacja optymalna..... 23
Kategoryzacja optymalna: Wynik..... 23
Kategoryzacja optymalna: Zapisz..... 24
Kategoryzacja optymalna: Braki danych..... 24
Kategoryzacja optymalna: Opcje..... 24
Dodatkowe właściwości komendy OPTIMAL BINNING..... 25

Uwagi..... 27
Znaki towarowe..... 28

Indeks..... 31

Rozdział 1. Przygotowanie danych

Następujące funkcje przygotowania danych są dostępne w programie Base Edition.

Wprowadzenie do przygotowania danych

Wraz ze wzrostem mocy obliczeniowej komputerów rośnie apetyt na informacje, co prowadzi do gromadzenia coraz większej ilości danych — obserwacji, zmiennych oraz błędów podczas wprowadzania danych danych. Błędy te są zmorą prognozowania za pomocą modelu predykcyjnego, które jest ostatecznym celem magazynowania danych. Z tego względu należy dbać o „czystość” danych. Jednakże ilość przechowywanych danych wzrosła na tyle, że nie jest możliwe ręczne zweryfikowanie obserwacji i konieczne jest wdrożenie zautomatyzowanych procesów walidacji danych.

Przygotowywanie danych pozwala na zidentyfikowanie nietypowych przypadków i niepoprawnych przypadków, zmiennych i wartości danych w aktywnym zbiorze danych oraz na przygotowaniu danych do modelowania.

Korzystanie z procedur przygotowywania danych

Sposób użycia procedur przygotowywania danych zależy od konkretnych potrzeb. Typowy przebieg działań po załadowaniu danych jest następujący:

- **Przygotowanie metadanych.** Przejrzyj zmienne w pliku danych i określ ich ważne wartości, etykiety i poziomy pomiaru. Zidentyfikuj kombinacje wartości zmiennych, które są niemożliwe, ale często błędnie kodowane. Zdefiniuj reguły walidacyjne na podstawie tych informacji. To zadanie może być czasochłonne, ale warto je wykonać, jeśli planuje się często walidację plików danych z podobnymi atrybutami.
- **Walidacja danych.** Wykonaj podstawowe kontrole oraz użyj zdefiniowanych reguł walidacyjnych w celu zidentyfikowania niepoprawnych obserwacji, zmiennych i wartości danych. Po znalezieniu niepoprawnych danych należy zbadać i usunąć przyczynę problemu. Może to wymagać powtórnego przygotowania metadanych.
- **Przygotowanie modelu.** Użyj technik zautomatyzowanego przygotowywania danych w celu uzyskania takich przekształceń pierwotnych zmiennych, które korzystnie wpłyną na tworzenie modelu. Zidentyfikuj potencjalne wartości odstające statystycznie, które mogą powodować problemy w wielu modelach predykcyjnych. Niektóre wartości odstające są wynikiem obecności niewykrytych wcześniej nieprawidłowych wartości zmiennych. Może to wymagać powtórnego przygotowania metadanych.

Gdy plik danych jest „czysty”, można przystąpić do tworzenia modeli przy użyciu innych modułów dodatkowych.

Reguły walidacyjne

Reguła jest używana do określenia ważności obserwacji. Istnieją dwa typy reguł walidacyjnych:

- **Reguły jednej zmiennej.** Reguły jednej zmiennej składają się ze stałego zbioru kontroli, które mają zastosowanie do jednej zmiennej, na przykład kontroli, czy wartości mieszczą się w poprawnym zakresie. W przypadku reguł jednej zmiennej poprawne wartości mogą być wyrażane jako zakres lub lista dopuszczalnych wartości.
- **Reguły wielu zmiennych.** Reguły wielu zmiennych to reguły wprowadzone przez użytkownika, które mogą być stosowane na jednej zmiennej lub kombinacji zmiennych. Reguły wielu zmiennych są definiowane przez wyrażenie logiczne, które oznacza niepoprawne wartości.

Reguły walidacji są zapisywane w słowniku danych w pliku danych. Dzięki temu raz zdefiniowanej reguły można wielokrotnie używać.

Ładowanie predefiniowanych reguł walidacyjnych

Można szybko uzyskać zestaw gotowych do użytku reguł walidacyjnych, ładując predefiniowane reguły z zewnętrznego pliku danych wchodzącego w skład instalacji.

Ładowanie predefiniowanych reguł walidacyjnych

1. Wybierz z menu następującą opcję:

   Dane > **Walidacja** > **Załaduj predefiniowane reguły...**

Alternatywnym sposobem jest użycie kreatora Kopiowanie właściwości danych do załadowania reguł z dowolnego pliku danych.

Definiowanie reguł walidacyjnych

Okno dialogowe Definiuj reguły walidacyjne umożliwia tworzenie i przeglądanie reguł walidacyjnych pojedynczej zmiennej i reguł walidacyjnych wielu zmiennych.

Aby utworzyć i przeglądać reguły walidacyjne

1. Z menu wybierz:

   Dane > **Walidacja** > **Definiuj reguły...**

   Okno dialogowe jest zapełniane regułami walidacyjnymi pojedynczej zmiennej i wielu zmiennych odczytywanymi ze słownika danych. Gdy nie ma żadnych reguł, automatycznie tworzona jest nowa reguła zastępcza, którą można zmodyfikować do własnych celów.
2. Wybierz poszczególne reguły na kartach Reguły pojedynczej zmiennej i Reguły wielu zmiennych, aby wyświetlić i zmodyfikować ich właściwości.

Definiowanie reguł pojedynczej zmiennej

Karta Reguły pojedynczej zmiennej umożliwia tworzenie, przeglądanie i modyfikowanie reguł walidacyjnych pojedynczej zmiennej.

Reguły. Na liście wymienione są reguły walidacyjne pojedynczej zmiennej uporządkowane według nazw oraz typów zmiennych, do których reguły mogą być stosowane. Po otwarciu okna dialogowego wyświetlane są reguły zdefiniowane w słowniku danych lub, jeśli obecnie nie ma zdefiniowanych reguł, reguła zastępcza o nazwie „RegułaPojedynczejZmiennej 1”. Pod listą Reguły wyświetlane są następujące przyciski:

- **Nowa.** Dodaje nowy wpis na końcu listy Reguły. Reguła jest wybierana i otrzymuje nazwę „RegułaPojedynczejZmiennej *n*”, gdzie *n* jest liczbą całkowitą zapewniającą unikalność nazwy reguły wśród wszystkich reguł pojedynczej zmiennej i wielu zmiennych.
- **Duplikuj.** Dodaje kopię wybranej reguły na końcu listy reguł. Nazwa reguły jest modyfikowana w taki sposób, aby była unikalna wsród reguł pojedynczej zmiennej i wielu zmiennych. Na przykład po powieleniu reguły „RegułaPojedynczejZmiennej 1” nazwa pierwszej kopii będzie miała poastać „Kopia RegułaPojedynczejZmiennej 1”, nazwa drugiej kopii będzie miała postać „Kopia (2) RegułaPojedynczejZmiennej 1” i tak dalej.
- **Usuń.** Usuwa wybraną regułę.

Definicja reguły. Te elementy sterujące umożliwiają przeglądanie i ustawianie właściwości wybranej reguły.

- **Name.** Nazwa reguły musi być unikalna wsród reguł pojedynczej zmiennej i wielu zmiennych.
- **Typ.** Jest to typ zmiennych, do których reguła może być stosowana. Wybierz jeden z typów: **Liczbowa**, **Łańcuchowa** albo **Data i czas**.
- **Format.** Umożliwia wybranie formatu daty dla reguł, które mogą być stosowane do zmiennych typu Data.
- **Ważne wartości.** Można określić ważne wartości jako przedział albo listę.

Definicja przedziału

Elementy sterujące do definiowania przedziału umożliwiają określenie przedziału, w którym wartości są ważne. Wartości spoza przedziału są oznaczane jako nieważne.

Aby określić przedział, wprowadź wartość minimalną i/lub maksymalną. Pola wyboru umożliwiają oznaczanie w przedziale wartości bez etykiet i wartości niecałkowitych.

Definicja listy

Elementy sterujące do definiowania listy umożliwiają zdefiniowanie listy ważnych wartości. Wartości spoza listy są oznaczane jako nieważne.

Wprowadź wartości w tabeli. Pola wyboru określają, czy dana obserwacja ma znaczenie przy porównywaniu wartości łańcuchowych z listą wartości dopuszczalnych.

- **Dopuszczaj braki danych użytkownika.** Określa, czy braki danych użytkownika mają być oznaczane jako nieważne.
- **Zezwalaj na systemowe braki danych.** Określa, czy systemowe braki danych mają być oznaczane jako nieważne. Nie ma zastosowania do reguł operujących na łańcuchach.
- **Zezwalaj na wartości puste.** Określa, czy puste (tj. niezawierające żadnych znaków) wartości łańcuchowe mają być oznaczane jako nieważne. Nie ma zastosowania do reguł operujących na danych innych niż łańcuchowe.

Definiowanie reguł wielu zmiennych

Karta Reguły wielu zmiennych umożliwia tworzenie, przeglądanie i modyfikowanie reguł walidacyjnych wielu zmiennych.

Reguły. Na liście wymienione są reguły walidacyjne wielu zmiennych pogrupowane według nazw. Po otwarciu okna dialogowego wyświetlana jest jedna reguła zastępcza o nazwie „RegułaZmiennychKrzyżowych 1”. Pod listą Reguły wyświetlane są następujące przyciski:

- **Nowa.** Dodaje nowy wpis na końcu listy Reguły. Reguła jest wybierana i otrzymuje nazwę „RegułaZmiennychKrzyżowych *n*”, gdzie *n* jest liczbą całkowitą zapewniającą unikalność nazwy reguły wśród wszystkich reguł pojedynczej zmiennej i wielu zmiennych.
- **Duplikuj.** Dodaje kopię wybranej reguły na końcu listy reguł. Nazwa reguły jest modyfikowana w taki sposób, aby była unikalna wsród reguł pojedynczej zmiennej i wielu zmiennych. Na przykład po powieleniu reguły „RegułaZmiennychKrzyżowych 1” nazwa pierwszej kopii będzie miała poastać „Kopia RegułaZmiennychKrzyżowych 1”, nazwa drugiej kopii będzie miała postać „Kopia (2) RegułaZmiennychKrzyżowych 1” i tak dalej.
- **Usuń.** Usuwa wybraną regułę.

Definicja reguły. Te elementy sterujące umożliwiają przeglądanie i ustawianie właściwości wybranej reguły.

- **Name.** Nazwa reguły musi być unikalna wsród reguł pojedynczej zmiennej i wielu zmiennych.
- **Wyrażenie logiczne.** Jest to zasadniczo definicja reguły. Wyrażenie należy skonstruować w taki sposób, aby w przypadku nieważnej obserwacji dawało wynik 1.

Tworzenie wyrażeń

1. Aby utworzyć wyrażenie, można w pole Wyrażenie wkleić składowe lub wpisać je tam bezpośrednio.

- Można wkleić zmienne lub często używane zmienne zostaną wklejone poprzez wybranie grupy z listy grupy funkcji i klikając dwukrotnie funkcję lub zmienną na liście funkcji i zmiennych specjalnych (lub wybrać funkcję lub zmienną i klikając przycisk **Wstaw**). Wprowadź wszystkie parametry ze znakiem zapytania (dotyczy tylko funkcji). Grupa funkcji o nazwie **Wszystkie** to lista wszystkich dostępnych funkcji i zmiennych systemowych. Krótki opis aktualnie wybranej funkcji lub zmiennej jest wyświetlony w zarezerwowanym obszarze ona dialogowego.
- Stałe łańcuchowe muszą być ujęte w cudzysłów lub apostrofy.
- Jeśli wartości zawierają dziesiętne, do wskazania dziesiętnej należy użyć kropki (.).

Sprawdzenie poprawności danych

Okno dialogowe Walidacja danych umożliwia wykrywanie podejrzanych lub nieważnych obserwacji, zmiennych i wartości danych w aktywnym zbiorze danych.

Przykład. Analityk danych musi co miesiąc sporządzać raport o zadowoleniu klientów. Jakość danych wejściowych z każdego miesiąca musi być kontrolowana w celu wykrycia niekompletnych danych identyfikujących klientów, wartości zmiennych będących poza dozwolonym zakresem i kombinacji wartości zmiennych, które często wprowadzane są w wyniku pomyłki. W oknie dialogowym Walidacja danych analityk może określić zmienne, które jednoznacznie identyfikują klienta, zdefiniować reguły pojedynczej zmiennej wychwytujące wartości spoza dozwolonego zakresu i zdefiniować reguły wielu zmiennych wychwytujących niedozwolone kombinacje. Procedura zwraca raport z listą problematycznych obserwacji i zmiennych. Ponadto dane z każdego miesiąca zawierają te same elementy, dlatego analityk będzie mógł stosować te same reguły do danych z następnych miesięcy.

Statystyki. Procedura generuje listę zmiennych, obserwacji i wartości danych, które nie przeszły pomyślnie różnych kontroli, liczby naruszeń reguł pojedynczej zmiennej i wielu zmiennych, a także proste opisowe podsumowania analizowanych zmiennych.

Wagi. Procedura ignoruje specyfikację zmiennej ważącej i traktuje tę zmienną tak, jak pozostałe analizowane zmienne.

Przeprowadzanie walidacji danych

1. Wybierz z menu następującą opcję:

   Dane > **Walidacja** > **Walidacja danych...**
2. Wybierz co najmniej jedną analizowaną zmienną do walidacji przez zastosowanie prostych kontroli zmiennych lub reguł walidacji pojedynczej zmiennej.

   Zamiast tego można:
3. Kliknąć kartę **Reguły wielu zmiennych** i zastosować jedną lub więcej reguł wielu zmiennych.

Opcjonalnie można wykonać następujące kroki:

- Wybrać co najmniej jedną zmienną identyfikującą, aby wykryć ewentualne zduplikowane lub niekompletne dane identyfikacyjne. Zmienne identyfikujące obserwacje są także używane jako etykiety wyników generowanych obserwacjami. Jeśli określone są dwie lub większa liczba zmiennych identyfikujących obserwacje, to jako identyfikator obserwacji traktowana jest kombinacja ich wartości.

Zmienne z nieznanym poziomem pomiaru

Alert poziomu pomiaru wyświetla się, gdy poziom pomiaru dla jednej lub większej ilości zmiennych w zbiorze danych jest nieznany. Ponieważ poziom pomiaru wpływa na wyliczenie wyników dla tej procedury, wszystkie zmienne muszą mieć zdefiniowany poziom pomiaru.

Skanowanie danych. Odczytuje dane w aktywnym zbiorze danych i przypisuje domyślny poziom pomiaru do wszystkich zmiennych, które mają aktualnie nieznany poziom pomiaru. Jeśli zbiór danych jest duży, może to zająć trochę czasu.

Przypisz ręcznie. Otwiera okno dialogowe, które zestawia wszystkie zmienne z nieznanym poziomem pomiaru. Można użyć tego okna dialogowego do przypisania poziomu pomiaru do tych zmiennych. Można również przypisać poziom pomiaru w Widoku zmiennych Edytora danych.

Ponieważ poziom pomiaru jest ważny dla tej procedury, nie można wejść do tego okna dialogowego w celu uruchomienia tej procedury, dopóki wszystkie zmienne nie będą miały zdefiniowanego poziomu pomiaru.

Walidacja danych: Sprawdzenia podstawowe

Karta Sprawdzenia podstawowe umożliwia wybranie podstawowych sprawdzeń analizowanych zmiennych, identyfikatorów obserwacji i całych obserwacji.

Zmienne analizowane. Jeśli na karcie Zmienne wybrano jakiekolwiek zmienne analizowane, można wybrać dowolne z następujących sprawdzeń ich ważności. Sprawdzenia można włączać i wyłączać za pomocą pól wyboru.

- **Maksymalny procent braków danych.** Zgłasza analizowane zmienne, w których odsetek braków danych przekracza podaną wartość. Podana wartość musi być liczbą dodatnią nie większą niż 100.
- **Maksymalny procent obserwacji w pojedynczej kategorii.** Jeśli którakolwiek z analizowanych zmiennych jest jakościowa, ta opcja zgłasza jakościowe analizowane zmienne z odsetkiem obserwacji należących do jednej niebrakującej kategorii większym od podanej wartości. Podana wartość musi być liczbą dodatnią nie większą niż 100. Odsetek obliczany jest na podstawie obserwacji z niebrakującymi wartościami zmiennej.
- **Maksymalny procent kategorii z liczebnością równą 1.** Jeśli którakolwiek z analizowanych zmiennych jest jakościowa, ta opcja zgłasza jakościowe analizowane zmienne z odsetkiem kategorii zmiennej zawierających tylko jedną obserwację większym od podanej wartości. Podana wartość musi być liczbą dodatnią nie większą niż 100.
- **Minimalny współczynnik zmienności.** Jeśli którakolwiek z analizowanych zmiennych jest ilościowa, ta opcja zgłasza ilościowe analizowane zmienne, których współczynnik zmienności ma wartość bezwzględną mniejszą od podanej wartości. Ta opcja ma zastosowanie tylko do zmiennych o niezerowej średniej. Wartość musi być liczbą całkowitą nieujemną. Określenie wartości 0 powoduje wyłączenie sprawdzania współczynnika zmienności.
- **Minimalne odchylenie standardowe.** Jeśli którakolwiek z analizowanych zmiennych jest ilościowa, ta opcja zgłasza ilościowe analizowane zmienne, których odchylenie standardowe jest mniejsze od podanej wartości. Wartość musi być liczbą całkowitą nieujemną. Określenie wartości 0 powoduje wyłączenie sprawdzania odchylenia standardowego.

Identyfikatory obserwacji. Jeśli na karcie Zmienne wybrano jakiekolwiek zmienne identyfikujące obserwacje, można wybrać dowolne z następujących sprawdzeń ich ważności.

- **Flaga niekompletnych identyfikatorów.** Ta opcja zgłasza obserwacje z niekompletnymi identyfikatorami obserwacji. Identyfikator konkretnej obserwacji uznawany jest za niekompletny, jeśli wartość którejkolwiek ze zmiennych identyfikujących jest pusta lub nie istnieje.
- **Flaga zduplikowanych identyfikatorów.** Ta opcja zgłasza obserwacje ze zduplikowanymi identyfikatorami obserwacji. Niekompletne identyfikatory są wykluczane ze zbioru możliwych duplikatów.

Flaga pustych obserwacji. Ta opcja zgłasza obserwacje, w których wszystkie zmienne są puste. Na potrzeby identyfikacji pustych obserwacji można używać wszystkich zmiennych w pliku (z wyjątkiem zmiennych identyfikujących) lub tylko analizowanych zmiennych zdefiniowanych na karcie Zmienne.

Walidacja danych: Reguły pojedynczej zmiennej

Na karcie Reguły pojedynczej zmiennej wyświetlane są dostępne reguły walidacyjne pojedynczej zmiennej i możliwe jest zastosowanie tych reguł do analizowanych zmiennych. Aby zdefiniować dodatkowe reguły pojedynczej zmiennej, kliknij przycisk **Definiuj reguły**. Więcej informacji zawiera temat “Definiowanie reguł pojedynczej zmiennej” na stronie 2.

Zmienne analizowane. Na liście wyświetlane są analizowane zmienne, podsumowanie ich rozkładów oraz liczby reguł zastosowanych do każdej zmiennej. Należy zwrócić uwagę, że w podsumowaniach nie są uwzględniane braki danych użytkownika ani systemowe braki danych. Lista rozwijana wyświetlania steruje wyświetlaniem zmiennych. Można wybrać **Wszystkie zmienne**, **Zmienne numeryczne**, **Zmienne łańcuchowe** i **Zmienne daty i czasu**.

Reguły. Aby zastosować reguły do analizowanych zmiennych, wybierz jedną lub więcej zmiennych i na liście reguł zaznacz wszystkie reguły, które mają być zastosowane. Na liście reguł wyświetlane są tylko reguły odpowiednie dla wybranych analizowanych zmiennych. Na przykład, jeśli wybrane są analizowane zmienne liczbowe, wyświetlane są tylko reguły liczbowe; jeśli wybrana jest zmienna łańcuchowa, wyświetlane są tylko reguły łańcuchowe. Jeśli nie są wybrane żadne analizowane zmienne lub wybrane są zmienne różnych typów, nie są wyświetlane żadne reguły.

Rozkłady zmiennej. Podsumowania rozkładów widoczne na liście Zmienne analizowane mogą być oparte na wszystkich obserwacjach lub na przeglądzie pierwszych *n* obserwacji, zgodnie z wpisem w polu tekstowym Obserwacje. Kliknięcie przycisku **Skanuj ponownie** powoduje zaktualizowanie podsumowań rozkładu.

Walidacja danych: Reguły wielu zmiennych

Na karcie Reguły wielu zmiennych wyświetlane są dostępne reguły walidacyjne wielu zmiennych i możliwe jest zastosowanie tych reguł do analizowanych zmiennych. Aby zdefiniować dodatkowe reguły wielu zmiennych, kliknij przycisk **Definiuj reguły**. Więcej informacji zawiera temat "Definiowanie reguł wielu zmiennych" na stronie 3.

Walidacja danych: Raport

Raport dla poszczególnych obserwacji. Jeśli zastosowano jakiekolwiek reguły walidacji jednej zmiennej lub wielu zmiennych, można wygenerować raport z listą naruszeń reguł walidacyjnych przez poszczególne obserwacje.

- **Minimalna liczba naruszeń.** Ta opcja określa minimalną liczbę naruszeń reguły, aby obserwacja została uwzględniona w raporcie. Określ dodatnią liczbę całkowitą.
- **Maksymalna liczba obserwacji.** Ta opcja określa maksymalną liczbę obserwacji uwzględnianych w raporcie z obserwacji. Określ liczbę dodatnią nie większą niż 1000.

Reguły walidacyjne pojedynczej zmiennej. Jeśli zastosowano jakiekolwiek reguły walidacyjne dotyczące jednej zmiennej, można wybrać sposób wyświetlania wyników lub w ogóle zrezygnować z ich wyświetlania.

- **Podsumuj naruszenia dla analizowanej zmiennej.** Ta opcja powoduje, że dla każdej analizowanej zmiennej podawane są wszystkie naruszone reguły walidacyjne pojedynczej zmiennej i liczba wartości, które naruszyły każdą z reguł. Podawana jest także łączna liczba naruszeń reguł pojedynczej zmiennej dla każdej zmiennej.
- **Podsumuj naruszenia dla reguły.** Ta opcja powoduje, że dla każdej reguły walidacyjnej pojedynczej zmiennej podawane są zmienne, które naruszyły regułę, oraz liczba nieważnych wartości przypadających na każdą zmienną. Podawana jest także łączna liczba wartości, które naruszyły każdą regułę wielu zmiennych.

Przedstaw statystyki opisowe dla analizowanych zmiennych. Ta opcja umożliwia uzyskanie statystyk opisowych analizowanych zmiennych. Dla każdej zmiennej jakościowej generowana jest tabela częstości. Dla zmiennych ilościowych generowana jest tabela ze statystykami podsumowującymi, obejmującymi średnią, odchylenie standardowe, minimum i maksimum.

Przesuń obserwacje naruszające reguły walidacyjne na początek aktualnego zbioru danych. Ta opcja przesuwa obserwacje z naruszeniami reguł walidacyjnych pojedynczej zmiennej lub wielu zmiennych na początek aktywnego zbioru danych, aby łatwo można było je odszukać.

Walidacja danych: Zapisz

Karta Zapisz umożliwia zapisanie w aktywnym zbiorze danych zmiennych rejestrujących naruszenia reguł walidacji.

Charakteryzowane zmienne. Są to poszczególne zmienne, które można zapisać. Zaznacz pole wyboru zmiennej, aby ją zapisać. Wprowadzone są domyślne nazwy zmiennych; można je edytować.

- **Wskaźnik pustej obserwacji.** Do pustych obserwacji przypisywana jest wartość 1. W przypadku wszystkich pozostałych obserwacji kodowana jest wartość 0. Wartości zmiennej odzwierciedlają zasięg podany na karcie kontroli podstawowych.
- **Powtórzony identyfikator grupy.** Obserwacjom o tym samym identyfikatorze obserwacji (ale nie obserwacjom z niekompletnymi identyfikatorami) przypisywany jest ten sam numer grupy. Obserwacjom z unikalnymi lub niekompletnymi identyfikatorami przypisywany jest kod 0.

- **Niekompletny identyfikator.** Obserwacje z pustymi lub niekompletnymi identyfikatorami mają przypisywaną wartość 1. W przypadku wszystkich pozostałych obserwacji kodowana jest wartość 0.
- **Naruszenia reguły walidacyjnej.** Jest to łączna liczba naruszeń reguł walidacyjnych dotyczących jednej zmiennej i wielu zmiennych, liczona według obserwacji.

Zastąp już istniejące charakteryzowane zmienne. Zmienne zapisane w pliku danych muszą mieć unikalne nazwy lub zastępować zmienne o tych samych nazwach.

Zapisz zmienne wskaźnikowe. Ta opcja umożliwia zapisanie kompletnych informacji o naruszeniach reguł walidacyjnych. Każda zmienna odpowiada zastosowaniu reguły walidacyjnej i ma wartość 1, jeśli obserwacja narusza regułę, albo wartość 0, jeśli nie narusza.

Automatyczne przygotowywanie danych

Przygotowywanie danych do analizy jest jednym z najbardziej istotnych kroków w każdym projekcie — i również jednym z najbardziej czasochłonnych. Automatyczne przygotowanie danych (Automated Data Preparation — ADP) ma za zadanie analizę danych i identyfikację stałych, klasyfikację pól (zmiennych), które są problematyczne lub mają małe prawdopodobieństwo bycia użytecznymi, w razie potrzeby obliczanie nowych atrybutów i zwiększanie wydajności poprzez wykorzystywanie inteligentnych technik klasyfikowania. Można używać tego algorytmu w sposób **w pełni automatyczny**, pozwalając mu na wybór i zastosowanie stałych, lub korzystać z niego w sposób **interaktywny**, przeglądając zmiany przed ich dokonaniem i zaakceptować je lub odrzucać.

Użycie funkcji automatycznego przygotowania danych (ADP) umożliwia przygotowanie danych do szybkiego i łatwego budowania modelu, bez konieczności uzyskiwania wiedzy na temat użytych koncepcji statystycznych. Budowa i ocena modeli będzie odbywać się szybciej; ponadto, korzystanie z automatycznego przygotowywania danych (ADP) zwiększa elastyczność procesów automatycznego modelowania.

Uwaga: kiedy proces automatycznego przygotowywania danych przygotowuje zmienną do analizy, tworzona jest nowa zmienna zawierająca korekty lub transformacje, zamiast zastępowania istniejących wartości i właściwości starej zmiennej. Stara zmienna nie jest używana w dalszej analizie; jej rola jest ustawiona na Brak. Należy również pamiętać, że informacje na temat braków danych zdefiniowanych przez użytkownika nie są przenoszone do nowo utworzonych zmiennych, a braki danych w nowej zmiennej są systemowymi brakami danych.

Przykład. Firma ubezpieczeniowa o ograniczonych środkach na sprawdzenie roszczeń chce zbudować model do flagowania podejrzanych, potencjalnie oszukańczych roszczeń. Przed utworzeniem modelu dane będą przygotowane do modelowania przy użyciu automatycznego przygotowywania danych. Ponieważ firma chce mieć możliwość przejrzenia zaproponowanych transformacji przed ich zastosowaniem, użyte zostanie automatyczne przygotowywanie danych w trybie interaktywnym.

Grupa z branży motoryzacyjnej śledzi sprzedaż różnych pojazdów osobowych. Podejmując próbę zidentyfikowania najbardziej i najmniej rentownych modeli, chcą ustalić relacje pomiędzy sprzedażą pojazdów a charakterystykami pojazdów. Automatyczne przygotowywanie danych umożliwia przygotowanie danych do analizy, a utworzenie modeli z użyciem danych „przed” i „po” przygotowaniu pozwoli zobaczyć różnice w wynikach.

Jaki jest cel? Automatyczne przygotowywanie danych rekomenduje kroki przygotowania danych, które będą wpływały na szybkość, z jaką inne algorytmy mogą budować modele i które ulepszą jakość predykcji tych modeli. Mogą one zawierać przekształcenia istniejących i tworzenie nowych predyktorów, a także ich wybór. Zmienna przewidywana również może być przekształcona. Można określić priorytety budowania modelu, na jakich proces przygotowywania danych powinien się skoncentrować.

- **Zrównoważenie szybkości i dokładności.** Ta opcja umożliwia przygotowanie danych, tak aby nadać jednakowy priorytet szybkości przetwarzania danych przez algorytmy budowania modelu oraz dokładności predykcji.
- **Optymalizacja dla szybkości.** Ta opcja umożliwia przygotowanie danych, tak aby nadać priorytet szybkości przetwarzania danych przez algorytmy budowania modelu. Opcję tę należy wybrać w przypadku pracy z dużymi zbiorami danych lub poszukiwania szybkiej odpowiedzi.

- **Optymalizacja dla dokładności.** Ta opcja umożliwia przygotowywanie danych, taka aby nadać priorytet dokładności predykcji tworzonych przez algorytmy budowania modelu.
- **Analiza użytkownika.** Opcję tę należy wybrać, aby ręcznie zmienić algorytm na karcie Ustawienia. Jeśli w późniejszym czasie na karcie Ustawienia dokonane zostaną zmiany opcji, które są niekompatybilne z jednym z pozostałych celów, należy pamiętać, że to ustawienie jest zaznaczane automatycznie.

Otrzymywanie automatycznego przygotowania danych

Z menu wybierz:

1. Z menu wybierz:

   Przekształcenia > **Przygotuj dane do modelowania** > **Automatyczne...**
2. Kliknij opcję **Uruchom**.

Opcjonalnie można wykonać następujące czynności:

- Określ cel w zakładce Cel.
- Określ przypisania zmiennych w zakładce Zmienne.
- Określ ustawienia zaawansowane w zakładce Zaawansowane.

Otrzymywanie interaktywnego przygotowania danych

1. Z menu wybierz:

   Przekształcenia > **Przygotuj dane do modelowania** > **Interaktywne...**
2. Kliknij opcję **Analiza** na pasku narzędzi u góry okna dialogowego.
3. Kliknij zakładkę Analiza i przejrzyj sugerowane kroki przygotowania danych.
4. Jeśli kroki są zgodne z oczekiwaniami, kliknij opcję **Uruchom**. W przeciwnym wypadku kliknij opcję **Wyczyść analizę**, zmień dowolne ustawienia i kliknij opcję **Analiza**.

Opcjonalnie można wykonać następujące czynności:

- Określ cel w zakładce Cel.
- Określ przypisania zmiennych w zakładce Zmienne.
- Określ ustawienia zaawansowane w zakładce Zaawansowane.
- Aby zapisać sugerowane kroki przygotowania danych w pliku XML, kliknij opcję **Zapisz XML**.

Karta Zmienne

Karta Zmienne określa, które zmienne powinny zostać przygotowane na potrzeby dalszej analizy.

Użyj wstępnie zdefiniowanych ról. Ta opcja wykorzystuje istniejące informacje i zmiennych. Jeśli dostępna jest pojedyncza zmienna z rolą zmiennej przewidywanej, zostanie użyta jako zmienna przewidywana; w przeciwnym razie nie będzie użyta żadna zmienna przewidywana. Wszystkie zmienne z predefiniowaną rolą zmiennych wejściowych, będą używane jako zmienne wejściowe. Wymagana jest co najmniej jedna zmienna wejściowa. .

Użyj niestandardowych przypisań. Po zastąpieniu ról zmiennych poprzez przeniesienie zmiennych z ich list domyślnych, okno dialogowe automatycznie przełącza się na tę opcję. W przypadku niestandardowych przypisań zmiennych należy określić następujące zmienne:

- **Wartość docelowa (opcjonalnie).** Jeśli planowane jest utworzenie modeli, które wymagają zmiennej przewidywanej, należy wybrać zmienną przewidywaną. Jest to podobne do ustawienia roli zmiennej na zmienną przewidywaną.
- **Zmienne wejściowe.** Należy wybrać co najmniej jedną zmienną wejściową. Jest to podobne do ustawienia roli zmiennej na zmienną wejściową.

Karta Ustawienia

Zakładka Ustawienia zawiera wiele różnych grup ustawień, które można zmieniać w celu precyzyjnego określenia sposobu przetwarzania danych użytkownika przez algorytm. W przypadku wprowadzenia zmian w ustawieniach domyślnych, które są niezgodne z innymi celami, zakładka Cel zostanie automatycznie zaktualizowana do zaznaczenia opcji **Analiza niestandardowa**.

Przygotowanie daty i czasu

Wiele algorytmów modelowania nie jest w stanie bezpośrednio obsługiwać danych daty i czasu; te ustawienia umożliwiają wyliczenie nowych danych czasu trwania, które mogą być użyte jako dane wejściowe modelu, na podstawie informacji o dacie i czasie dostępnych w istniejących danych. Zmienne zawierające datę i czas muszą zostać wstępnie zdefiniowane z użyciem typów składowania data lub czas. Oryginalne zmienne daty i czasu nie będą zalecane przez proces automatycznego przygotowywania danych jako dane wejściowe modelu.

Przygotuj datę i czas do modelowania. Usunięcie zaznaczenia tej opcji wyłączy pozostałe elementy sterujące w obszarze Przygotowanie daty i czasu podczas dokonywania wyboru ustawień.

Wylicz czas jaki upłynął od daty odniesienia. Ta opcja wyznacza liczbę lat/miesięcy/dni od daty odniesienia dla każdej zmiennej zawierającej daty.

- **Data odniesienia.** Umożliwia określenie daty, od której obliczany będzie czas trwania, z odniesieniem do informacji na temat daty dostępnych w danych wejściowych. Wybranie opcji **Dzisiejsza data** oznacza, że podczas wykonywania automatycznego przygotowywania danych zawsze używana będzie bieżąca data systemowa. Aby użyć konkretnej daty, należy zaznaczyć opcję **Ustalona data** i wprowadzić wymaganą datę.
- **Jednostki czasu trwania.** Należy określić, czy proces automatycznego przygotowywania danych będzie automatycznie dobierał jednostkę czasu trwania, lub wybrać opcję **Ustalone jednostki** i zaznaczyć Lata, Miesiące lub Dni.

Wylicz czas jaki upłynął od czasu odniesienia. Ta opcja powoduje wyznaczenie liczby godzin/minut/sekund od czasu odniesienia dla każdej zmiennej zawierającej dane o czasie.

- **Czas odniesienia.** Umożliwia określenie czasu, od którego obliczany będzie czas trwania, z odniesieniem do informacji na temat czasu dostępnych w danych wejściowych. Wybranie opcji **Bieżący czas** oznacza, że podczas wykonywania automatycznego przygotowywania danych zawsze używany będzie bieżący czas systemowy. Aby użyć konkretnego czasu, należy wybrać opcję **Ustalony czas** i wprowadzić odpowiednie szczegóły.
- **Jednostki czasu trwania.** Należy określić, czy proces automatycznego przygotowywania danych będzie automatycznie dobierał jednostkę czasu trwania, lub wybrać opcję **Ustalone jednostki** i zaznaczyć Godziny, Minuty lub Sekundy.

Wyodrębnij cykliczne elementy czasu. Te ustawienia umożliwiają podzielenie pojedynczej zmiennej daty lub czasu na jedną lub więcej zmiennych. Przykładowo, jeśli zaznaczone zostaną wszystkie trzy pola wyboru dla daty, wejściowa zmienna daty "1954-05-23" zostanie podzielona na trzy zmienne: 1954, 5 i23, a dla każdej z nich zastosowany zostanie przedrostek zdefiniowany w panelu **Nazwy zmiennych**, a oryginalna zmienna daty zostanie zignorowana.

- **Pobierz z daty.** Dla dowolnych danych wejściowych daty należy określić lata, miesiące lub dni do wyodrębnienia lub dowolną ich kombinację.
- **Pobierz z czasu.** Dla dowolnych danych wejściowych czasu należy określić godziny, minuty lub sekundy do wyodrębnienia lub dowolną ich kombinację.

Statystyki wykluczonych zmiennych

Słaba jakość danych może wpływać na dokładność predykcji; dlatego można określić akceptowalny poziom jakości dla wejściowych predykcji. Wszystkie zmienne, które są stałe lub mają 100% braków danych, zostają automatycznie wykluczone.

Wyklucz zmienne wejściowe niskiej jakości. Usunięcie zaznaczenia tej opcji wyłączy pozostałe elementy sterujące w obszarze Statystyki wykluczonych zmiennych podczas dokonywania wyboru ustawień.

Wyklucz zmienne wejściowe o zbyt dużej liczbie braków. Zmienne, w których liczba braków danych przekracza określoną wartość procentową, zostają usunięte z dalszej analizy. Określenie wartości większej niż lub równej 0 jest równoznaczne z usunięciem zaznaczenia tej opcji, a wartości mniejszej niż lub równej 100 spowoduje, że zmienne ze wszystkimi brakami danych będą automatycznie wykluczane. Domyślną wartością jest 50.

Wyklucz zmienne nominalne o zbyt dużej liczbie unikatowych kategorii. Zmienne nominalne z liczbą kategorii większą od podanej będą wykluczone z dalszej analizy. Podaj dodatnią liczbę całkowitą. Wartość domyślna to: 100. Ta opcja jest przydatna do automatycznego usuwania z modelowania zmiennych zawierających informacje unikalne dla rekordu, takie jak identyfikator, adres lub nazwa.

Wyklucz zmienne kategorialne o zbyt dużej liczbie wartości w jednej kategorii. Zmienne porządkowe i nominalne z kategorią, która zawiera więcej rekordów od określonej wartości procentowej, są usuwane z dalszej analizy. Określenie wartości większej niż lub równej 0 jest równoznaczne z usunięciem zaznaczenia tej opcji, a wartości mniejszej niż lub równej 100 spowoduje, że zmienne o stałych wartościach będą automatycznie wykluczane. Domyślną wartością jest 95.

Korekta pomiaru

Korekta poziomu pomiaru. Usunięcie zaznaczenia tej opcji wyłączy pozostałe elementy sterujące przeznaczone do korygowania pomiarów, a pozostałe wybory pozostaną.

Poziom pomiaru. Określ, czy poziom pomiaru ze zmiennymi ciągłymi ze zbyt małą ilością wartości („za mało”) może zostać skorygowany do porządkowego, a zmienne porządkowe ze zbyt dużą ilością wartości („zbyt wiele”) mogą zostać skorygowane do ciągłych.

- **Maksymalna liczba wartości dla zmiennych porządkowych.** Zmienne porządkowe z liczbą kategorii większą niż podana liczba są uznawane za zmienne ciągłe. Podaj dodatnią liczbę całkowitą. Wartością domyślną jest 10. Ta wartość musi być większa od minimalnej liczby wartości dla zmiennych ilościowych lub równa tej liczbie.
- **Minimalna liczba wartości dla zmiennych ilościowych.** Zmienne ciągłe z liczbą kategorii większą niż podana liczba unikatowych wartości są uznawane za zmienne porządkowe. Podaj dodatnią liczbę całkowitą. Wartością domyślną jest 5. Ta wartość musi być mniejsza lub równa maksymalnej liczbie wartości dla pól porządkowych.

Poprawa jakości danych

Przygotowanie zmiennych w celu poprawy jakości danych. Usunięcie zaznaczenia tej opcji wyłączy pozostałe elementy sterujące przeznaczone do poprawy jakości danych, a pozostałe wybory pozostaną.

Obsługa wartości odstających. Określ, czy wartości odstające będą zastępowane dla zmiennych wejściowych i zmiennych przewidywanych; jeśli tak, należy podać kryterium odcięcia dla wartości odstających (mierzone w odchyleniach standardowych) oraz metodę zastępowania wartości odstających. Wartości skrajne mogą być zastępowane przez przycięcie (ustawienie wartości odcięcia) lub poprzez ustawienie ich jako braków danych. Każda dowolna wartość odstająca ustawiona na brak danych jest obsługiwana zgodnie z ustawieniami obsługi danych wybranymi poniżej.

Zastąp braki danych. Określ, czy zastępowane będą braki danych zmiennych ciągłych, nominalnych, czy porządkowych.

Zmień kolejność zmiennych nominalnych. Wybranie tej opcji umożliwia rekodowanie zmiennych nominalnych od najmniejszej (najrzadziej występującej) do największej (najczęściej występującej) kategorii. Wartości nowej zmiennej zaczynają się od 0, ponieważ jest to kategoria najrzadsza. Należy zwrócić uwagę na to, że nowa zmienna będzie liczbowa, nawet jeśli zmienna oryginalna jest łańcuchowa. Jeśli na przykład wartości danych zmiennej nominalnej to "A", "A", "A", "B", "C", "C", automatyczne przygotowywanie danych spowoduje rekodowanie "B" na 0, "C" na 1, a "A" na 2.

Przeskalowanie zmiennych

Przeskalowanie zmiennych. Usunięcie zaznaczenia tej opcji wyłączy pozostałe elementy sterujące przeznaczone do przeskalowania zmiennych, a pozostałe wybory pozostaną.

Waga analizy. Ta zmienna zawiera wagi analizy (regresji lub doboru prób). Wagi analizy są używane w celu uwzględnienia różnic w wariancji po wszystkich poziomach zmiennej przewidywanej. Wybierz zmienną ciągłą.

Zmienne o ciągłych danych wejściowych. Wybranie tej opcji spowoduje normalizację zmiennych o ciągłych danych wejściowych przy użyciu **przekształcenia statystyki z** lub **przekształcenia min./maks.** Przeskalowanie danych wejściowych jest szczególnie użyteczne w przypadku wyboru opcji **Konstruuj predyktory** w ustawieniach wyboru i konstruowania.

- **Przekształcenie statystyki z.** Gdy obserwowana średnia i odchylenie standardowe są używane jako oszacowania parametrów populacji, statystyki *z* są mapowane na odpowiednie wartości rozkładu normalnego z podaną **średnią ostateczną** i **ostatecznym odchyleniem standardowym**. Podaj liczbę dla **średniej ostatecznej** oraz liczbę dodatnią dla **ostatecznego odchylenia standardowego**. Wartości domyślne wynoszą odpowiednio 0 i 1 — odpowiednio do standaryzowanego przeskalowania.
- **Przekształcenie min./maks.** Gdy obserwowane minimum i maksimum są używane jako oszacowania parametrów populacji, zmienne są mapowane na odpowiednie wartości rozkładu jednostajnego z podanym **Minimum** i **Maksimum**. Podaj liczby w taki sposób, aby **Maksimum** było większe niż **Minimum**.

Ciągła zmienna przewidywana. W tym przypadku następuje przekształcenie ciągłej zmiennej przewidywanej z użyciem transformacji Boxa-Coxa w zmienną, która ma w przybliżeniu rozkład normalny oraz podaną **średnią ostateczną** i **ostateczne odchylenie standardowe**. Podaj liczbę dla **średniej ostatecznej** oraz liczbę dodatnią dla **ostatecznego odchylenia standardowego**. Wartości domyślne wynoszą odpowiednio 0 i 1.

Uwaga: jeśli zmienna przewidywana została przekształcona przez ADP, kolejne modele budowane z użyciem przekształconej zmiennej przewidywanej oceniają przekształcone jednostki. Aby możliwe było interpretowanie i korzystanie z wyników, należy przekonwertować wartość predykcyjną z powrotem do oryginalnej skali. Aby uzyskać dodatkowe informacje, patrz temat . Więcej informacji można znaleźć w temacie “Przywrócone oceny” na stronie 19.

Przekształcenia zmiennych

W celu poprawy jakości predykcji danych można przekształcić zmienne wejściowe.

Przekształć zmienną do modelowania. Usunięcie zaznaczenia tej opcji wyłączy pozostałe elementy sterujące przeznaczone do przekształcania zmiennych, a pozostałe wybory pozostaną.

Jakościowe zmienne wejściowe Dostępne są następujące opcje:

- **Połącz małoliczne kategorie w celu zwiększenia związku ze zmienną przewidywaną.** Zaznacz tę opcję, aby uzyskać skromniejszy model poprzez zmniejszenie liczby zmiennych do przetworzenia w powiązaniu z docelową. Podobne kategorie identyfikuje się na podstawie relacji między wejściem a zmienną przewidywaną. Scalane są kategorie, które znacząco się nie różnią (to znaczy takie, których wartość *p* jest większa niż wartość podana). Podaj wartość większą od 0 i mniejszą lub równą 1. Jeśli wszystkie kategorie zostaną scalone w jedną, oryginalne i pochodne wersje tej zmiennej są wykluczane z dalszej analizy, ponieważ nie mają wartości jako predyktora.
- **Przy braku zmiennej przewidywanej połącz małoliczne kategorie na podstawie liczebności.** Jeśli zbiór danych nie zawiera zmiennej przewidywanej, należy wybrać opcję scalenia rozrzuconych kategorii zmiennych porządkowych i nominalnych. Metoda równej częstotliwości jest używana w celu scalenia kategorii z mniejszym niż określony minimalny procent łącznej liczby rekordów. Podaj wartość większą lub równą 0 i mniejszą od 100 lub równą 100. Wartością domyślną jest 10. Scalanie zatrzymuje się, gdy nie istnieją kategorie z mniejszym niż określony minimalny procent obserwacji albo gdy zostały tylko dwie kategorie.

Zmienne o ciągłych danych wejściowych. Jeśli zbiór danych zawiera jakościową zmienną przewidywaną, można skategoryzować ciągłe dane wejściowe z mocnymi powiązaniami, aby poprawić wydajność

przetwarzania. Kategorie są tworzone w oparciu o właściwości „jednorodnych podzbiorów”, które są identyfikowane przez metodę Scheffe z użyciem podanej wartości *p* jako alfa dla wartości krytycznej w celu ustalenia jednorodnych podzbiorów. Podaj wartość większą od 0 i mniejszą lub równą 1. Domyślną wartością jest 0,05. Jeśli w wyniku kategoryzacji tworzona jest pojedyncza kategoria dla konkretnej zmiennej, oryginalne i skategoryzowane wersje zmiennej są wykluczane, ponieważ nie mają wartości jako predyktora.

Uwaga: Kategoryzacja w ADP różni się od kategoryzacji optymalnej. Kategoryzacja optymalna wykorzystuje informacje o entropii w celu przekształcenia zmiennej ciągłej w zmienną jakościową; w takim przypadku wymagane jest posortowanie danych i zapisanie ich wszystkich w pamięci. ADP wykorzystuje jednorodne podzbiory w celu kategoryzacji zmiennych ciągłych, co oznacza, że kategoryzacja ADP nie wymaga sortowania danych i nie zapisuje wszystkich danych w pamięci. Zastosowanie metody jednorodnego podzbioru w celu kategoryzacji zmiennej ciągłej oznacza, że liczba kategorii po kategoryzacji jest zawsze mniejsza lub równa liczbie kategorii w zmiennej przewidywanej.

Wybór i tworzenie

W celu poprawy jakości predykcji danych można tworzyć nowe zmienne oparte na istniejących zmiennych.

Dokonaj wyboru predyktora. Ciągła zmienna wejściowa jest usuwana z analizy, jeśli wartość *p* dla jej korelacji ze zmienną przewidywaną jest większa niż podana wartość *p*.

Konstruuj predyktory. Wybierz tę opcję, aby wyprowadzić nowe funkcje z kombinacji kilku istniejących funkcji. Stare funkcje nie są używane w dalszej analizie. Ta opcja ma zastosowanie tylko do ilościowych zmiennych wejściowych, w których zmienna przewidywana jest ilościowa lub w których nie ma zmiennych przewidywanych.

Nazwy zmiennych

Aby w prosty sposób zidentyfikować nowe i przekształcone predyktory, proces automatycznego przygotowywania danych tworzy i stosuje podstawowe nowe nazwy, przedrostki lub przyrostki. Nazwy te można zmienić, tak aby były bardziej odpowiednie do potrzeb użytkownika i danych, którymi dysponuje.

Zmienne transformowane i konstruowane. Należy określić rozszerzenie nazw, jakie będą stosowane do przekształcanych zmiennych przewidywanych i zmiennych wejściowych.

Ponadto, należy określić nazwę przedrostka, jaki będzie stosowany do dowolnych predyktorów tworzonych za pośrednictwem ustawień tworzenia i wyboru. Nowa nazwa jest tworzona poprzez dołączenie numerycznego przyrostka do trzonu nazwy z przyrostkiem. Format liczbowy zależy od liczby wyznaczanych nowych predyktorów, na przykład:

- Utworzone predyktory od 1 do 9 będą miały następujące nazwy: od feature1 (predyktor1) do feature9 (predyktor9).
- Utworzone predyktory od 10 do 99 będą miały następujące nazwy: od feature01 (predyktor01) do feature99 (predyktor99).
- Utworzone predyktory od 100 do 999 będą miały nazwy: od feature001 (predyktor001) do feature999 (predyktor999) itd.

Dzięki temu utworzone predyktory będą posortowane w sensowny sposób, niezależnie od ich liczby.

Czasy trwania obliczone na podstawie daty i czasu. Należy określić nazwy rozszerzeń, jakie będą stosowane do czasów trwania obliczonych na podstawie daty i czasu.

Elementy cykliczne wyodrębnione z daty i czasu. Należy określić nazwy rozszerzeń, jakie będą stosowane do elementów cyklicznych wyodrębnionych na podstawie daty i czasu.

Stosowanie i zapisywanie przekształceń

Ustawienia dotyczące stosowania i zapisywania przekształceń różnią się nieznacznie w zależności od tego, czy używane są okna dialogowe interaktywnego, czy automatycznego przygotowania danych.

Interaktywne przygotowanie danych — stosowanie ustawień transformacji

Przekształcone dane. Te ustawienia określają miejsce zapisu przekształconych danych.

- **Dodaj nowe zmienne do aktywnego zbioru danych.** Dowolne zmienne utworzone na skutek automatycznego przygotowywania danych są dodawane jako nowe pola do aktywnego zbioru danych. Opcja **Aktualizuj role dla analizowanych zmiennych** spowoduje ustawienie roli na Brak dla dowolnych zmiennych wykluczonych z dalszej analizy przez automatyczne przygotowywanie danych.
- **Utwórz nowy zbiór danych lub plik zawierający przekształcone dane.** Zmienne rekomendowane przez automatyczne przygotowywanie danych są dodawane do nowego zestawu danych lub pliku. Opcja **Dołącz zmienne nieanalizowane** powoduje dodanie z oryginalnego zbioru danych zmiennych, które nie zostały określone na karcie Zmienne, do nowego zbioru danych. Ta opcja jest użyteczna w przypadku przenoszenia zmiennych zawierających informacje, które nie są używane w modelowaniu — np. identyfikator, adres, nazwisko, do nowego zbioru danych.

Automatyczne przygotowanie danych — stosowanie i zapisywanie ustawień

Grupa przekształconych danych jest taka sama, jak w procedurze interaktywnego przygotowywania danych. W automatycznym przygotowaniu danych dostępne są następujące opcje dodatkowe:

Zastosuj przekształcenia. W automatycznym przygotowywaniu danych usunięcie zaznaczenia tej opcji powoduje wyłączenie wszystkich pozostałych elementów sterujących Zastosuj i Zapisz oraz zachowanie dokonanych wyborów.

Zapisz przekształcenia jako składnię. Wybranie tej opcji powoduje zapisanie zalecanych przekształceń jako składni w zewnętrznym pliku. Ten element sterujący jest niedostępny w oknie dialogowym Interaktywne przygotowanie danych, ponieważ to okno wkleja przekształcenia jako składnię komend do okna składni po kliknięciu opcji **Wklej**.

Zapisz przekształcenia jako XML. Ta opcja powoduje zapisanie zalecanych przekształceń w postaci kodu XML do zewnętrznego pliku, który można scalić z PMML modelu za pomocą `TMS MERGE` albo zastosować względem innego zbioru danych za pomocą `TMS IMPORT`. Okno dialogowe Interaktywne przygotowanie danych nie zawiera tego elementu sterującego, ponieważ zapisuje ono przekształcenia w postaci kodu XML po kliknięciu opcji **Zapisz XML** na pasku narzędziu u góry okna dialogowego.

Karta Analiza

Uwaga: Karta Analiza jest używana w oknie dialogowym Interaktywne przygotowanie danych w celu umożliwienia wyświetlenia podglądu zalecanych przekształceń. Okno Automatyczne przygotowanie danych nie obejmuje tego kroku.

1. Jeśli ustawienia automatycznego przygotowywania danych są zadowalające, w tym zmiany dokonane na kartach Cele, Zmienne i Ustawienia, należy kliknąć przycisk **Analizuj dane**; algorytm zastosuje ustawienia do danych wejściowych i wyświetli wyniki na karcie Analiza.

Karta Analiza zwiera wyniki w postaci tabeli i wykresu, podsumowujące przetwarzanie danych oraz wyświetla zalecenia co do możliwych sposobów modyfikacji lub udoskonalenia danych do przeprowadzenia oceny. Następnie można wyświetlić podgląd i zaakceptować lub odrzucić zalecenia.

Karta Analiza składa się z dwóch paneli, widoku głównego z lewej strony i powiązanego lub dodatkowego widoku z prawej strony. Istnieją trzy główne widoki:

- Podsumowanie przetwarzania zmiennych (ustawienie domyślne). Więcej informacji można znaleźć w temacie “Podsumowanie przetwarzania zmiennej ” na stronie 14.
- Zmienne. Więcej informacji można znaleźć w temacie “Zmienne ” na stronie 14.
- Podsumowanie kroku. Więcej informacji można znaleźć w temacie “Podsumowanie kroku ” na stronie 15.

Istnieją cztery połączone/dodatkowe widoki:

- Jakość predykcji (ustawienie domyślne). Więcej informacji można znaleźć w temacie “Jakość predykcji ” na stronie 16.
- Tabela zmiennych. Więcej informacji można znaleźć w temacie “Tabela zmiennych ” na stronie 16.
- Szczegóły zmiennej. Więcej informacji można znaleźć w temacie “Szczegóły zmiennej ” na stronie 16.

- Szczegóły działania. Więcej informacji można znaleźć w temacie “Szczegóły działania ” na stronie 17.

Łącza pomiędzy widokami

W widoku głównym podkreślony tekst w tabelach umożliwia sterowanie wyświetlaniem w powiązanym widoku. Kliknięcie tekstu umożliwia uzyskanie szczegółowych informacji o konkretnej zmiennej, zestawie zmiennych lub kroku przetwarzania. Ostatnio wybrane łącze jest wyświetlane w ciemniejszym kolorze; ułatwi to identyfikację połączenia pomiędzy zawartością dwóch paneli widoku.

Resetowanie widoków

Aby ponownie wyświetlić oryginalne rekomendacje z karty Analiza i zrezygnować ze zmian wprowadzonych w widokach analizy, należy kliknąć przycisk **Resetuj** w dolnej części panelu głównego widoku.

Podsumowanie przetwarzania zmiennej

Tabela podsumowania przetwarzania zmiennej stanowi obraz stanu planowanego ogólnego wpływu przetwarzania, z uwzględnieniem zmian stanu predyktorów i liczby tworzonych predyktorów.

Należy pamiętać, że w rzeczywistości nie jest tworzony żaden model, dlatego nie jest dostępna żadna miara ani wykres zmiany ogólnej jakości predykcji przed przygotowaniem danych i po ich przygotowaniu; można jednak wyświetlić wykresy jakości predykcji dla pojedynczych zalecanych predyktorów.

Tabela zawiera następujące informacje:

- Liczba zmiennych przewidywanych.
- Liczba oryginalnych (wejściowych) predyktorów.
- Predyktory zalecane do użycia w analizie i modelowaniu. W tym łączna liczba zalecanych zmiennych; liczba zalecanych oryginalnych, nieprzekształconych zmiennych; liczba zalecanych przekształconych zmiennych (z wykluczeniem pośrednich wersji zmiennych, zmiennych pochodnych wyznaczonych na podstawie predyktorów typu data/czas oraz utworzonych predyktorów); liczba zalecanych zmiennych pochodnych, które zostały wyznaczone na podstawie zmiennej typu data/czas; oraz liczba zalecanych utworzonych predyktorów.
- Liczba predyktorów wejściowych niezalecanych do użycia w żadnej postaci, niezależnie od tego, czy są w oryginalnej postaci, takiej jak zmienna pochodna, czy stanowią dane wejściowe dla utworzonego predyktora.

Jeśli informacje w tabeli **Zmienne** są podkreślone, można je kliknąć, aby w powiązanym widoku wyświetlić dodatkowe szczegóły. W powiązanym widoku tabeli Zmienne wyświetlane są szczegóły na temat **zmiennej przewidywanej**, **predyktorów wejściowych** i **nieużywanych predyktorów wejściowych**. Więcej informacji można znaleźć w temacie “Tabela zmiennych ” na stronie 16. **Predyktory zalecane do użycia w analizie** są wyświetlane w powiązanym widoku jakości predykcji. Więcej informacji można znaleźć w temacie “Jakość predykcji ” na stronie 16.

Zmienne

W widoku głównym Zmienne wyświetlane są zmienne przetworzone oraz informacja, czy proces automatycznego przygotowania danych zaleca ich użycie w dalszej części modeli. Zalecenia dla każdej zmiennej można zastąpić; na przykład, aby wykluczyć utworzone predyktory lub uwzględnić predyktory, które proces automatycznego przygotowania danych zaleca wykluczyć. Jeśli zmienna została przekształcona, można zdecydować, czy zaakceptować zalecane przekształcenie, czy też użyć oryginalnej wersji.

Widok Zmienne składa się z dwóch tabel, jednej dla zmiennych przewidywanych, drugiej dla predyktorów, które zostały przetworzone lub utworzone

Tabela docelowa

Tabela **Zmienna przewidywana** jest wyświetlana tylko po zdefiniowaniu zmiennej w danych.

Tabela składa się z dwóch kolumn:

- **Name.** Jest to nazwa lub etykieta zmiennej przewidywanej; zawsze używana jest oryginalna nazwa, nawet jeśli zmienna została przekształcona.
- **Poziom pomiaru.** Wyświetla ikonę reprezentującą poziom pomiaru; należy ustawić wskaźnik myszy nad ikoną, aby wyświetlić etykietę (ilościowy, porządkowy, nominalny itd.), która opisuje dane.

  Jeśli zmienna przewidywana została przekształcona, ostatnia przekształcona wersja ma odzwierciedlenie w kolumnie **Poziom pomiaru**. *Uwaga*: nie można wyłączyć przekształceń dla zmiennych przewidywanych.

Tabela predyktorów

Tabela **Predyktory** jest zawsze wyświetlana. Każdy wiersz w tabeli odpowiada jednej zmiennej. Domyślnie wiersze są posortowane w porządku malejącym jakości predykcji.

W przypadku predyktorów porządkowych nazwa oryginalna zawsze używana jest jako nazwa wiersza. W tabeli wyświetlane są wersje oryginalne i pochodne zmiennej typu data/czas (w osobnych wierszach); tabela zawiera również predyktory utworzone.

Należy pamiętać, że przekształcone wersje zmiennych wyświetlane w tabeli zawsze reprezentują wersje końcowe.

Domyślnie w tabeli predyktorów wyświetlane są tylko zmienne zalecane. Aby wyświetlić pozostałe zmienne, należy zaznaczyć pole **Umieść nierekomendowane zmienne w tabeli** nad tabelą; zmienne te zostaną wówczas wyświetlone u dołu tabeli.

Tabela zawiera następujące kolumny:

- **Wersja do użycia.** Wyświetla listę rozwijaną, która pozwala kontrolować, czy zmienna będzie używana w dalszej części strumienia i czy użyć sugerowanych przekształceń. Domyślnie lista rozwijana odzwierciedla zalecenia.

  W przypadku predytorów porządkowych, które zostały przekształcone, lista rozwijana zawiera trzy opcje do wyboru: **Przekształcone**, **Oryginalne** i **Nie używaj**.

  W przypadku predytorów porządkowych, które nie zostały przekształcone, dostępne opcje to: **Oryginalne** i **Nie używaj**.

  Dla zmiennych pochodnych typu data/czas i predyktorów tworzonych dostępne są następujące opcje: **Przekształcone** i **Nie używaj**.

  Dla oryginalnych zmiennych daty lista rozwijana jest wyłączona i ustawiona jest opcja **Nie używaj**.

  Uwaga: W przypadku predyktorów oryginalnych i przekształconych zmiana wersji pomiędzy **Oryginalne** i **Przekształcone** powoduje automatyczną aktualizację ustawień **Poziom pomiaru** i **Jakość predykcji** dla tych zmiennych.
- **Name.** Każda nazwa zmiennej stanowi odsyłacz. Kliknięcie nazwy pozwala wyświetlić dodatkowe informacje na temat zmiennej w powiązanym widoku. Więcej informacji zawiera temat “Szczegóły zmiennej ” na stronie 16.
- **Poziom pomiaru.** Wyświetla ikonę reprezentującą typ danych; należy ustawić wskaźnik myszy nad ikoną, aby wyświetlić etykietę (ilościowy, porządkowy, nominalny itd.), która opisuje dane.
- **Jakość predykcji.** Jakość predykcji jest wyświetlana tylko dla zmiennych zalecanych przez proces automatycznego przygotowania danych. Ta kolumna nie jest wyświetlana, jeśli nie zdefiniowano żadnej zmiennej przewidywanej. Jakość predykcji może należeć do zakresu od 0 do 1, przy czym wyższe wartości oznaczają „lepsze” predyktory. Ogólnie jakość predykcji jest przydatna do porównywania predyktorów z analizą procesu automatycznego przygotowania danych, ale wartości jakości predykcji nie powinny być porównywane w ramach analizy.

Podsumowanie kroku

W każdym kroku wykonywanym w ramach automatycznego przygotowania danych predyktory wejściowe są przekształcane i/lub filtrowane; zmienne, które pozostaną po wykonaniu jednego kroku, są wykorzystywane w kolejnym. Zmienne, które pozostaną do ostatniego kroku, są wówczas zalecane do

użycia w modelowaniu, natomiast wartości wejściowe dla przekształconych i tworzonych predyktorów zostają odfiltrowane.

Podsumowanie kroku to prosta tabela, która zawiera listę czynności przetwarzania wykonywanych w procesie automatycznego przygotowania danych. Jeśli jakiś **krok** jest podkreślony, można go kliknąć, aby w powiązanym widoku wyświetlić dodatkowe szczegóły na temat wykonywanej czynności. Więcej informacji można znaleźć w temacie “Szczegóły działania ” na stronie 17.

Uwaga: Wyświetlane są tylko oryginalne i końcowe przekształcone wersje poszczególnych zmiennych; wersje pośrednie, używane w czasie analizy, nie są wyświetlane.

Jakość predykcji

Wykres jest wyświetlany domyślnie po pierwszym uruchomieniu analizy lub po wybraniu opcji **Predyktory zalecane do wykorzystania w analizie** w widoku głównym Podsumowanie przetwarzania zmiennych; przedstawia jakość predykcji zalecanych predyktorów. Zmienne są uporządkowane według jakości predykcji, przy czym zmienna o najwyższej wartości jest wyświetlana na samej górze.

W przypadku przekształconych wersji predyktorów porządkowych nazwa zmiennej zawiera przyrostek wybrany w panelu Nazwy zmiennych na karcie Ustawienia; przykładowo: *_transformed*.

Ikony poziomów pomiaru są wyświetlane za nazwami poszczególnych zmiennych.

Jakość predykcji każdego zalecanego predyktora jest obliczana na podstawie regresji liniowej lub modelu Naïve Bayes, w zależności od tego, czy zmienna przewidywana jest ilościowa, czy jakościowa.

Tabela zmiennych

Widok Tabela zmiennych jest wyświetlany po kliknięciu opcji **Zmienna przewidywana**, **Predyktory** lub **Niewykorzystane predyktory** w widoku głównym Podsumowanie przetwarzania zmiennych; zawiera prostą tabelę z listą odpowiednich funkcji.

Tabela składa się z dwóch kolumn:

- **Name.** Nazwa predyktora.

  W przypadku zmiennych przewidywanych używana jest oryginalna nazwa lub etykieta zmiennej, nawet jeśli zmienna została przekształcona.

  W przypadku przekształconych wersji predyktorów porządkowych nazwa zawiera przedrostek wybrany w panelu Nazwy zmiennych na karcie Ustawienia; przykładowo: *_transformed*.

  W przypadku zmiennych pochodnych wyznaczonych na podstawie daty i czasu używana jest nazwa końcowej przekształconej wersji; na przykład: *bdate_years*.

  W przypadku tworzonych predyktorów używana jest nazwa utworzonego predyktora; na przykład: *Predictor1*.
- **Poziom pomiaru.** Wyświetla ikonę reprezentującą typ danych.

  Dla zmiennych przewidywanych opcja **Poziom pomiaru** zawsze odzwierciedla wersję przekształconą (o ile zmienna przewidywana została przekształcona); na przykład, zmiana typu z porządkowego (zbiór uporządkowany) na ilościowy (zakres, skala) lub odwrotnie.

Szczegóły zmiennej

Widok Szczegóły zmiennej jest wyświetlany po kliknięciu dowolnej **nazwy** w widoku głównym Zmienne; zawiera wykres rozkładu, braków danych i jakości predykcji (o ile ma zastosowanie) dla wybranej zmiennej. Ponadto wyświetlana jest również historia przetwarzania zmiennej oraz nazwa zmiennej przekształconej (o ile ma zastosowanie).

Dla każdego zestawu wykresów wyświetlane są obok siebie dwie wersje do porównania zmiennej: z zastosowanym przekształceniem i bez przekształcenia; jeśli wersja zmiennej po przekształceniu nie istnieje, wyświetlany jest tylko wykres dla oryginalnej zmiennej. Dla zmiennych pochodnych typu data lub czas oraz utworzonych predyktorów wyświetlane są tylko wykresy dla nowego predyktora.

Uwaga: Jeśli zmienna została wykluczona z powodu zbyt dużej liczby kategorii, wyświetlana jest tylko historia przetwarzania.

Wykres rozkładu

Rozkład zmiennych ilościowych jest wyświetlany w postaci histogramu, z nałożoną krzywą normalną i pionową linią odniesienia dla wartości średniej; zmienne jakościowe są wyświetlane w postaci wykresu słupkowego.

Histogramy są opatrzone etykietami, które wskazują odchylenie standardowe i skośność; skośność nie jest jednak wyświetlana, jeśli liczba wartości wynosi 2 lub mniej lub wariancja oryginalnej zmiennej jest mniejsza niż 10-20.

Ustawienie wskaźnika myszy nad wykresem pozwala wyświetlić średnią dla histogramów lub liczebność i wartość procentową łącznej liczby rekordów dla kategorii na wykresach słupkowych.

Wykres braków danych

Na wykresach kołowych porównywana jest wartość procentowa braków danych z zastosowanym przekształceniem lub bez przekształcenia; etykiety na wykresie wskazują wartość procentową.

Jeśli proces automatycznego przygotowania danych obejmował traktowanie braków danych, na wykresie kołowym po przekształceniu w postaci etykiety przedstawiana jest również wartość zastępcza — czyli wartość użyta zamiast braków danych.

Ustawienie wskaźnika myszy nad wykresem spowoduje wyświetlenie liczebności oraz wartości procentowej braków danych dla łącznej liczby rekordów.

Wykres jakości predykcji

Dla zmiennych zalecanych wykresy słupkowe przedstawiają jakość predykcji przed przekształceniem i po przekształceniu. Jeśli zmienna przewidywana została przekształcona, obliczona jakość predykcji odnosi się do przekształconej zmiennej.

Uwaga: Wykresy jakości predykcji nie są wyświetlane, jeśli nie zdefiniowano żadnej zmiennej przewidywanej lub jeśli zmienna przewidywana została kliknięta w panelu widoku głównego.

Ustawienie wskaźnika myszy nad wykresem pozwala wyświetlić wartość jakości predykcji.

Tabela historii przetwarzania

Tabela przedstawia, w jaki sposób wyliczona została przekształcona wersja zmiennej. Kroki wykonane w procesie automatycznego przygotowania danych są wyświetlane w kolejności, w jakiej zostały wykonane; jednak w przypadku niektórych kroków dla danej zmiennej wykonanych mogło być kilka czynności.

Uwaga: Ta tabela nie jest wyświetlana dla zmiennych, które nie zostały przekształcone.

Informacje w tabeli są podzielone na dwie lub trzy kolumny:

- **Podjęte działania.** Nazwa podjętego działania. Na przykład Predyktory ilościowe. Więcej informacji można znaleźć w temacie “Szczegóły działania ” na stronie 17.
- **Details.** Lista przeprowadzonych procesów. Na przykład Transformuj do jednostek standardowych.
- **Funkcja.** Opcja wyświetlana tylko dla utworzonych predyktorów; wyświetla kombinację liniową zmiennych wejściowych, na przykład: 0,06*age + 1,21*height, gdzie age oznacza wiek, a height wzrost.

Szczegóły działania

Widok powiązany Szczegóły działania jest wyświetlany po wybraniu dowolnego podkreślonego **działania** w widoku głównym Podsumowanie kroku; wyświetla informacje specyficzne dla działania oraz informacje wspólne dla wszystkich wykonanych kroków przetwarzania; szczegóły specyficzne dla działania są wyświetlane jako pierwsze.

Dla każdego kroku w górnej części powiązanego widoku zamieszczany jest opis, który stanowi jego tytuł. Szczegóły specyficzne dla działania są wyświetlane pod tytułem i mogą zawierać informacje,

takie jak liczba wyliczonych predyktorów, zmienne rekategoryzowane, przekształcenia zmiennych przewidywanych, kategorie połączone lub uporządkowane oraz predyktory utworzone lub wykluczone.

W trakcie przetwarzania poszczególnych działań liczba użytych predyktorów może ulec zmianie, na przykład po wykluczeniu lub połączeniu predyktorów.

Uwaga: Jeśli działanie zostało wyłączone lub nie określono żadnej zmiennej przewidywanej, po kliknięciu działania w widoku głównym Podsumowanie kroku zamiast szczegółów działania wyświetlany jest komunikat o błędzie.

Dostępnych jest dziewięć działań; jednak nie wszystkie muszą być aktywowane dla każdej analizy.

Tabela zmiennych tekstowych

W tabeli wyświetlane są:

- Predyktory wykluczone z analizy.

Tabela predyktorów daty i czasu

W tabeli wyświetlane są:

- Czasy trwania wyznaczone na podstawie predyktorów daty i czasu.
- Elementy daty i czasu.
- Wyliczone predyktory daty i czasu, łącznie.

Data lub czas odniesienia są wyświetlane jako przypis, o ile czasy trwania zostały obliczone.

Tabela monitorowania predyktorów

W tabeli wyświetlana jest liczba następujących predyktorów wykluczonych z przetwarzania:

- Stałe.
- Predyktory ze zbyt dużą liczbą braków danych.
- Predyktory ze zbyt dużą obserwacji w jednej kategorii.
- Zmienne nominalne (zbiory) ze zbyt dużą liczbą kategorii.
- Predyktory monitorowane, łącznie.

Tabela sprawdzania poziomu pomiaru

W tabeli wyświetlana jest liczba rekategoryzacji zmiennych, podzielona w następujący sposób:

- Zmienne porządkowe (zbiór uporządkowany) uznane za ilościowe.
- Zmienne ilościowe uznane za porządkowe.
- Łączna liczba rekategoryzacji.

Jeśli żadna zmienna wejściowa (zmienne przewidywane lub predyktory) nie była ilościowa lub porządkowa, informacja ta jest wyświetlana jako przypis.

Tabela wartości odstających

W tej tabeli przedstawiana jest liczebność wartości odstających, jakie były obsługiwane.

- Liczba zmiennych ilościowych, dla których wartości odstające zostały wykryte i odcięte, lub liczba zmiennych ilościowych, dla których wartości odstające zostały wykryte i ustawione jako braki danych, w zależności od ustawień w panelu Zmienne wejściowe i przewidywana na karcie Ustawienia.
- Liczba zmiennych ilościowych wykluczonych, ponieważ były stałe, po zakończeniu działań związanych z obsługą wartości odstających.

Jeden przypis przedstawia wartość odcięcia dla wartości odstających; drugi przypis jest wyświetlany, jeśli żadna zmienna wejściowa (przewidywana lub predyktor) nie była ilościowa.

Tabela braków danych

W tabeli wyświetlana jest liczba zmiennych z zastąpionymi brakami danych, podzielona na następujące części:

- Zmienna przewidywana. Ten wiersz nie jest wyświetlany, jeśli nie określono żadnej zmiennej przewidywanej.
- Predyktory. Ten obszar jest następnie podzielony na liczbę predyktorów nominalnych (zbiór), porządkowych (zbiór uporządkowany) i ilościowych.
- Łączna liczba zastąpionych braków danych.

Tabela zmiennych przewidywanych

Ta tabela przedstawia, czy zmienna przewidywana została przekształcona, w następujący sposób:

- Transformacja Boxa-Coxa na normalność. Ten obszar jest podzielony na kolumny, w których przedstawiane są określone kryteria (średnia i odchylenie standardowe) oraz parametr Lambda.
- Kategorie zmiennej przewidywanej ze zmienioną kolejnością w celu zwiększenia stabilności.

Tabela predyktorów jakościowych

W tej tabeli wyświetlana jest liczba predyktorów jakościowych:

- Których kolejność kategorii została zmieniona z najniższej na najwyższą w celu zwiększenia stabilności.
- Których kategorie zostały połączone w celu zmaksymalizowania powiązań ze zmienną przewidywaną.
- Których kategorie zostały połączone w celu umożliwienia obsługi kategorii małolicznych.
- Wykluczonych z powodu słabego powiązania ze zmienną przewidywaną.
- Wykluczonych, ponieważ były stałe po połączeniu.

Przypis jest wyświetlany, jeśli nie wystąpiły żadne predyktory jakościowe.

Tabela predyktorów ilościowych

Dostępne są dwie tabele. Pierwsza wyświetla jedno z następujących przekształceń:

- Wartości predyktora przekształcone na jednostki standardowe. Ponadto przedstawia liczbę przekształconych predyktorów, określoną średnią oraz odchylenie standardowe.
- Wartości predyktora zmapowane do wspólnego przedziału. Ponadto przedstawia liczbę predyktorów przekształconych z zastosowaniem transformacji min.-maks. oraz określone wartości minimalne i maksymalne.
- Wartości predykcyjne skategoryzowane oraz liczba skategoryzowanych predyktorów.

W drugiej tabeli przedstawiane są szczegóły dotyczące tworzenia przestrzeni predyktorów, wyświetlane jako liczba predyktorów:

- Utworzonych.
- Wykluczonych z powodu słabego powiązania ze zmienną przewidywaną.
- Wykluczonych, ponieważ były stałe po kategoryzacji.
- Wykluczonych, ponieważ były stałe po utworzeniu.

Przypis jest wyświetlany, jeśli żaden z predyktorów wejściowych nie był ilościowy.

Przywrócone oceny

Jeśli zmienna przewidywana została przekształcona przez ADP, kolejne modele budowane z użyciem przekształconej zmiennej przewidywanej oceniają przekształcone jednostki. Aby możliwe było interpretowanie i korzystanie z wyników, należy przekonwertować wartość predykcyjną z powrotem do oryginalnej skali.

1. W celu przywrócenia oceny wybierz z menu następujące opcje:

   Transformacja > **Przygotuj dane do modelowania** > **Przywrócone oceny...**
2. Wybierz zmienną do przywrócenia oceny. Ta zmienna powinna zawierać przewidziane przez model wartości przekształconej zmiennej przewidywanej.

3. Określ przyrostek dla nowej zmiennej. Ta nowa zmienna będzie zawierała wartości przewidziane przez model w oryginalnej skali nieprzekształconej zmiennej przewidywanej.
4. Określ lokalizację pliku XML zawierającego przekształcenia ADP. Powinien to być plik z okna dialogowego Interaktywne przygotowanie danych albo Automatyczne przygotowanie danych. Więcej informacji można znaleźć w temacie “Stosowanie i zapisywanie przekształceń ” na stronie 12 .

Zidentyfikuj obserwacje nietypowe

Procedura wykrywania anomalii umożliwia wyszukiwanie nietypowych obserwacji na podstawie odchyleń od norm właściwych dla ich grup. Procedura została zaprojektowana w taki sposób, aby szybko wykrywać nietypowe obserwacje na potrzeby celów kontroli danych — na etapie eksploracyjnej analizy danych, przed przeprowadzeniem jakiejkolwiek analizy służącej do wnioskowania. Ten algorytm jest przeznaczony do ogólnego wykrywania anomalii; tj. definicja nietypowej obserwacji nie jest specyficzna dla konkretnego zastosowania, takiego jak wykrywanie nietypowych wzorców płatności w służbie zdrowia lub wykrywanie prania brudnych pieniędzy w branży finansowej, w których to branżach definicje anomalii mogą być dobrze zdefiniowane.

Przykład. Analityk danych zatrudniony do tworzenia modeli predykcyjnych wyników leczenia pacjentów po udarze martwi się o jakość danych, ponieważ takie modele mogą być wrażliwe na anormalne obserwacje. Niektóre z takich obserwacji reprezentują prawdziwie unikatowe przypadki, które tym samym nie są odpowiednie do celów prognostycznych, natomiast inne obserwacje są powodowane przez błędy wprowadzania danych, w których wartości są technicznie „poprawne” i nie mogą zostać odnalezione za pomocą procedur walidacji danych. Procedura Odkryj obserwacje anormalne znajduje i raportuje takie odstające obserwacje, dzięki czemu analityk może zdecydować, co z nimi zrobić.

Statystyki. Procedura tworzy grupy o zbliżonych wartościach, normy grup o zbliżonych wartościach dla zmiennych ciągłych i jakościowych, indeksy anomalii na podstawie odchyleń od norm grup o zbliżonych wartościach oraz wartości wpływu zmiennych dla zmiennych, które w największym stopniu przyczyniły się do wystąpienia obserwacji, która została sklasyfikowana jako nietypowa.

Zagadnienia dotyczące danych

Dane. Procedura znajduje zastosowanie zarówno w przypadku zmiennych ilościowych, jak i zmiennych kategorialnych. Każdy wiersz reprezentuje odrębną obserwację, a każda kolumna reprezentuje odrębną zmienną, na której oparte są grupy o zbliżonych wartościach. Zmienna identyfikacji obserwacji może być dostępna w pliku danych na potrzeby oznaczania danych wyjściowych, ale nie będzie używana w analizie. Dozwolone są brakujące wartości. Zmienna ważąca, jeśli została określona, jest ignorowana.

Model wykrywania może zostać zastosowany do nowego pliku danych testowych. Elementy danych testowych muszą być takie same, jak elementy danych uczących. W zależności od ustawień algorytmu metoda obsługi braków danych, która jest używana przy tworzenia modelu, może zostać przed dokonaniem oceny zastosowana do pliku danych testowych.

Kolejność obserwacji. Należy pamiętać, że rozwiązanie może zależeć od kolejności obserwacji. Aby zminimalizować wpływ kolejności, należy losowo ustawić obserwacje. Aby sprawdzić stabilność danego rozwiązania, można uzyskać kilka różnych rozwiązań z obserwacjami posortowanymi w różnej, losowej kolejności. W sytuacji, gdy wielkość pliku jest ekstremalnie duża, wykonywanych jest wiele uruchomień z użyciem próby obserwacji posortowanych w różnej kolejności losowej.

Założenia. W algorytmie zakłada się, że wszystkie zmienne są niestałe i niezależne, a w żadnej obserwacji nie brakuje wartości żadnej ze zmiennych wejściowych. Zakłada się, że każda zmienna ciągła ma rozkład normalny (Gaussa), a każda zmienna kategorialna ma rozkład wielomianowy. Empiryczne testy wewnętrzne wskazują, że procedura jest dość odporna na naruszenia zarówno założeń niezależności, jak i założeń co do rozkładu, ale użytkownik powinien orientować się, na ile dobrze te założenia są spełnione w jego konkretnym przypadku.

Aby zidentyfikować nietypowe obserwacje

1. Z menu wybierz:

   Dane > **Zidentyfikuj obserwacje nietypowe...**
2. Wybierz co najmniej jedną zmienną do analizy.

3. Opcjonalnie można wybrać zmienną identyfikatora obserwacji, aby użyć jej do etykietowania wyniku.

Zmienne z nieznanym poziomem pomiaru

Alert poziomu pomiaru wyświetla się, gdy poziom pomiaru dla jednej lub większej ilości zmiennych w zbiorze danych jest nieznany. Ponieważ poziom pomiaru wpływa na wyliczenie wyników dla tej procedury, wszystkie zmienne muszą mieć zdefiniowany poziom pomiaru.

Skanowanie danych. Odczytuje dane w aktywnym zbiorze danych i przypisuje domyślny poziom pomiaru do wszystkich zmiennych, które mają aktualnie nieznany poziom pomiaru. Jeśli zbiór danych jest duży, może to zająć trochę czasu.

Przypisz ręcznie. Otwiera okno dialogowe, które zestawia wszystkie zmienne z nieznanym poziomem pomiaru. Można użyć tego okna dialogowego do przypisania poziomu pomiaru do tych zmiennych. Można również przypisać poziom pomiaru w Widoku zmiennych Edytora danych.

Ponieważ poziom pomiaru jest ważny dla tej procedury, nie można wejść do tego okna dialogowego w celu uruchomienia tej procedury, dopóki wszystkie zmienne nie będą miały zdefiniowanego poziomu pomiaru.

Identyfikowanie obserwacji nietypowych: Wyniki

Lista wyjątkowych obserwacji i przyczyny, dla których zostały uznane za nietypowe. W przypadku wybrania tej opcji zostaną utworzone trzy tabele:

- Na liście indeksów obserwacji nietypowych wyświetlane są obserwacje, które są identyfikowane jako nietypowe, wraz z odpowiadającymi im wartościami indeksów anomalii.
- Lista identyfikatorów grup o zbliżonych wartościach zawiera nietypowe obserwacje i informacje dotyczące odpowiadających im grup o zbliżonych wartościach.
- Lista przyczyn anomalii zawiera numer obserwacji, zmienną przyczyny, wartość zmiennej wpływu, wartość zmiennej oraz normę zmiennej dla każdej przyczyny.

Wszystkie tabele są sortowane wg indeksu anomalii w porządku malejącym. Ponadto wyświetlane są identyfikatory obserwacji, jeśli zmienna identyfikatora obserwacji została określona na karcie Zmienne.

Podsumowania. Elementy sterujące w tej grupie generują podsumowania rozkładu.

- **Normy grup obserwacji o zbliżonych wartościach.** Ta opcja powoduje uwidocznienie tabeli norm zmiennych ciągłych (jeśli w analizie używane są jakiekolwiek zmienne ciągłe) i tabeli norm zmiennych kategorialnych (jeśli w analizie używane są jakiekolwiek zmienne kategorialne). Tabela norm zmiennych ciągłych zawiera średnią i odchylenie standardowe każdej zmiennej ciągłej dla każdej grupy o zbliżonych wartościach. Tabela norm zmiennych kategorialnych zawiera dominantę (najbardziej popularna kategoria), częstotliwość oraz procent częstotliwości każdej zmiennej kategorialnej dla każdej grupy o zbliżonych wartościach. Średnia ze zmiennej ilościowej i dominanta zmiennej kategorialnej są używane jako wartości norm w analizie.
- **Indeksy anomalii.** Podsumowanie indeksu anomalii zawiera statystyki opisowe dla indeksu anomalii obserwacji zidentyfikowanych jako najbardziej nietypowe.
- **Wystąpienia analizowanej zmiennej jako przyczyny.** Dla każdej przyczyny w tabeli wyświetlana jest częstość i procent częstości występowania każdej zmiennej jako przyczyny. Tabela zawiera także statystyki opisowe wpływu każdej zmiennej. Jeśli na karcie Opcje ustawiona jest maksymalna liczba przyczyn równa 0, ta opcja nie jest dostępna.
- **Obserw. przetworzono.** W podsumowaniu przetwarzania obserwacji są wyświetlane liczebności i procenty liczebności wszystkich obserwacji w aktywnym zbiorze danych, obserwacje uwzględnione w analizie i wykluczone z analizy, a także obserwacje w każdej grupie o zbliżonych wartościach.

Identyfikowanie obserwacji nietypowych: Zapisz

Zapisz zmienne. Elementy sterujące w tej grupie pozwalają na zapisanie zmiennych modelu w aktywnym zbiorze danych. Można również zastąpić istniejące zmienne, których nazwy powodują konflikt ze zmiennymi, które mają zostać zapisane.

- **Indeks anomalii.** Zapisuje wartość indeksu anomalii dla każdej obserwacji w zmiennej o podanej nazwie.
- **Grupy o zbliżonych wartościach.** Zapisuje identyfikator grupy o zbliżonych wartościach, liczebność obserwacji i wielkość grupy jako procent dla każdej obserwacji w zmiennych z określonym trzonem nazwy. Na przykład, jeśli trzon nazwy *Peer* jest określony, zostaną wygenerowane zmienne *Peerid*, *PeerSize* i *PeerPctSize*. *Peerid* jest identyfikatorem grupy o zbliżonych wartościach, *PeerSize* jest wielkością grupy, a *PeerPctSize* jest wielkością grupy wyrażoną procentowo.
- **Przyczyny.** Zapisuje zestawy zmiennych przyczynowych o określonym trzonie nazwy. Zestaw zmiennych przyczynowych zawiera nazwy zmiennych przyczynowych, ich miary wpływu, ich własne wartości oraz wartości normatywnej. Liczba zestawów zależy od liczby przyczyn żądanych na karcie Opcje. Na przykład, jeśli określono trzon nazwy *Reason*, to wygenerowane zostaną zmienne *ReasonVar_k*, *ReasonMeasure_k*, *ReasonValue_k* i *ReasonNorm_k*, gdzie *k* jest *k*-tą przyczyną. Opcja ta nie jest dostępna, jeśli liczba przyczyn jest ustawiona na 0.

Eksportuj plik modelu. Umożliwia zapisanie modelu w formacie XML.

Identyfikowanie obserwacji nietypowych: Braki danych

Karta Braki danych służy do sterowania obsługą braków danych zdefiniowanych przez użytkownika i systemowych braków danych.

- **Wyklucz brakujące wartości z analizy.** Obserwacje z brakującymi wartościami są wykluczane z analizy.
- **Uwzględnij brakujące wartości w analizie.** Brakujące wartości zmiennych ciągłych są zastępowane odpowiednimi średnimi ogółem, a brakujące kategorie zmiennych kategorialnych są zgrupowane i traktowane jako ważna kategoria. Przetworzone zmienne są następnie używane w analizie. Opcjonalnie można zażądać utworzenia dodatkowej zmiennej, która reprezentować będzie odsetek brakujących zmiennych w każdej obserwacji, a następnie użyć tej zmiennej w analizie.

Identyfikowanie obserwacji nietypowych: Opcje

Kryteria identyfikacji nietypowych obserwacji. Poniższe ustawienia określają, ile obserwacji znajduje się na liście anomalii.

- **Odsetek obserwacji z największymi wartościami wskaźnika anomalii.** Podaj liczbę dodatnią, która jest mniejsza lub równa 100.
- **Ustalona liczba obserwacji z największymi wartościami wskaźników anomalii.** Podaj dodatnią liczbę całkowitą, mniejszą lub równą łącznej liczbie obserwacji w aktywnym zbiorze danych, które są używane w analizie.
- **Identyfikuj tylko te obserwacje, których wartość indeksu anomalii jest równa bądź przekracza wartość minimalną.** Podaj nieujemną liczbę całkowitą. Obserwacja jest uznawana za nietypową, jeśli jej wartość indeksu anomalii jest większa lub równa podanej wartości punktu odcięcia. Tej opcji używa się razem z opcjami **Procent obserwacji** i **Stała liczba obserwacji**. Na przykład, jeśli zostanie podana stała liczba 50 obserwacji i wartość odcięcia 2, lista anomalii będzie składać się z 50 obserwacji, z których każda będzie miała wartość indeksu anomalii równą lub większą od 2.

Liczba grup o zbliżonych wartościach. Procedura wyszukuje najlepszą liczbę grup o zbliżonych wartościach między podaną wartością minimalną i maksymalną. Wartości muszą być dodatnimi liczbami całkowitymi, a wartość minimalna nie może przekraczać wartości maksymalnej. Jeśli podane wartości są równe, procedura przyjmuje stałą liczbę grup o zbliżonych wartościach.

Uwaga: W zależności od zmienności danych mogą wystąpić sytuacje, w których liczba grup o zbliżonych wartościach potencjalnie pokrywanych przez dane jest mniejsza niż liczba określona jako minimalna. W takiej sytuacji procedura może wygenerować mniej grup o zbliżonych wartościach.

Maksymalna liczba przyczyn. Przyczyna składa się z miary wpływu zmiennej, nazwy zmiennej związanej z tą przyczyną, wartości zmiennej i zmiennej odpowiedniej grupy o zbliżonych wartościach. Określ nieujemną liczbę całkowitą. Jeśli ta wartość jest równa liczbie przetworzonych zmiennych używanych w analizie lub ją przekracza, zostaną wyświetlone wszystkie zmienne.

Dodatkowe właściwości komendy DETECTANOMALY

Język składni komend umożliwia również:

- Pomiń kilka zmiennych w aktywnym zbiorze danych z analizy bez jawnego określenia wszystkich zmiennych analizy (za pomocą opcji EXCEPT).
- Określ korektę, aby zrównoważyć wpływ zmiennych ilościowych i jakościowych (za pomocą słowa kluczowego MLWEIGHT w opcji CRITERIA).

Pełne informacje na temat składni znajdują się w podręczniku *Command Syntax Reference*.

Kategoryzacja optymalna

Procedura kategoryzacji optymalnej dyskretyzuje jedną lub więcej zmiennych ilościowych (zwanych dalej **wejściowymi zmiennymi kategoryzowanymi**) przez przekazanie wartości każdej zmiennej do przedziałów. Tworzenie przedziałów jest optymalne w odniesieniu do kategorialnej zmiennej nadzorującej proces umieszczania w przedziałach. Przedziałów można używać do dalszej analizy zamiast oryginalnych danych.

Przykłady. Zredukowanie liczby różnych wartości zmiennych ma wiele zastosowań, na przykład:

- Wymagania dotyczące danych w innych procedurach. Zmienne po dyskretyzacji można traktować jako jakościowe na potrzeby procedur, które wymagają zmiennych jakościowych. Przykładowo, procedura tabel krzyżowych wymaga, aby wszystkie zmienne były jakościowe.
- Ochrona prywatności danych. Tworzenie raportów z podzielonych wartości zamiast wartości rzeczywistych może ułatwić ochronę prywatności źródeł danych. Procedura Kategoryzacja optymalna pomoże w doborze kategorii.
- Szybkość i wydajność. Niektóre procedury są bardziej efektywne, jeśli są wykonywane z ograniczoną liczbą różnych wartości. Przykładowo, szybkość wielomianowej regresji logistycznej może zostać zwiększona poprzez użycie zmiennych po dyskretyzacji.
- Kompletna lub quasi-kompletna separacja danych.

Kategoryzacja optymalna a kategoryzacja wizualna. Okno dialogowe Kategoryzacja optymalna oferuje kilka automatycznych metod tworzenia kategorii bez użycia zmiennej optymalizującej. Te "nienadzorowane" reguły są przydatne podczas tworzenia statystyk opisowych, takich jak tabele częstości, natomiast kategoryzacja optymalna jest rozwiązaniem nadrzędnym, jeśli celem jest utworzenie modelu predykcyjnego.

Wynik. Procedura tworzy tabele punktów podziału dla kategorii oraz statystyki opisowe dla każdej wejściowej zmiennej dzielącej. Ponadto, można zapisać nowe zmienne do aktywnego zbioru danych zawierającego skategoryzowane wartości wejściowych zmiennych dzielących i zapisać reguły kategoryzacji w postaci składni komend w celu jej użycia podczas dyskretyzacji nowych danych.

Wymagania dotyczące danych dla kategoryzacji optymalnej

Dane. Ta procedura wymaga, aby wejściowe zmienne dzielące były ilościowymi zmiennymi numerycznymi. Zmienna optymalizująca powinna być jakościowa i może być łańcuchowa lub numeryczna.

Wykonanie kategoryzacji optymalnej

1. Z menu wybierz:

   Przekształcanie > **Kategoryzacja optymalna...**
2. Wybierz co najmniej jedną wejściową zmienną dzielącą.
3. Wybierz zmienną optymalizującą.

Zmienne zawierające skategoryzowane wartości danych nie są generowane domyślnie. Korzystając z karty Zapisz, zapisz te zmienne.

Kategoryzacja optymalna: Wynik

Karta Wynik steruje wyświetlaniem wyników.

- **Punkty brzegowe kategorii.** Wyświetla zestaw punktów brzegowych każdej wejściowej zmiennej kategoryzacji.
- **Statystyki opisowe dla kategoryzowanych zmiennych.** Dla każdej wejściowej zmiennej kategoryzacji ta opcja wyświetla liczbę obserwacji z ważnymi wartościami, liczbę obserwacji z brakami danych, liczbę różnych ważnych wartości oraz wartość minimalną i maksymalną. W przypadku zmiennej optymalizującej ta opcja wyświetla rozkłady klas dla każdej powiązanej wejściowej zmiennej kategoryzacji.
- **Entropia modelu dla kategoryzowanych zmiennych.** Dla każdej wejściowej zmiennej kategoryzacji opcja ta prezentuje miarę dokładności predykcyjnej zmiennej wejściowej w odniesieniu do zmiennej optymalizującej.

Kategoryzacja optymalna: Zapisz

Zapisz zmienne w aktywnym zbiorze danych. W dalszej analizie zamiast pierwotnych zmiennych można używać zmiennych zawierających kategoryzowane wartości danych.

Zapisz reguły kategoryzacji jako komendy. Generuje komendy, których można użyć do kategoryzacji innych zbiorów danych. Reguły rejestrowania bazują na punktach podziału określonych przez algorytm kategoryzacji.

Kategoryzacja optymalna: Braki danych

Na karcie Braki danych określa się, czy braki danych są usuwane obserwacjami, czy parami. Braki danych użytkownika są zawsze traktowane jako wartości nieważne. Podczas rekodowania pierwotnych wartości zmiennej na nową zmienną braki danych użytkownika są przekształcane w systemowe braki danych.

- **Parami.** Ta opcja operuje na każdej parze zmiennej optymalizującej i wejściowej zmiennej kategoryzacji. Procedura wykorzysta wszystkie obserwacje z niebrakującymi wartościami w zmiennej optymalizującej i wejściowej zmiennej kategoryzacji.
- **Obserwacjami** Ta opcja operuje na wszystkich zmiennych określonych na karcie Zmienne. Jeśli w obserwacji brakuje którejkolwiek zmiennej, cała obserwacja jest wykluczana.

Kategoryzacja optymalna: Opcje

Przetwarzanie wstępne. „Wstępna kategoryzacja” zmiennych wejściowych do kategoryzacji mających wiele różnych wartości może skrócić czas przetwarzania nie pogarszając istotnie jakości końcowych kategorii. Maksymalna liczba kategorii określa górny limit liczby tworzonych kategorii. Na przykład, jeśli podasz 1000 jako maksymalną liczbę, ale zmienna wejściowa do kategoryzacji ma mniej niż 1000 różnych wartości, liczba wstępnie utworzonych kategorii tej zmiennej będzie równa liczbie jej różnych wartości.

Kategorie małoliczne. Niekiedy procedura może wygenerować kategorie z bardzo małą liczbą obserwacji. Następująca strategia umożliwia usunięcie tych fałszywych punktów podziału:

Załóżmy, że dla danej zmiennej algorytm znalazł $n_{\text{końcowych}}$ punktów podziału, a zatem utworzył $n_{\text{końcowych}}+1$ kategorii. W przypadku $i = 2, \ldots, n_{\text{końcowych}}$ kategorii (druga kategoria o najniższej wartości do drugiej kategorii o najwyższej wartości), należy obliczyć

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}), sizeof(b_{i+1}))}$$

gdzie *sizeof(b)* to liczba obserwacji w kategorii.

Gdy ta wartość jest mniejsza niż określony próg łączenia, kategoria b_i jest uważana za małoliczną i jest łączona z kategorią b_{i-1} lub b_{i+1} (zależnie od tego, która z nich ma niższą entropię informacji).

Procedura obejmuje pojedyncze przetworzenie kategorii.

Punkty końcowe kategorii. Ta opcja określa, w jaki sposób ma być zdefiniowany dolny limit przedziału. Ponieważ procedura automatycznie określa wartości punktów podziału, jest to w dużej mierze kwestia preferencji.

Pierwsza (najniższa)/ostatni (najwyższa) kategoria. Te opcje określają, w jaki sposób zdefiniowane są minimalne i maksymalne punkty podziału dla każdej zmiennej wejściowej kategoryzacji. Ogólnie procedura zakłada, że zmienne wejściowe kategoryzacji mogą przyjmować dowolną wartość w wierszu liczb rzeczywistych, ale jeśli istnieje powód ograniczenia zakresu, można to zrobić przy użyciu najniższych/najwyższych wartości.

Dodatkowe właściwości komendy OPTIMAL BINNING

Język składni komend umożliwia również:

- Wykonać kategoryzację nienadzorowaną metodą równych częstotliwości (przy użyciu opcji komendy `CRITERIA`).

Pełne informacje na temat składni znajdują się w podręczniku *Command Syntax Reference*.

Uwagi

Niniejsza publikacja została przygotowana z myślą o produktach i usługach oferowanych w Stanach Zjednoczonych. IBM może udostępniać ten materiał 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 produkty, o których mowa w niniejszej publikacji. Przedstawienie tej 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 10504-1785U.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 (W TYM TAKŻE RĘKOJMI), WYRAŹNYCH LUB DOMNIEMANYCH, A W SZCZEGÓLNOŚCI DOMNIEMANYCH GWARANCJI PRZYDATNOŚCI HANDLOWEJ, PRZYDATNOŚCI DO OKREŚLONEGO CELU ORAZ GWARANCJI, ŻE PUBLIKACJA NIE NARUSZA PRAW STRON 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 tej 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 firm innych niż IBM zostały wprowadzone wyłącznie dla wygody użytkowników i w żadnym razie 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) umożliwienia wymiany informacji między niezależnie utworzonymi programami i innymi programami

(łącznie z opisywanym) oraz (ii) wykorzystywania wymienianych informacji, powinni skontaktować się z:

IBM Director of Licensing

IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785U.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 produktami firm innych niż IBM. Pytania dotyczące możliwości produktów firm innych niż IBM 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 codziennych operacjach działalności gospodarczej. W celu kompleksowego zilustrowania tej działalności podane przykłady zawierają nazwy osób, firm i ich produktów. Wszystkie te nazwy/nazwiska są fikcyjne i jakiekolwiek podobieństwo do istniejących nazw/nazwisk jest całkowicie przypadkowe.

LICENCJA W ZAKRESIE PRAW AUTORSKICH:

Niniejsza publikacja zawiera przykładowe aplikacje w kodzie źródłowym ilustrujące techniki programowania w różnych systemach operacyjnych. Użytkownik może kopiować, modyfikować i rozpowszechniać te programy przykładowe w dowolnej formie bez uiszczania opłat na rzecz IBM, w celu rozbudowy, użytkowania, handlowym lub w celu rozpowszechniania aplikacji zgodnych z aplikacyjnym interfejsem programowym dla tego systemu operacyjnego, dla którego napisane były programy przykładowe. Programy przykładowe nie zostały gruntownie przetestowane. IBM nie może zatem gwarantować ani sugerować niezawodności, użyteczności i funkcjonalności tych programów. Programy przykładowe są dostarczane w stanie, w jakim się znajdują ("AS IS"), bez jakichkolwiek gwarancji (rękojmię również wyłącza się). IBM nie ponosi odpowiedzialności za jakiekolwiek szkody wynikające z używania programów przykładowych.

Każda kopia programu przykładowego lub jakikolwiek jego fragment, jak też jakiekolwiek prace pochodne muszą zawierać następujące uwagi dotyczące praw autorskich:

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, 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 w Stanach 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 Open Group w Stanach Zjednoczonych i w innych krajach.

Java oraz wszystkie znaki towarowe i logo dotyczące Java są znakami towarowymi firmy i jej firm zależnych.

Indeks

A

Automatyczne przygotowanie danych 7
automatyczne przygotowywanie danych
    analiza zmiennych 14
    cele 7
    jakość predykcji 16
    konstruowanie predyktorów 12
    korekta poziomu pomiaru 10
    łącza pomiędzy widokami 13
    normalizacja docelowych wartości ilościowych 11
    podsumowanie kroku 15
    podsumowanie przetwarzania zmiennej 14
    pola 8
    poprawa jakości danych 10
    przekształcanie zmiennych 11
    przeskalowanie zmiennych 11
    przygotowanie daty i czasu 9
    przywracanie ocen 19
    resetowanie widoków 13
    statystyki wykluczonych zmiennych 9
    stosowanie przekształceń 12
    szczegóły działania 17
    szczegóły zmiennej 16
    tabela zmiennych 16
    widok modelu 13
    wybór predyktorów 12
    zmienne nazwy 12

B

braki danych
    w procedurze identyfikacji obserwacji nietypowych 22

C

cykliczne elementy czasu
    automatyczne przygotowywanie danych 9
czasy trwania, obliczanie
    automatyczne przygotowywanie danych 9

D

Definiowanie reguł walidacyjnych
    reguły jednej zmiennej 2
    reguły wielu zmiennych 3

G

grupy elementów równorzędnych
    w procedurze identyfikacji obserwacji nietypowych 21

I

indeksy anomalii
    w procedurze identyfikacji obserwacji nietypowych 21
Interaktywne przygotowanie danych 7

K

kategoryzacja nadzorowana
    a kategoryzacja nienadzorowana 23
    w kategoryzacji optymalnej 23
kategoryzacja nienadzorowana
    a kategoryzacja nadzorowana 23
Kategoryzacja optymalna
    brakujące wartości 24
    opcje 24
    wyniki 23
    zapisywanie 24
konstruowanie predyktorów
    w czasie automatycznego przygotowywania danych 12

M

MDLP
    w kategoryzacji optymalnej 23

N

naruszenia reguł walidacji
    w oknie Walidacja danych 6
naruszone reguły walidacyjne
    w oknie Walidacja danych 6
niekompletne identyfikatory obserwacji
    w oknie Walidacja danych 6
normalizacja docelowych wartości ilościowych 11

O

obliczanie czasów trwania
    automatyczne przygotowywanie danych 9

P

powtórzone identyfikatory obserwacji
    w oknie Walidacja danych 6
przyczyny
    w procedurze identyfikacji obserwacji nietypowych 21
punkty brzegowe kategorii
    w kategoryzacji optymalnej 23
puste obserwacje
    w oknie Walidacja danych 6

R

reguły kategoryzacji
    w kategoryzacji optymalnej 24
reguły walidacyjne 1
reguły walidacyjne pojedynczej zmiennej
    w oknie Definiowanie reguł walidacyjnych 2
    w oknie Walidacja danych 5

reguły walidacyjne wielu zmiennych
w oknie Definiowanie reguł walidacyjnych 3
w oknie Walidacja danych 6

S

sprawdzenie poprawności danych
w oknie Walidacja danych 4
Sprawdzenie poprawności danych
reguły wielu zmiennych 6
sprawdzenia podstawowe 4
wyniki 6
zapisywanie zmiennych 6

T

transformacja Boxa-Coxa
w czasie automatycznego przygotowywania danych 11

W

waga analizy
w czasie automatycznego przygotowywania danych 11
Walidacja danych
reguły jednej zmiennej 5
widok modelu
w czasie automatycznego przygotowywania danych 13
wstępna kategoryzacja
w kategoryzacji optymalnej 24
wybór predyktorów
w czasie automatycznego przygotowywania danych 12

Z

Zidentyfikuj obserwacje nietypowe
braki danych 22
dane wyjściowe 21
eksportowanie pliku modelu 21
opcje 22
zapisywanie zmiennych 21