

ITA-102 Hurtownie Danych

Marcin Gorawski, Michał Jan Gorawski

Moduł 9

Wersja 1.0

Data Mining I

Spis treści

Data Mining I	1
Informacje o module	2
Przygotowanie teoretyczne	3
Podstawy teoretyczne	3
Laboratorium podstawowe	6
Problem 1 (czas realizacji 25 min)	6
Problem 2 (czas realizacji 15min) -	7
Laboratorium podstawowe	11
Problem 2 (czas realizacji 25 min) – kontynuacja modułu 1	11
Laboratorium rozszerzone	12
Zadanie 1 (czas realizacji 30 min)	12
Zadanie 1 (czas realizacji 60 min)	12

Informacje o module

Opis modułu

W module poznasz zaawansowane zastosowania modeli eksploracji danych. Dowiesz się czym jest Data Mining (Eksploracja Danych), w jakich przypadkach można ją wykorzystać, jakie są korzyści wykorzystania tych struktur oraz jak korzystać z Data Miningu w SQL Server 2008.

Cel modułu

Przekazanie informacji na temat modeli służących do określenia potencjalnych odbiorców akcji reklamowej, oraz metody porównań różnych algorytmów.

Uzyskane kompetencje

Po zrealizowaniu modułu będziesz:

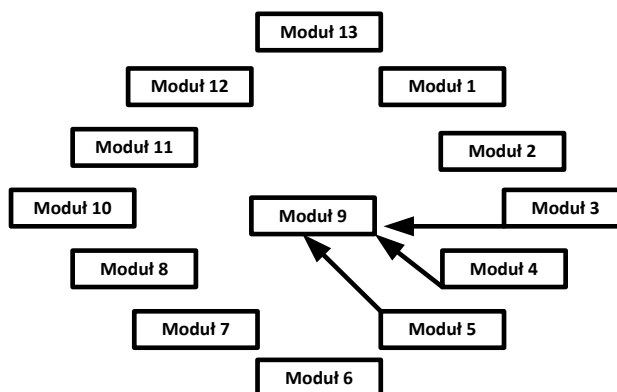
- wiedział czym jest Data Mining i do czego jest wykorzystywany
- potrafił stworzyć modele Dataminingowe w SQL Server 2008

Wymagania wstępne

Przed przystąpieniem do pracy z tym modulem powinieneś:

- wiedział teorię Ekstrakcji Danych
- potrafić utworzyć modele i przeanalizować modele Ekstrakcji danych
- rozumiał zasady tworzenia i analizy modeli Ekstrakcji Danych

Zgodnie z mapą zależności przedstawioną na Rys. 1, przed przystąpieniem do realizacji tego modułu należy zapoznać się z materiałem zawartym w modułach 3, 4, 5.



Rys. 1 Mapa zależności modułu

Przygotowanie teoretyczne

Podstawy teoretyczne

Początek dziedziny **odkrywania wiedzy** (ang. *Knowledge Discovery in Databases KDD*) poprzez **eksplorację danych** (ang. *Data Mining DM*) sięgają 5000 lat wstecz, kiedy to ludzie kultury sumeryjskiej gromadzili zapisy podatkowe na glinianych tabliczkach. Od tego czasu, rozwijano techniki gromadzenia i analizy tego typu informacji. Poniżej przykłady KDD.

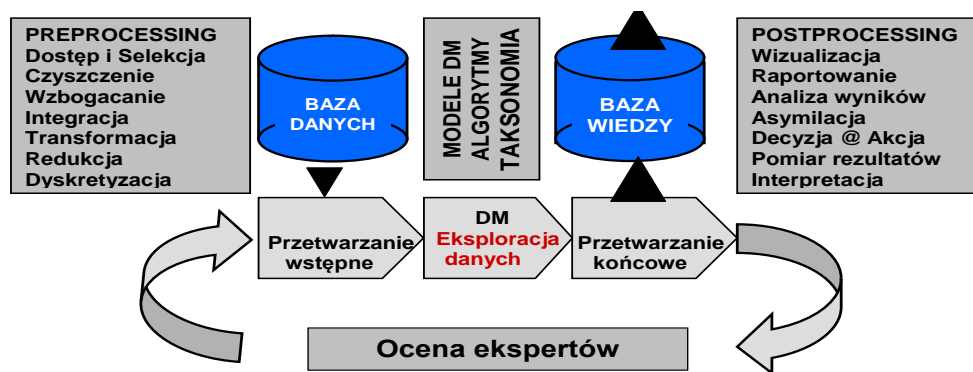
Analizy korporacyjne i sterowanie ryzykiem

- Planowanie finansowe i ewolucja aktywów:
 - Analiza i predykcja przepływu pieniędzy.
 - Analiza żądań warunkowych zapewniających rozwój aktywów.
 - Analiza przekrojowo-profilowana i szeregów czasowych (finansowe wskaźniki, analiza trendów, etc).
- Konstruowanie zasobów:
 - Sumaryzacja - porównywanie zasobów i ich zużycia.
- Konkurencja:
 - Monitor konkurencji i kierunki rynku (CI: inteligentna konkurencja).
 - Segmentowanie klientów w klasy i klasy bazujące na procedurach wyceny.
 - Ustawienie strategii wyceny dla wysoce konkurencyjnego rynku.

Zarządzanie rynkiem

- Lokowanie danych do analizy:
 - transakcje kart kredytowych, karty lojalności klienta, kupony dyskontowe, zgłoszenia skarg klientów, badania stylu życia (publicznego).
- Marketing ukierunkowany - poszukiwanie klasterów „modelu” klientów, którzy dzielą pewne charakterystyki np.: zainteresowań, poziom dochodu, przyzwyczajień, itp.;
- Określenie wzorców nadprogramowo kupowania dla klienta -Konwersja pojedynczego rachunku bankowego na rachunek wspólny : skojarzenie, etc.
- Analiza przekrojowa rynku:
 - Skojarzenia/korelacje pomiędzy sprzedażami produktu.
 - Predykcja bazująca na skojarzonej informacji.

KDD odwołuje się do całościowego procesu odkrywania użytecznej wiedzy z danych, podczas, gdy DM to szczególny krok w tym procesie – aplikacja specyficznych algorytmów ekstrakcji wiedzy z danych. Rys. 1 prezentuje ogólny schemat procesu KDD z użyciem algorytmów DM. Zauważmy, że proces ten jest dość skomplikowany - wieloiteracyjny i wieloeksperymentalny.



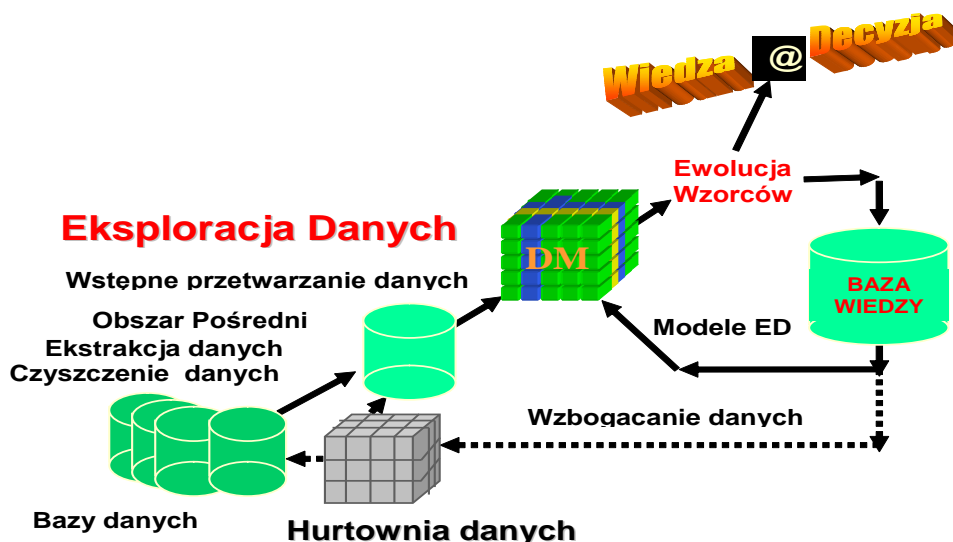
Rys. 1. Schemat procesu KDD z użyciem algorytmów DM

Największym wyzwaniem dla DM jest bardzo duży rozmiar kolekcji danych. Obecne obserwacje pokazują, że rozmiar danych rośnie podobnie jak moc obliczeniowa komputerów – podwaja się co 12 miesięcy. Ma to ogromny wpływ na ewolucję algorytmów DM z powodu tak szybko rosnących ilości danych.

Pod pojęciem eksploracji danych (DM) rozumiemy:

- Ekstrakcję niejawnych, wcześniej nieznanych i potencjalnie użytecznych informacji z danych.
- Ekstrakcję i analizę dużej liczby danych, przy pomocy środków automatycznych lub pół automatycznych w celu odkrycia znaczących wzorców i reguł.
- Technika używana do znajdowania struktur i związków w dużych zbiorach danych.

Eksploracją danych nie jest dedukcyjnym przetwarzaniem zapytań, systemem eksperckim czy standardowym programem statystycznym.



Rys. 2. Schemat procesu KDD z użyciem algorytmów DM i hurtowni danych

Linia procesu KDD z użyciem algorytmów DM i hurtowni danych (DW) składa się z najpierw z wyboru modelu DM (metod (taksonomia) a następnie z ich użycia w DW (rys.102/9/2). Metody DM mają na celu odkrycie wiedzy w hurtowniach danych, która przybiera formę wzorców, związków i faktów, które wcześniej nie były oczywiste. Wybrana metoda DM jest nierozdzielnie związana z wykrytymi wzorcami. Nie oczekuje się, że wszystkie te metody będą działały równie dobrze na wszystkich zbiorach danych. Wizualizacja zbiorów danych może być połączona z, lub użyta przed

modelowaniem i może wspomagać wybór metody oraz wskazywać, jakie wzorce mogą być prezentowane

Wyróżniamy następujące podejście w metodach DM:

- Kierowanie weryfikacją:
 - Analiza zapytań,
 - Analiza statystyczna.
- Nadzorowane kierowanie odkryciami:
 - Predykcja,
 - Klasyfikacja.
- Nienadzorowane kierowanie odkryciami:
 - Sieci neuronowe map samo-organizujących się,
 - Asocjacja,
 - Klasteryzacja,
 - Wykrywanie odchyleń.

Metody kierowania weryfikacją wymagają, aby użytkownik postawił pewną hipotezę a odpowiedzi na zapytania i raportowanie lub analiza statystyczna potwierdzają następnie tę hipotezę. Statystyka w DM jest w pewnym stopniu niedoceniana w porównaniu do mniej tradycyjnych technik takich jak: sieci neuronowe, algorytmy genetyczne i klasyfikacja regułowa. Wiele z tych mniej tradycyjnych technik posiada swoją statystyczną interpretację. Metody statystyczne są najbardziej użyteczne dla zagadnień dobrze ustrukturyzowanych. Wiele problemów DM nie zalicza się do tej klasy - techniki statystyczne załamują się lub wymagają zbyt dużych nakładów, żeby być efektywne.

Nadzorowane kierowanie odkryciami polega na odkryciu związków między wejściami i wyjściami systemu. Związki te mogą być wykorzystywane do predykcji, estymacji lub klasyfikacji. Do uczenia sieci wykorzystywany jest znany zbiór trenujący par wejść/wyjść z dołączonymi etykietami wskazującymi klasę obserwacji, a nowe dane są klasyfikowane w oparciu o niego. Metody nadzorowanego kierowania odkryciami to **predykcja** oraz **klasyfikacja**. Metody predykcyjne budują wzorce przewidując nieznane wartości atrybutów na podstawie znanych wartości innych atrybutów. Główną metodą predykcji jest **regresja** liniowa i wielokrotna lub nie liniowa (**sieci neuronowe z propagacją wsteczną**). **Klasyfikacja** to proces dwuetapowy, który tworzy *konstrukcja modelu* (opisywanie predefiniowanych klas) oraz *użycie modelu do predykcji* (klasyfikacja przyszłych lub nieznanych obiektów). Wyróżniamy **metody klasyfikacji** związane z:

- sieciami neuronowymi z propagacją wsteczną,
- sieciami Bayesowskimi,
- k-najbliższymi sąsiadami
- wnioskowaniem opartym o przypadki,
- algorytmami genetycznymi,
- zbiorami przybliżonymi,
- zbiorami rozmytymi.

Nienadzorowane kierowanie odkryciami. Często zadaniem DM jest odkrycie struktur w zbiorze danych, bez jakiegokolwiek wiedzy wstępnej o nim. Stąd *uczenie nienadzorowane*, gdzie występują nieznane etykiety danych w zbiorze trenującym, a celem jest wykrycie istnienia klas lub klastrów


w danych. Do tego zadania można wykorzystać klasę sieci neuronowe zwane mapami samo-organizującymi się. Nienadzorowane kierowanie odkryciami zapewniają:


- sieci neuronowe: mapy samo-organizujące się,
- asocjacje,
- klasteryzacja,
- wykrywanie odchyleń.

Laboratorium podstawowe

Problem 1 (czas realizacji 25 min)


Ćwiczenie prezentuje sposób stworzenia projektu dla Data Miningu. Krok po kroku opisane zostały definiowanie źródła danych, widoku źródła danych i jego modyfikacji.

Zadanie	Tok postępowania
1. Stworzenie nowego projektu Analysis Services w SS BIDS (SQL Server Business Intelligence Development Studio)	<ul style="list-style-type: none">• Uruchom Business Intelligence Development Studio (BIDS).• Stwórz nowy projekt Analysis Services File->New->Projects-> Analysis Services Project Template.• Wpisz nazwę projektu AdventureWorks.
2. Stworzenie źródła danych (Data Source)	<ul style="list-style-type: none">• W Solution Explorer kliknij ppm na katalogu DataSource i wybierz opcję New Data Source.• Wybierz New aby dodać nowe połączenie do bazy Adventure Works.<ul style="list-style-type: none">• Provider: Native OLE\DB\Microsoft OLE DB Provider for SQL Server• Select or enter a database name: AdventureWorksDW• Server Name: localhost.• Jako sposób logowania wybierz Use the service account• Wybierz Finish i zamknij edytor. <p> Nowe źródło danych, baza danych AdventureWorksDW pojawia się w folderze Data Sources.</p>
3. Stworzenie widoku źródła danych (Data Source View)	<ul style="list-style-type: none">• W Solution Explorer kliknij ppm i wybierz opcję New Data Source View<ul style="list-style-type: none">• Select a Data Source : AdventureWorksDW• Select or enter a database name: AdventureWorksDW• Na stronie Select Tables and Views wybierz następujące tabele:<ul style="list-style-type: none">• dbo.ProspectiveBuyer

	<ul style="list-style-type: none"> • <code>dbo.vAssocSeqLineItems</code> • <code>dbo.vAssocOrders</code> • <code>dbo.vTargetMailing</code> • <code>dbo.vTimeSeries</code> <ul style="list-style-type: none"> • Wybierz Finish i zamknij kreatora <p>Widok źródła danych pozwala modyfikować strukturę danych aby były one bardziej znaczące dla projektu. Używając widoków można: wybrać tabele najistotniejsze dla tworzonego projektu, stworzyć relacje pomiędzy tabelami oraz dodać kolumny obliczeniowe bez potrzeby modyfikowania oryginalnego źródła danych</p>
4. Modyfikacja widoku źródła danych (Data Source View)	<p> Aby stworzyć zaplanowane w ćwiczeniu modele market basket i sequence clustering musisz utworzyć związek jeden-do-wielu pomiędzy tabelami vAssocSeqOrders i vAssocSeqLineItems. Pozwoli to na uczynienie tabeli vAssocSeqLineItems zagnieżdżoną tabelą vAssocSeqOrders na potrzeby w/w modeli.</p> <ul style="list-style-type: none"> • Kliknij podwójnie widok źródła danych Adventure Works DW.dsv • W panelu widoku źródła danych wybierz kolumnę OrderNumber z tabeli vAssocSeqLineItems <ul style="list-style-type: none"> • Przeciągnij kolumnę do tabeli vAssocSeqOrders do kolumny OrderNumber <p>Stworzyłeś nowy związek jeden-do-wielu pomiędzy tabelami vAssocSeqOrders i vAssocSeqLineItems</p>





Problem 2 (czas realizacji 15min) -

Ćwiczenie prezentuje sposób użycia różnych modeli danych na przykładzie fikcyjnej firmy **Adventure Works**. Firma ma zamiar zwiększyć sprzedaż rowerów kierując ofertę do konkretnych klientów za pomocą kampanii pocztowej. Analizując cechy znanych klientów można zauważyć pewne zależności wskazujące potencjalnych nowych klientów, oraz określić prawdopodobieństwo zakupu towarów danej firmy przez potencjalnego klienta. Dodatkowo możliwe jest logiczne grupowanie istniejących klientów, za pomocą zależności np. demograficznych bądź nabywczych. Baza danych **Adventure Works DW** (instalowana wraz z produktem SQL Server 2005 jako przykład) zawiera informacje o istniejących i potencjalnych klientach firmy.

Zadanie	Tok postępowania
1. Stworzenie struktury eksploracji danych dla modelu Targeted Mailing	<p> Pierwszym krokiem jest stworzenie w BIDS nowej struktury eksploracji danych oraz modelu drzewa decyzyjnego.</p> <ul style="list-style-type: none"> • W Solution Explorer kliknij ppm na struktury eksploracji (Mining Structures) i wybierz New Mining Structures. • Jako metodę definicji (Select the Definition Method) wybierz istniejącą relacyjną bazę lub hurtownię danych (From existing relational database or data warehouse) i wybierz Next. • Jako metodę eksploracji danych (Select a Data Mining Technique) wybierz drzewo decyzyjne (Microsoft Decision Trees) i wybierz Next. • Aby określić typy tabel (Specify the Table Types) zaznacz pole Case przy kolumnie vTargetMail i wybierz Next

	<ul style="list-style-type: none"> Aby określić dane treningowe (Specify the Training Data) zaznacz pole Key przy kolumnie CustomerKey. Wybierz pola Input i Predictible przy kolumnie BikeBuyers Po zaznaczeniu, że kolumna jest przewidywana (predictible) uaktywnia się przycisk Suggest. Wybranie przycisku powoduje uruchomienie okna Suggest Related Columns, w którym wyświetlone są kolumny o bliskiej relacji z kolumną przewidywaną. Jako kolumny wejściowe (Input) wybierz następujące kolumny i wybierz Next: <ul style="list-style-type: none"> Age CommuteDistance EnglishEducation EnglishOccupation <ul style="list-style-type: none"> FirstName Gender GeographyKey HouseOwnerFlag LastName MaritalStatus NumberCarsOwned NumberChildrenAtHome Region TotalChildren YearlyIncome Aby wybrać zawartość kolumn i typ danych (Specify Columns' Content and DataType) wybierz Wykryć (Detect) a następnie Next. Na stronie (Split data into training and testing sets), jako Percentage of testing data (procent danych testowych), zostaw domyślną wartość 30. Jako Maximum number of cases in testing data set (maksymalna ilość przypadków w zbiorze danych testowych), wpisz 1000. Wybierz Next <ul style="list-style-type: none"> Mining structure name: TargetedMailing Mining model name : TM_Decision_Tree Zaznacz pole Allow drill through i wybierz Finish
2. Przetwarzanie modelu ekstrakcji danych	<ul style="list-style-type: none"> W BIDS rozwiń menu Mining Models i wybierz Process Jeżeli pokaże się okno o potrzebie konwersji projektu kliknij Yes Kliknij Run i poczekaj na informację o zakończeniu procesu. Wybierz Close.
3. Analiza modelu opartego o drzewa decyzyjne	<ul style="list-style-type: none"> Dwukrotnie kliknij na TargetedMailing.dmm Wybierz zakładkę Mining Models Na tej zakładce możesz sprawdzić jakie kolumny są używane w modelu. Możesz ustawiać kolumny ignorowane, przez co właściwości modelu zmieniają się. Wybierz zakładkę Mining Model Viewer, jeżeli pojawi się okienko informujące o konieczności przebudowania projektu należy kliknąć Yes.





	<p>Na tej zakładce możesz zaobserwować w jaki sposób parametry wejściowe wpływają na parametr wyjściowy modelu. Przy drzewie decyzyjnym możesz ocenić które kolumny mają największy wpływ na ostateczną decyzję o kupnie roweru.</p> <ul style="list-style-type: none"> Wybierz zakładkę Decision Tree. <p>Drzewo składa się z wielu hierarchicznie połączonych węzłów decyzyjnych. W każdym węźle jest warunek jaki musi być spełniony. Po zatrzymaniu kursora myszki nad węzłem, pokazuje się informacja szczegółowa o danej grupie. Po prawej stronie nad diagramem masz możliwość wybrania z listy rozwijanej liczby poziomów jakie są wyświetlane. Jeżeli chcesz wyświetlić węzły potomne dla danego węzła, to musisz podwójnie kliknąć w kwadrat po prawej stronie węzła (jeżeli węzeł jest zwinięty, to w środku jest plus, jeżeli węzeł jest rozwinięty to w środku jest minus).</p>
4. Mapowanie kolumn	<ul style="list-style-type: none"> Wybierz zakładkę Mining Accuracy Chart. Na zakładce Column Mapping sprawdź kolumny w oknie Mining Structure. W oknie Select Input Table(s) naciśnij przycisk Select Case Table i wybierz tabelę vTargetMail. Sprawdź jakie kolumny z TargetMailing wiążą się z kolumnami tabeli vTargetMail.
5. Macierz klasyfikacji	<ul style="list-style-type: none"> Przełączyć na zakładkę Classification Matrix <p>Jak można zinterpretować otrzymaną macierz?</p> <p>W kolumnach są wartości rzeczywiste, natomiast w wierszach wartości wynikające z predykcji. Wartość 7157 w krotce (0,0) oznacza, że w danych testowych tyle przypadków zostało zaklasyfikowanych dobrze do grupy osób nie zainteresowanych kupnem roweru. Wartość 2240 w krotce (1,0) oznacza, że tyle przypadków zostało zaklasyfikowanych do grupy osób które nie są zainteresowane, choć w rzeczywistości jest inaczej. Wartość (0,1) to liczba przypadków zaklasyfikowanych jako osoby zainteresowane kupniem roweru, choć w rzeczywistości tak nie jest. Ostatnią krotką jest (1,1) gdzie mamy przypadek poprawnego zaklasyfikowania osób które chcą kupić rower.</p> <p>Całkowity błąd predykcji można wyznaczyć poprzez podzielenie sumy przypadków leżących we wszystkich krotkach poza główną przekątną do sumy liczb we wszystkich krotkach macierzy.</p> <p>Błąd predykcji dla kupna roweru należy rozumieć jako liczbę osób które nie przyjmą oferty do liczby wszystkich osób do których zostanie wysłana oferta. Ten błąd jest dla nas ważniejszy, ponieważ on decyduje o ostatecznym zysku z kampanii (przykładowo, jeżeli na 100 osób do których zostanie wysłana oferta tylko 5 skorzysta z niej, to zysk ze sprzedaży rowerów może być mniejszy od całkowitego kosztu wysyłania ofert). W naszym przypadku błąd dotyczy tylko tych krotek, dla których wartość predykcji wynosiła 1.</p>


	<p>Jaki jest całkowity błąd predykcji drzewa decyzyjnego, a także jaka część osób odrzuci ofertę z kampanii.</p> <p></p>
6. Wykres przewidywanego zysku	<p> <ul style="list-style-type: none"> Wybierz zakładkę Lift Chart. Z rozwijanej listy Chart Type wybierz Profit Chart. Wpisać następujące wartości: <ul style="list-style-type: none"> Population: 50000 Fixed Cost: 5000 Individual Cost: 10 Revenue per Individual: 15 <p> Jak można zinterpretować otrzymany wykres?</p> <ul style="list-style-type: none"> Zbadaj kształt wykresu dla różnych wartości Individual Cost. (3, 5, 10, 12, 14) – w tym celu naciśnij przycisk Settings i wpisz wybrane wartości. <p>Dlaczego przy ustawieniu Individual Cost powyżej wartości 15 wykres zysku nie posiada wartości dodatnich?</p> </p>
7. Zmiany kolumn w modelu	<ul style="list-style-type: none"> Wybierz zakładkę Mining Models. Dla kolumn First Name, Last Name, Region zaznacz w pozycji rozwijanej wartość Ignore zamiast Input. Przejdź do zakładki Mining Accuracy Chart Classification Matrix. <p> Oblicz na nowo całkowity błąd predykcji i błąd dla kupna roweru. Czy otrzymane wyniki są lepsze czy gorsze od poprzednio uzyskanych?</p>

Laboratorium podstawowe

Problem 2 (czas realizacji 25 min) – kontynuacja modułu 1

Ćwiczenie prezentuje tworzenie różnych modeli dla vTargetMail.

Zadanie	Tok postępowania
1. Algorytm naiwny Bayesowski	<ul style="list-style-type: none"> Przejdź na zakładkę Mining Models, z menu kontekstowego (ppm) wybierz New Mining Model. Jako Model name wpisz TM_Bayes, jako Algorithm name wybrać Naive Bayes. <p> Niektóre typy danych nie są wspierane przez różne metody. Dla wyżej wybranego algorytmu kolumny Age, Geography Key i Yerly Income zostaną pominięte, ponieważ nie mają wartości dyskretnych tylko ciągłe.</p> <ul style="list-style-type: none"> Przy pojawieniu się okienka ostrzegającego wybierz Yes. Z menu kontekstowego dla TM_Bayes wybierz opcję Process. Aby wybrać ten model, możesz kliknąć ppm na pierwszym wierszu w nowo utworzonej kolumnie (czyli na Microsoft Neural Network). Przejdź na zakładkę Mining Model Viewer Attribute Profiles. Sprawdź dla których atrybutów jest największa różnica pomiędzy liczbą osób zainteresowanych kupnem roweru. Przejdź na zakładkę Attribute Characteristics. Sprawdź jakie wartości dla określonych atrybutów wpływają na wynik predykcji. Przejdź na zakładkę Mining Accuracy Chart Classification Matrix. <p> Czy aktualnie stworzony model jest lepszy czy gorszy od drzew decyzyjnych?</p> <p> Zmienić kolumny wejściowe dla tego modelu, czy można w ten sposób poprawić wyniki dla tego algorytmu?</p>
2. Algorytm oparty o Sztuczne Sieci Neuronowe	<ul style="list-style-type: none"> Przejdź na zakładkę Mining Models, z menu kontekstowego (ppm) wybierz New Mining Model. Jako Model name wpisz Neural_Network, jako Algorithm name wybierz Neural Network. <p>Podobnie jak w poprzednim przypadku, niektóre kolumny zostaną automatycznie wyłączone.</p> <p> Jakie kolumny nie mogą być użyte przez sieć neuronową?</p> <ul style="list-style-type: none"> Przetwórz model – z menu kontekstowego dla tego modelu wybierz Process. <p>Zmień kolumny wejściowe dla tego modelu, czy można w ten sposób poprawić wyniki dla tego algorytmu?</p>

	<p>Porównaj opracowane modele. Które kolumny można zignorować?</p> 
--	--

Laboratorium rozszerzone

Zadanie 1 (czas realizacji 30 min)

Otrzymałeś zadanie dokładnej analizy modelu Target Mail dataminingowego opartego o drzewa decyzyjne który jest prezentowany w laboratorium podstawowym. W tym celu musisz najpierw zbudować ten model zgodnie z wskazówkami z laboratorium podstawowego a następnie Wykonać następujące zadania: Na zakładce **Mining Model Viewer** znaleźć po trzy takie ścieżki w drzewie decyzyjnym, które jednoznacznie wskazują (albo prawie jednoznacznie, np. 90%), że osoba z tej grupy kupi lub nie kupi roweru, następnie przejdź zakładkę Dependency Network i zinterpretuj jej zawartość.

Kolejnym twoim zadaniem jest przejść na zakładkę **Mining Accuracy Chart/Lift Chart** i sprawdź jaki wpływ mają wartości **Population** i **Fixed Cost** na końcowy kształt wykresu. Ostatnim twoim zadaniem jest na zakładce **Mining Models** tak dobrać kolumny wejściowe, aby zminimalizować błąd predykcji dla kupna roweru. Czy zmiana kolumn wejściowych znacząco wpływa na błąd predykcji?

Zadanie 1 (czas realizacji 60 min)

Firma Adventure Works w której pracujesz prowadzi szeroko zakrojoną akcję reklamową. Twój przełożeni są bardzo zainteresowani wykorzystaniem modeli dataminingowych aby określić grupę docelową do której należy skierować kampanie ulotkową. Twoim zadaniem jest porównanie wszystkich dostępnych struktur dataminingowych dostępnych dla modelu **Target Mail**. Określ grupę docelową, oraz przeanalizuj, porównaj i skrytykuj struktury dostępne dla modelu **TM**. Które z testowanych struktur są najbardziej odpowiednie dla testowanych danych? W ćwiczeniu skorzystaj z tworzonego na laboratorium podstawowym projektu.