



ANALIZA DANYCH I DATA MINING W CRM

Grzegorz Migut¹

Wstęp

Nieustanne zmiany na rynku, wzrost konkurencji spowodowały, że firmy, chcąc przetrwać na rynku, muszą przystąpić do zmiany strategii marketingowych. Dużą popularność zdobył marketing relacyjny, który oznacza działania polegające na budowaniu indywidualnych, trwałych relacji pomiędzy dostawcą a klientem. W wymiarze operacyjnym filozofia marketingu relacyjnego związana jest z systemami służącymi do zarządzania relacjami z klientem, działającymi według metodyki CRM.

W CRM możemy wyróżnić trzy podstawowe podsystemy: operacyjny, analityczny oraz interaktywny.

CRM operacyjny ma na celu zbieranie danych transakcyjnych i danych o klientach, ich opinii o produktach, sprzedawcach i komunikacji. Zadaniem operacyjnego CRM jest również wsparcie telemarketingu. Do tego celu służą teleinformacyjne systemy automatycznego rozdzielania rozmów przychodzących (*Automatic Call Distribution*) czy też interaktywnej obsługi głosowej (*Interactive Voice Response*).

Zadaniem **CRM analitycznego** jest przetwarzanie i analiza danych, data mining w celu planowania marketingowego, segmentacji i strategii instrumentalnych.

CRM interaktywny ma na celu kształtowanie bezpośrednich kontaktów z klientem i bieżące reagowanie za pomocą różnorodnych kanałów dystrybucji i komunikacji, na pojawiające się wymagania.

CRM Analityczny

W niniejszym artykule skupię się na systemach zaliczanych do CRM analitycznego. Celem tych systemów jest dostarczenie, na podstawie analizy danych, informacji niezbędnych do podejmowania efektywnych działań skierowanych na klienta. Jeśli pragniemy odnieść

¹ StatSoft Polska Sp. z o.o.



sukces podczas wdrażania systemu analitycznego CRM, musimy zadbać o trzy niezwykle ważne aspekty:

Aspekt danych – stosowanie tego typu systemu ma sens jedynie w przypadku, gdy jesteśmy w stanie zapewnić odpowiednią jakość danych – ich dokładność, zgodność itp. W organizacji musi obowiązywać wysoka kultura gromadzenia i przechowywania danych. W aspekcie technicznym sprowadza się to najczęściej do wdrożenia bazy danych służącej celom analitycznym, tak zwanej hurtowni danych. Hurtownia danych może gromadzić dane z całej organizacji, wtedy taki system nazywamy korporacyjną hurtownią danych. Częściej jednak (ze względu na koszty) wdrażane są hurtownie danych o zasięgu departamentu (*data mart*). Spotyka się też opinie, że najlepszym źródłem danych dla CRM analitycznego są dane pochodzące z CRM operacyjnego. Bardzo ważne jest, by utworzona baza analityczna przechowywała dane atomowe, niezagregowane, ponieważ tylko na podstawie takich danych możemy przeprowadzać dogłębne analizy statystyczne i data mining.

Aspekt analizy – zgromadzone przez nas dane zawierają w sobie wiele ukrytych informacji, niedostępnych przy pobieżnej analizie. Ukryte zależności, trendy i wzorce, innymi słowy ukryta wiedza, która może w znacznym stopniu zoptymalizować nasze wysiłki marketingowe, może zostać odkryta dzięki zastosowaniu odpowiednich narzędzi do analizy danych. Narzędzia analityczne tego typu noszą łączne miano narzędzi do zgłębiania danych (*data mining*). Zgłębianie danych polega na intensywnym przeszukiwaniu baz danych w celu klasyfikacji, predykcji, dyskryminacji i odnajdywania powiązań między danymi. Do najczęściej wykorzystywanych metod i technik zgłębiania danych należą sieci neuronowe, analiza skupień, drzewa decyzyjne, analiza koszykowa, *Naive Bayes*, *Support Vector Machine* oraz *MARSplines*.

Aspekt organizacyjny – nawet najlepiej rozwiązany problem gromadzenia i analizy danych nie da efektu, jeśli równocześnie z ich wdrożeniem nie nastąpią zmiany organizacyjne w przedsiębiorstwie. Dobrze skonstruowany model zbudowany na danych o dobrej jakości nie będzie przydatny, jeśli jego wyniki nie będą dostępne dla osób, które ich potrzebują. Dlatego też w parze ze zmianami technologicznymi muszą iść zmiany w kulturze organizacyjnej.

Rodzaje zadań analitycznych

Warto się zastanowić, jakiego typu działania mogą zostać wsparte analizą danych. Patrząc na naszych klientów właśnie pod kątem analizy, możemy to robić na co najmniej dwa różne sposoby:

- ◆ Przekrojowy,
- ◆ Dynamiczny.

Patrząc na informacje dotyczące klientów w **sposób przekrojowy**, chcemy zbadać różnice i podobieństwa pomiędzy poszczególnymi klientami. Nasi klienci zwykle się między sobą



różnią, różne jest ich wykształcenie, zamożność, stan rodzinny itp. A tym samym różne są ich oczekiwania odnośnie naszych produktów. Analiza *data mining* może pomóc zidentyfikować grupy podobnych do siebie klientów, innymi słowy dokonać ich segmentacji. Tego typu analiza może być przydatna po pierwsze, by móc optymalizować nasze działania i bardziej indywidualnie traktować naszych klientów, po drugie może być wstępem do dalszych analiz. Dalsze analizy przeprowadzamy już jedynie na wyodrębnionych segmentach, dzięki czemu możemy uzyskać bardziej dokładne modele, odkryć bardziej szczegółową wiedzę. Możemy wyróżnić kilka rodzajów segmentacji. Pierwszy z nich to segmentacja opisowa, do której zaliczyć można:

- ◆ Segmentację demograficzną - jest przeprowadzana na podstawie danych takich jak: dochód klienta, wiek, płeć, wykształcenie, stan cywilny, ilość osób w rodzinie, status mieszkaniowy, typ mieszkania, grupa etniczna, wyznaniowa itp.
- ◆ Segmentację behawioralną - jest przeprowadzana w oparciu o dane reprezentujące zachowanie klientów. W przypadku sklepu mogą to być informacje na temat częstości zakupów, ilości oraz rodzaju zakupionych produktów.
- ◆ Segmentację pod względem motywacji - opiera się na zmiennych opisujących przyczyny, z powodu których klient dokonał zakupu. Dane tego typu zwykle uzyskiwane są na podstawie badań (np. ankietowych), gdyż z reguły nie jesteśmy w stanie odczytać motywacji klientów na podstawie danych opisujących ich zachowanie.

Drugi rodzaj segmentacji to segmentacja predykcyjna. Jest ona użyteczna do tego, by zrozumieć na przykład, jakie zmienne odróżniają dobrych klientów od złych. By móc przeprowadzić tego typu analizę, powinniśmy na początku określić zmienną określającą „dobrego” klienta - na przykład sumę dokonanych przez niego zakupów. Dla tej zmiennej określamy, jakie pozostałe zmienne mają na nią największy wpływ. Oczywiście należy pamiętać, że segmentacja klientów ma sens jedynie wtedy, gdy zakładamy, że nasze zachowanie będzie różne w odniesieniu do różnych segmentów.

Patrząc na klienta w **sposób dynamiczny**, możemy zauważyć, że nasi klienci podlegają pewnemu cyklowi: niektórzy z potencjalnych klientów, dokonując zakupu, zmieniają się w naszych nowych klientów, ci zaś, jeśli przez pewien czas korzystają z naszych usług, mogą być postrzegani jako nasi stali klienci. Następnie pewna ich część może zrezygnować, stając się naszymi byłymi klientami. Cykl ten możemy określić jako cykl życia klienta. Oczywiście w zależności od etapu cyklu zmieniają się typy analiz, jakie można przeprowadzić.

W odniesieniu do naszych potencjalnych klientów, czyli do osób, które znajdują się na rynku docelowym, lecz nie są naszymi klientami, możemy budować modele służące do optymalizowania kampanii polegającej na wysyłaniu ofert. Model wskazuje osoby, do których warto skierować ofertę. Budowa tego typu modeli jest dosyć trudna, ponieważ z reguły opieramy się na danych zewnętrznych (osoby nie są naszymi klientami).



Kolejna grupa (faza cyklu życia klienta) to osoby, które odpowiedziały na naszą ofertę i są nią zainteresowane, ewentualnie dokonały zakupu po raz pierwszy. Modele budowane dla tej grupy klientów mają na celu wskazanie, czy nowy klient będzie wysoce dochodowy, przeciętny czy też będzie on przynosił niski dochód. Tego typu analiza jest możliwa z tego względu, że bardzo dobrym predyktorem dalszego zachowania klienta jest jego pierwszy zakup, niejako inicjujący kontakty z nami.

Kolejny etap to osoby, które są naszymi stałymi klientami. Naszym celem jest spowodowanie, by byli oni jak najbardziej dochodowi. Możemy to próbować robić na wiele różnych sposobów, w szczególności sprawić, by:

- ◆ kupowali oni więcej towarów, usług – można zbudować model wskazujący, jacy klienci byliby skłonni w większym stopniu korzystać z naszej oferty, ewentualnie którym warto zaproponować rozszerzoną wersję produktu (*up-selling*),
- ◆ kupowali szerszy asortyment - budowane są modele wskazujące, którym z klientów można zaproponować dodatkowe towary lub usługi (*cross-selling*),
- ◆ dłużej byli naszymi klientami – przeprowadzane są analizy mające na celu wskazanie, którzy klienci mogą odejść, co może dać szansę dostawcy na podjęcie odpowiednich kroków, mających na celu zatrzymanie ich (*churn*).

Ostatni etap życia klienta to grupa naszych byłych klientów, w odniesieniu do których możemy stosować działania mające ich skłonić, by ponownie zostali naszymi klientami. Budowane modele mogą nam pomóc w tego typu działaniach (*winback campaign*).

Warto zauważyć, że im późniejszy etap cyklu życia klienta, tym bardziej pełne i rzetelne są analizy dla niego przeprowadzane. Wynika to z faktu, iż dysponujemy coraz większym zestawem informacji o kliencie. W przypadku potencjalnych klientów w najlepszym wypadku dysponujemy danymi demograficznymi, w późniejszych fazach dochodzi cały szereg informacji związanych z zachowaniem klienta, które zwykle są znacznie lepszymi predyktorami (jeśli podczas budowy modelu dysponujemy zarówno zmiennymi demograficznymi jak i opisującymi zachowanie klientów, to zwykle zmienne demograficzne rzadziej są uznawane za istotne).

Metody data mining w CRM analitycznym

Wszystkie przedstawione powyżej rodzaje analiz w praktyce wykonuje się za pomocą technik *data mining*. Pod tą nazwą kryje się w rzeczywistości szereg różnego rodzaju metod analitycznych. Ich zastosowanie zwykle się różni, dlatego też wprowadzono szereg klasyfikacji porządkujących te metody. Najbardziej odpowiedni z punktu widzenia przedstawionych zastosowań wydaje się podział na ukierunkowany *data mining* (uczenie z nauczycielem) oraz nieukierunkowany *data mining* (uczenie bez nauczyciela).

Metody ukierunkowanego *data mining* są wykorzystywane w przypadkach, gdy interesująca nas klasyfikacja została już zaobserwowana i zapisana w próbie uczącej. Wykorzystując te dane, budujemy model, który ma na celu przewidzieć tę klasyfikację dla



nowych danych. Mamy na przykład zbiór zawierający pewne dane o klientach oraz informację, czy dany klient odpowiedział na naszą specjalną promocję (mamy więc podział na klientów, którzy odpowiedzieli, i tych, którzy nie odpowiedzieli). Na podstawie zgromadzonych danych budujemy model, którego celem jest postawienie prognozy, kto z nowej grupy klientów (nieuczestniczących w starej promocji) byłby skłonny odpowiedzieć na podobną akcję promocyjną. Tego typu modele mogą znacznie obniżyć koszty naszych działań, nie obniżając naszej skuteczności, ponieważ nasze wysiłki kierujemy jedynie do osób, wobec których istnieje szansa, że na nie odpowiedzą. Techniki ukierunkowanego *data mining* stosujemy w segmentacji predykcyjnej, jak również we wszystkich analizach odnoszących się do cyklu życia klienta. Do metod ukierunkowanego *data mining* zalicza się: drzewa decyzyjne, sieci neuronowe, *MARSplines*, *Naive Bayes* oraz *Support Vector Machine*

W sytuacji nieukierunkowanego *data mining* sytuacja jest inna. Posiadamy pewne dane o kliencie, które jednak nie zawierają żadnego podziału narzuconego z góry. Naszym celem jest wyjaśnienie struktury tych danych, na przykład określenie skupisk czy zależności niewidocznych na pierwszy rzut oka. Możemy przykładowo dysponować bazą danych klientów zawierającą różnego typu zmienne, które potencjalnie mogą mieć duży wpływ na zachowanie klientów. Celem analizy może być odnalezienie segmentów rynku, czyli grup osób podobnych do siebie. Działania wobec naszych klientów możemy następnie uzależnić od segmentu, w jakim się znajdują. Do technik nieukierunkowanego *data mining* zaliczyć można: analizę skupień, analizę koszykową, sieci neuronowe oraz analizę czynnikową i składowych głównych

Przykład segmentacji klientów

Jako przykład wykorzystania analiz typu *data mining* do zarządzania relacjami z klientami zaprezentowany zostanie przykład segmentacji demograficznej za pomocą technik analizy skupień. Dane będące podstawą analizy są częścią zbioru danych powstałego w wyniku badań przeprowadzonych wśród klientów jednego z centrów handlowych w USA. Zawierają one 8995 przypadków zgromadzonych w arkuszu programu *STATISTICA* o nazwie *Marketing.sta*. Każdy przypadek zawiera 14 zmiennych, są to zmienne demograficzne, takie jak:

- ◆ *dochód* - roczny dochód klienta,
- ◆ *pleć* - płeć klienta,
- ◆ *stan cywilny* - stan cywilny klienta,
- ◆ *wiek* - wiek klienta,
- ◆ *wykształcenie* - wykształcenie klienta,
- ◆ *zawód* - zawód wykonywany przez klienta,



- ◆ *czas zamieszkania* - od jak dawna klient mieszka w danym miejscu,
- ◆ *podwójny dochód* - czy drugi członek rodziny pracuje,
- ◆ *członkowie rodziny* - ilość osób w rodzinie,
- ◆ *dzieci* - ilość dzieci w rodzinie,
- ◆ *status* - sposób utrzymywania się,
- ◆ *dom* - rodzaj domu, w jakim mieszka klient,
- ◆ *grupa* - grupa etniczna, do jakiej należy klient,
- ◆ *język* - język, jakim na co dzień posługuje się klient.

Wszystkie te zmienne są zmiennymi skategoryzowanymi.

Celem analizy jest wykrycie ewentualnych segmentów rynku oraz określenie cech osób należących do poszczególnych grup. Na tej podstawie można rozwijać określone działania marketingowe skierowane do poszczególnych grup. Wyniki analizy mogą także powiększyć wiedzę na temat naszych klientów. Ponieważ zgromadzone dane nie zawierają żadnego podziału znanego z góry, innymi słowy w danych nie możemy wyróżnić zmiennej zależnej, dlatego też do analizy zostanie wykorzystana metoda należąca do grupy metod nieukierunkowanego *data mining*: analiza skupień.

Wykorzystane metody

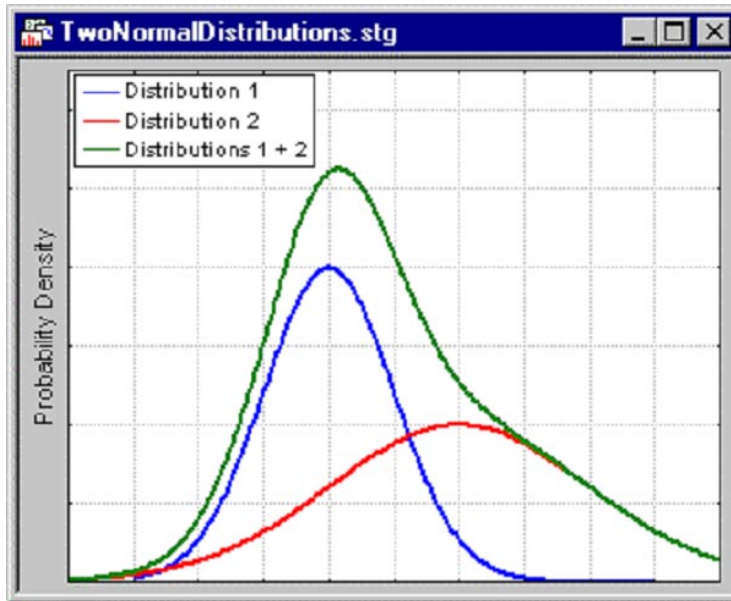
Celem **analizy skupień** (*cluster analysis*) jest wyodrębnienie ze zbioru danych obiektów, które byłyby podobne do siebie, i łączenie ich w grupy. W wyniku działania tej analizy z jednego niejednorodnego zbioru danych otrzymujemy grupę kilku jednorodnych zbiorów. Obiekty znajdujące się w tym samym zbiorze uznawane są za „podobne do siebie”, obiekty z różnych zbiorów traktowane są jako „niepodobne”. Pojęcie analizy skupień obejmuje faktycznie kilka różnych algorytmów klasyfikacji. Do najważniejszych należy zaliczyć metodę *k*-średnich oraz EM.

Stosowanie metody *k*-średnich wymaga od nas podania liczby grup, na które zostanie podzielony wejściowy zbiór danych. Jedną z wersji tej metody polega na losowym wyborze *k* obiektów z analizowanego zbioru i uznania ich za środki *k* grup. Każdy z pozostałych obiektów jest przypisywany do grupy o najbliższym mu środku. Następnie oblicza się nowe środki każdej podgrupy na podstawie średnich arytmetycznych ze współrzędnych zawartych w nich obiektów. W kolejnym kroku następuje przegrupowanie elementów grup, każdy obiekt jest przesuwany do tej grupy, do której środka ma najbliżej. Procedurę tę powtarzamy do momentu, gdy w danej iteracji żaden z obiektów nie zmieni swojej podgrupy. Pewną wadą tej metody jest konieczność odgórnego określenia liczby skupień występujących w danych, dlatego też zaleca się powtórzenie procedury dla różnych wartości *k* i wybranie tej, dla której zbiór danych jest podzielony najlepiej.

Metoda EM jest czasem nazywana analizą skupień bazującą na prawdopodobieństwie lub statystyczną analizą skupień. Program wyznacza skupienia, zakładając różnorodne



rozkłady prawdopodobieństwa zmiennych uwzględnianych w analizie. Na początku działania algorytmu, podobnie jak w metodzie k-średnich musimy podać liczbę skupień, jakie powinny być wyodrębnione ze zbioru wejściowego.



Załóżmy, że przeprowadziliśmy badania w pewnej dużej zbiorowości pod kątem jednej cechy ciągłej. Zaobserwowany rozkład tej cechy był zgodny z funkcją gęstości opisaną kolorem zielonym (*Distributions 1+2*) charakteryzującą się pewną średnią oraz odchyleniem standardowym. Wiemy też, że w zbiorowości tej występują dwa segmenty (na przykład kobiety i mężczyźni) o różnych parametrach funkcji gęstości w swoich segmentach. Algorytm EM ma na celu określenie parametrów rozkładów segmentów na podstawie rozkładu całej grupy oraz przydzielenie poszczególnych obserwacji do najbardziej odpowiadających im segmentów (klasyfikacja następuje na zasadzie prawdopodobieństwa). Na naszym rysunku rozkłady dwóch segmentów zostały opisane kolorem niebieskim (*Distribution 1*) oraz czerwonym (*Distribution 2*), które po zsumowaniu dają funkcję rozkładu całej zbiorowości (*Distributions 1+2*). Algorytm EM dokonuje klasyfikacji nie tylko przy założeniu normalności rozkładu, jak to zaprezentowano na rysunku, wykorzystując go można również określić inną funkcję gęstości dla badanej cechy (badanych cech).

Proces analizy

Wszystkie analizy zostaną przeprowadzone w środowisku *STATISTICA Data Miner*. Najwygodniej jest przeprowadzić je w specjalnie zaprojektowanej przestrzeni roboczej. Umieszczamy w niej arkusz wejściowy *Marketing.sta* i wybieramy zmienne zgodnie

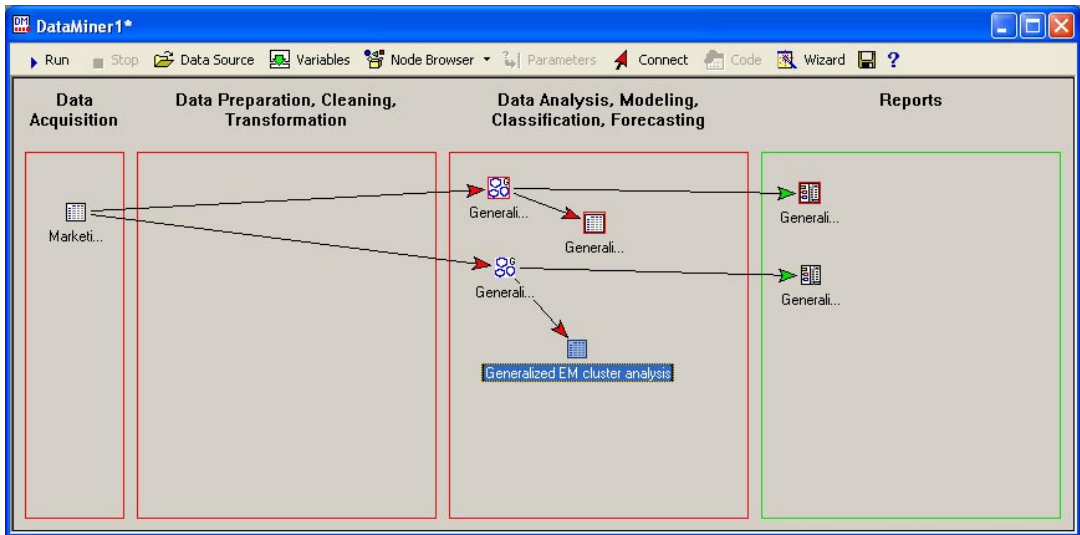


z zamieszczonym powyżej opisem danych. Jako narzędzia do analizy wybierzemy uogólnioną metodę EM (*Generalized EM cluster analysis*) oraz uogólnioną metodę k-średnich (*Generalized K-Means cluster analysis*). Wybieramy przeglądarkę węzłów, a następnie odnajdujemy interesujące nas procedury, dwukrotnie klikając na każdej z nich w celu umieszczenia ich w przestrzeni roboczej.

Dużym mankamentem obu metod jest fakt, iż musimy na wstępie analizy określić liczbę segmentów, na które algorytm ma podzielić zbiór wejściowy. W praktyce jednak zwykle nie znamy tej liczby, pragniemy dopiero odnaleźć najlepsze rozwiązanie. Niedogodność tą można ominąć stosując sprawdzian krzyżowy zaimplementowany w obu metodach. Dzięki sprawdzianowi krzyżowemu można wyznaczyć i ocenić najlepszy układ skupień - program automatycznie określa najbardziej odpowiednią liczbę skupień (segmentów). Dlatego też dla obu metod zaznaczamy opcję sprawdzianu krzyżowego (*V-Fold Crossvalidation*).

Warto zauważyć, że obie metody (podobnie jak inne zawierające w swojej nazwie *with deployment*) umożliwiają zapisanie zbudowanego modelu w postaci kodu C i PMML oraz *STATISTICA Visual Basic*.

Obydwa wybrane przez nas moduły podłączamy do wejściowego arkusza danych, a następnie uruchamiamy proces analizy, wybierając polecenie *Run*. Po wykonaniu analizy jej wyniki zostaną umieszczone w węzłach wynikowych utworzonych dla każdej metody.





Analiza uzyskanych wyników

Dla każdej metody wygenerowane zostały dwa węzły wynikowe (widoczne na rysunku powyżej). Pierwszy z nich zawiera pierwotny arkusz danych z nową zmienną określającą, do którego segmentu przydzieleni zostali poszczególni klienci. Drugi węzeł zawiera raporty przedstawiające szczegółowe wyniki analizy. Analizując te wyniki możemy zauważyć, że w obu przypadkach algorytm zidentyfikował trzy segmenty. Poniżej zamieszczono raport wyników dla metody EM.

Summary for EM clustering (Marketing2)	
Number of clusters: 3	
Total number of training cases: 8993	
Algorithm	EM
Continuous distribution	
MD casewise deletion	Yes
Cross-validation	5 folds
Testing sample	0
Training cases	8993
Training log-likelihood	-16,488859
Number of clusters	3

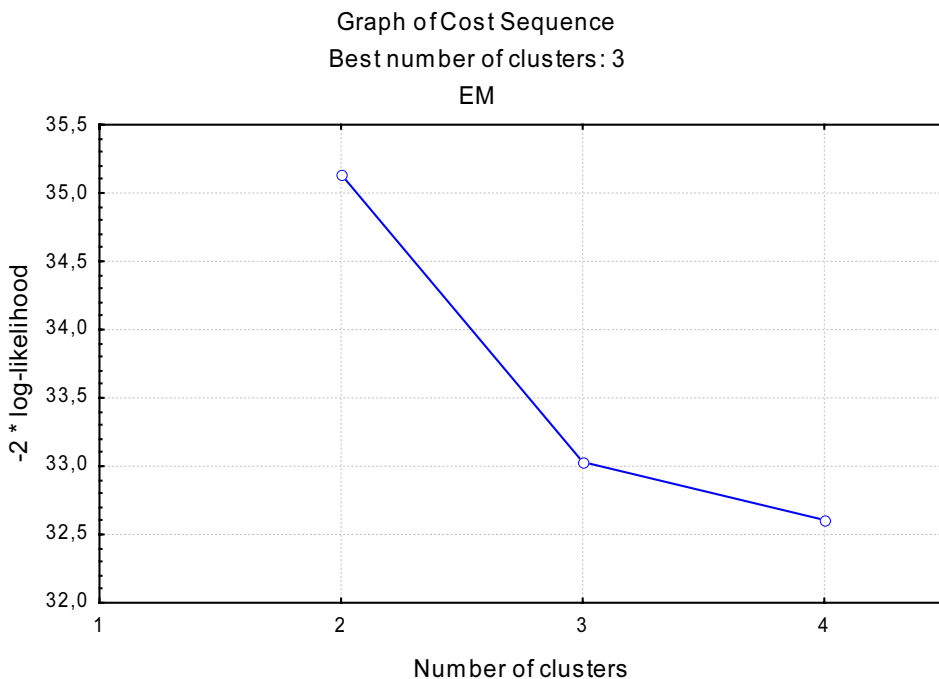
W przypadku metody k-średnich w węźle raportu znajduje się arkusz wskazujący odległość poszczególnych klientów od środka segmentu, w jakim się znajdują. W przypadku metody EM podane jest prawdopodobieństwo przynależności do poszczególnych segmentów. Poniższy rysunek przedstawia wynik klasyfikacji oraz poziom prawdopodobieństwa dla pierwszych ośmiu klientów, w przypadku wykorzystania metody EM.

Classification probabilities (weights) for EM clustering (Marketing2)				
Number of clusters: 3				
Total number of training cases: 8993				
	Cluster 1	Cluster 2	Cluster 3	Final classification
1	0,000001	0,999999	0,000000	2
2	0,000000	1,000000	0,000000	2
3	0,000159	0,999841	0,000000	2
4	0,000000	0,000000	1,000000	3
5	0,000000	0,000000	1,000000	3
6	0,000006	0,999994	0,000000	2
7	0,582448	0,000000	0,417552	1
8	1,000000	0,000000	0,000000	1



W węzle raportowym znajduje się również tak zwany wykres osypiska przedstawiający wynik działania mechanizmu walidacji krzyżowej. Wykorzystując test krzyżowy, dzielimy zbiór wejściowy kolejno na coraz większą liczbę segmentów, a następnie obserwujemy, jaka jest precyzja podziału dla każdego z nich. Dla metody *k-średnich* miarą precyzji podziału będzie przeciętna odległość elementów zbioru wejściowego od środka segmentu, w jakim się znajdują, w przypadku EM będzie to pewna miara oparta na prawdopodobieństwie obliczonym dla poszczególnych obserwacji.

Analizując wykres możemy zauważyć znaczną poprawę precyzji podziału przy zwiększeniu liczby segmentów z dwóch do trzech. Dodając jeszcze jeden segment uzyskujemy jedynie nieznaczną poprawę precyzji, stąd za optymalną liczbę segmentów należy uznać trzy.



W obrębie wykorzystywanej przez nas przestrzeni roboczej możemy łatwo przeprowadzić dodatkowe analizy, po prostu wybierając interesującą nas metodę z przeglądarki węzłów i podłączając ją do analizowanego zbioru danych.

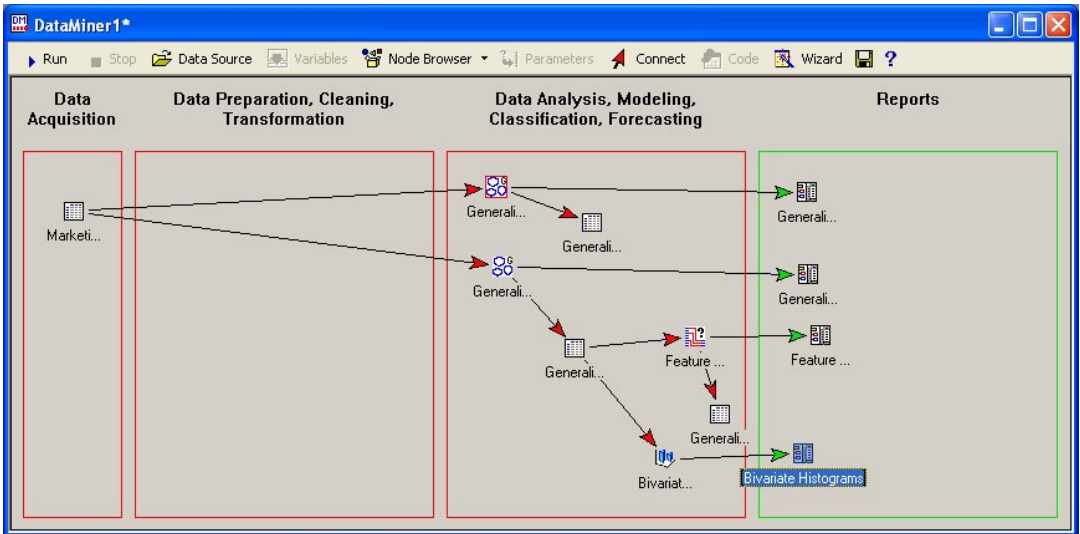
Na podstawie arkusza wyników uzyskanych dla metody EM zbadamy, jakie zmienne miały największy wpływ na proces podziału. W tym celu z przeglądarki węzłów wybieramy węzeł *Feature Selection and Root Cause Analysis*. W węzle tym zaznaczamy dodatkowo opcję nakazującą wygenerowanie wszystkich możliwych raportów.



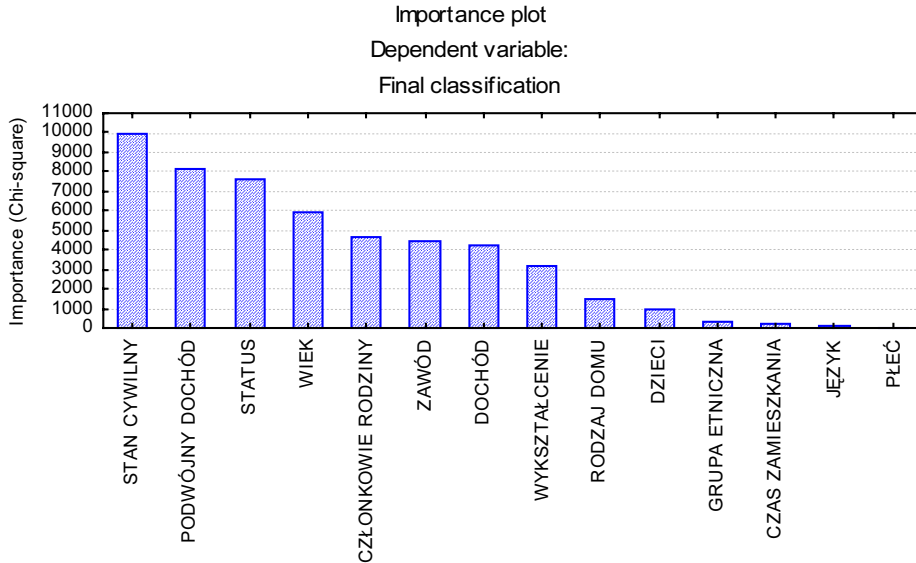
Zobaczymy też, jak wygląda rozkład poszczególnych zmiennych we wszystkich segmentach. W tym celu możemy użyć na przykład jednej z wielu metod wizualizacji danych - histogramów dwóch zmiennych 3W (*Bivariate Histograms*).

Oba węzły umieszczamy w przestrzeni roboczej i łączymy z arkuszem danych powstałym w wyniku analizy skupień metodą EM. W arkuszu tym musimy jeszcze określić zmienne, jakie będą analizowane. Skategoryzowaną zmienną zależną będzie zmienna określająca segment, do którego należą poszczególni klienci. Wszystkie zmienne, którymi dysponowaliśmy pierwotnie określamy jako skategoryzowane zmienne zależne. Ponieważ są to zmienne skategoryzowane, musimy określić dla nich kody – program robi to w sposób automatyczny po naciśnięciu przycisku *Codes*, a następnie *Select All*.

Proces analizy uruchamiamy wybierając polecenie *Run Dirty Nodes* z menu *Run* lub z menu podręcznego przestrzeni roboczej *STATISTICA Data Miner*, co spowoduje uruchomienie analiz tylko dla nowo dodanych lub zmienionych węzłów. Rysunek poniżej przedstawia przestrzeń roboczą już po przeprowadzonej analizie.

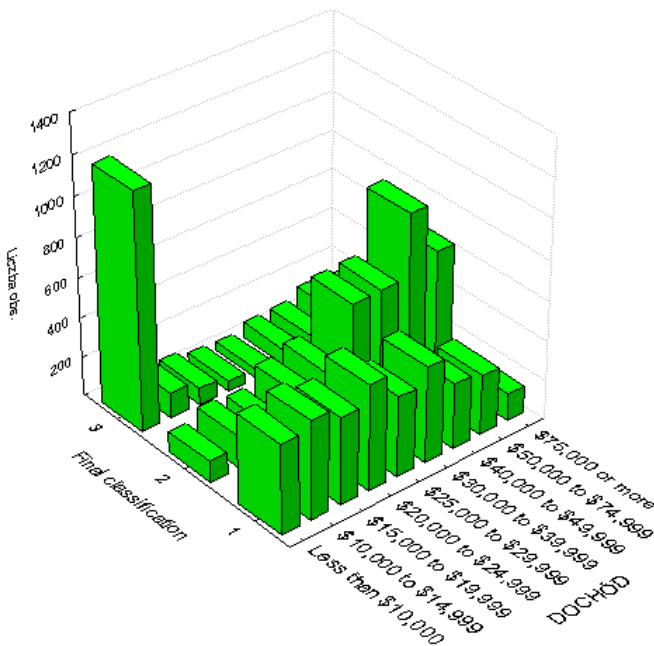


Analizę wyników rozpoczniemy od węzła *Feature Selection*. Poniższy histogram opisuje wpływ poszczególnych zmiennych na proces segmentacji. Można zauważyć, że największy wpływ na proces podziału miały zmienne *stan cywilny*, *podwójny dochód*, *status* oraz *wiek*, najmniejszy zmienne *pleć* oraz *język*.



W wyniku działania węzła *Bivariate Histograms* otrzymano szereg wykresów przedstawiających rozkład poszczególnych zmiennych biorących udział w analizie względem wyodrębnionych segmentów. Poniższy wykres przedstawia rozkład zmiennej *dochód* w poszczególnych segmentach.

Histogram dwu zmiennych (Generalized EM cluster analysis 16v*8993c)





Można zauważyć, że do pierwszego segmentu należą osoby o średnim dochodzie. Najwięcej osób najbogatszych znalazło się w drugim segmencie. Trzeci segment zdominowany został przez osoby z najniższym dochodem.

Oczywiście analizę wyników przeprowadzonej segmentacji możemy przeprowadzać za pomocą szeregu innych dostępnych metod analitycznych. Warta uwagi jest też możliwość zapisania całego projektu *Data Mining* w pliku i udostępnienie go innym użytkownikom *STATISTICA Data Miner*. Jeśli projekt ten umieścimy w repozytorium *Web STATISTICA*, to będzie można z niego korzystać i udoskonalać go w środowisku internetowym.

Literatura:

1. Berry M., Gordon L., *Mastering Data Mining. The Art and Science of Customer Relationship Management*, John Wiley & Sons, Inc, New York 2000.
2. Berson A., Smith S., Thearling K., *Building Data Mining Applications for CRM*, McGraw-Hill, New York 2000.
3. Coppock D.S. *Market Segmentation and "Best Customer"* DM Review www.dmreview.com/editorial/dmreview.