



## WYKORZYSTANIE SKORINGU MARKETINGOWEGO DO OPTIMALIZACJI KAMPANII SPRZEDAŻOWYCH

*Grzegorz Migut, StatSoft Polska Sp. z o.o.*

Znajomość wzorców zachowania klientów oraz czynników, jakie na nie wpływają, jest jednym z krytycznych warunków sukcesu każdej kampanii sprzedażowej. Bardzo pomocne w poznawaniu klientów są narzędzia służące do zgłębiania danych (data mining). Wykorzystanie tych technik jest szczególnie warte polecenia w sytuacji, gdy dysponujemy dużą liczbą cech każdego z klientów, takich jak: dane demograficzne klienta, historia jego transakcji itp. Dzięki analizie tych danych możemy odkryć ukryte, nieznane wcześniej zależności oraz zidentyfikować reguły zachowań klientów niemożliwe do wykrycia w inny sposób.

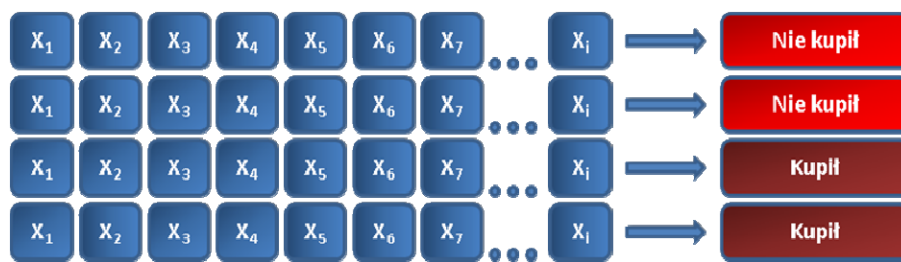
Jednym z najczęściej stosowanych podejść do optymalizacji kampanii marketingowych jest stworzenie modelu wskazującego klientów, do których warto skierować ofertę. Model taki tworzymy za pomocą technik zgłębiania danych na podstawie danych o klientach. Celem budowy modelu jest określenie, jaki produkt powinien zostać zaoferowany konkretnemu klientowi i jaki jest najlepszy kanał komunikacji z klientem.

Modele te określamy mianem modeli skoringowych, ponieważ rezultatem ich działania jest ocena (*scoring*) szansy zakupu przez danego klienta określonego produktu. Ocena ta może zostać wyrażona w formie prawdopodobieństwa bądź punktacji – im wyższa ocena, tym większa skłonność klienta do zakupu.<sup>24</sup>

Modele skoringowe budowane są na podstawie zachowań innych klientów w przeszłości. Wykorzystując dane historyczne zawierające cechy naszych bądź innych klientów (mogą to być zarówno cechy demograficzne, jak i behawioralne) oraz zmienną informującą o fakcie zakupu interesującego nas produktu, model określa wzorce zachowań klientów. Jeśli wzorce wychwycone przez model okażą się wartościowe, możemy je następnie zastosować dla nowych klientów. Model wskaże najbardziej odpowiednią grupę docelową planowanej kampanii (osoby z największą skłonnością do zakupu określonego produktu). Ogólny schemat budowy tego typu modeli przedstawia poniższy rysunek.

---

<sup>24</sup> Modele skoringowe są wykorzystywane również do szeregu innych zadań, takich jak: przewidywanie odejść klientów, wykrywanie nadużyć czy ocena wiarygodności kredytowej.



- Identyfikacja wzorców
- Budowa modeli
- Przewidywanie



W niniejszym artykule zaprezentowany zostanie przykład budowy modelu skoringowego przy użyciu regresji logistycznej oraz drzew wzmacnianych. Następnie modele te ocenimy pod kątem ich zdolności do przewidywania zachowania klientów i określimy optymalny punkt odcięcia dla lepszego z nich.

## Budowa modelu skoringowego

Przykład budowy modelu skoringowego przewidującego skłonność klientów do zakupu zaprezentujemy na podstawie nieco zmienionego zbioru CREDIT dostępnego z podręcznikiem [3]. Dane zawierają informacje o potencjalnych klientach (w większości są to różnego rodzaju wskaźniki opisujące aktywność klientów) wraz z informacją, czy klient dokonał zakupu karty kredytowej. Naszym zadaniem jest stworzenie modelu, który na podstawie cech klientów będzie w stanie przewidzieć ich odpowiedź na ofertę. Interesuje nas nie tylko samo przewidywanie decyzji klientów, ale również wiedza dotycząca czynników najmocniej wpływających na odpowiedź na ofertę oraz wzajemnych związków między zmiennymi; innymi słowy, chcemy wychwycić wzorce zachowań klientów.

Dysponujemy danymi o 13 996 osobach, którym zaproponowano kartę kredytową. W zbiorze znajduje się 39 zmiennych (cech potencjalnych klientów), na podstawie których będziemy chcieli przewidywać odpowiedź na ofertę. Zmienne te są predyktorami w naszej analizie. Zmienną zależną jest zmienna *Buyer* przyjmująca dwie wartości: 'T' (klient zakupił kartę) i 'N' (negatywna odpowiedź na ofertę).

### *Wstępna analiza danych*

Przed przystąpieniem do zasadniczej części analizy konieczne jest bliższe zapoznanie się z analizowanymi danymi w celu określenia ich charakteru, skali pomiaru oraz rozkładów



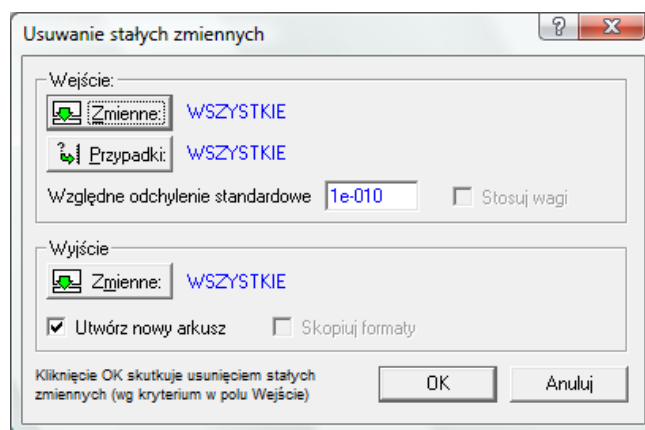
poszczególnych zmiennych, a także występowania w zbiorze danych błędów i problemów koniecznych do usunięcia przed etapem analizy.

Wstępna analiza zbioru danych została wnikliwie zaprezentowana w [1], w niniejszym artykule ograniczymy się do kilku aspektów szczególnie istotnych w kontekście budowy modeli skoringowych. Na tym etapie analizy przedmiotem naszego zainteresowania będzie:

- ◆ usunięcie ze zbioru danych cech niewykazujących zmienności,
- ◆ obsługa braków danych,
- ◆ eliminacja zmiennych nadmiernie skorelowanych z innymi zmiennymi (wejściowymi)
- ◆ eliminacja zmiennych, które nieistotnie wpływają na skłonność do zakupu karty kredytowej,
- ◆ dyskretyzacja zmiennych – podział zmiennych na jednorodne kategorie z punktu widzenia szansy zakupu.

Zmienne niewykazujące zmienności często występują w analizowanych zbiorach danych. Ich obecność może wynikać z analizy grupy jednorodnej pod względem danego czynnika (np. analizujemy jedynie mężczyzn, *pleć* będzie więc wartością stałą), bądź też braku dostatecznej pielęgnacji bazy danych i występowania w niej kolumn wypełnianych zawsze domyślnymi wartościami. Oczywiście zmienne (stałe) nie wnoszą żadnej informacji do modelu, w związku z tym zasadne jest ich usunięcie.

By usunąć stałe zmienne, z menu *Dane* wybieramy opcję *Czyszczenie danych*, a następnie *Usuń stałe zmienne*.



W wyświetlonym oknie *Usuwanie stałych zmiennych* wybieramy wszystkie zmienne i po naciśnięciu *OK* otrzymujemy arkusz, w którym usunięte zostały zmienne niewykazujące zmienności.

Bardzo częstym problemem występującym w analizowanych zbiorach są braki danych. Ponieważ występują one także w naszym zbiorze, przed przystąpieniem do kolejnych punktów wstępnej analizy musimy jeszcze rozwiązać problem ich występowania i określić optymalny sposób ich obsługi. Aby ocenić skalę występowania braków danych, skorzystamy ze statystyk opisowych. Z menu *Statystyka* wybieramy *Statystyki podstawowe*



i tabelę, a następnie opcję *Statystyki opisowe*. Po wybraniu wszystkich zmiennych na karcie *Więcej* wybieramy opcję *%Ważnych* i zatwierdzamy wykonanie analizy.

	% Ważnych
EQLIMIT	0,91
EQBAL	0,91
EQHIGHBAL	1,98
EQCURBAL	1,98
DOB_MONTH	8,14
UNSECLIMIT	43,71
UNSECBAL	43,71
ICURBAL	52,09
IHIGHBAL	52,10
MTHIGHBAL	55,17
MTCURBAL	55,17
DOB_YEAR	67,51
BCLIMIT	82,03
BCBAL	82,03
YEARS_RES	95,82
LST_R_OPEN	97,71
RBAL	97,71
RLIMIT	97,71
TBALNO	99,35
RBALNO	99,99

W powyższej tabeli widzimy fragment wyników dotyczący zmiennych z brakującymi danymi. Możemy zauważyć, że cztery pierwsze zmienne *EQLIMIT*, *EQBAL*, *EQHIGHBAL*, *EQCURBAL* są wypełnione w bardzo niewielkim stopniu (poniżej 5%) dlatego też usuniemy je ze zbioru danych.<sup>25</sup>

Z kolei sześć ostatnich zmiennych ma odsetek braków danych nie większy niż 5%. Ponieważ odsetek braków danych jest stosunkowo niewielki, zastąpienie ich odpowiednią stałą wartością (w naszym przypadku będzie to mediana) jedynie w niewielkim stopniu wpłynie na zmianę rzeczywistego rozkładu tych zmiennych.

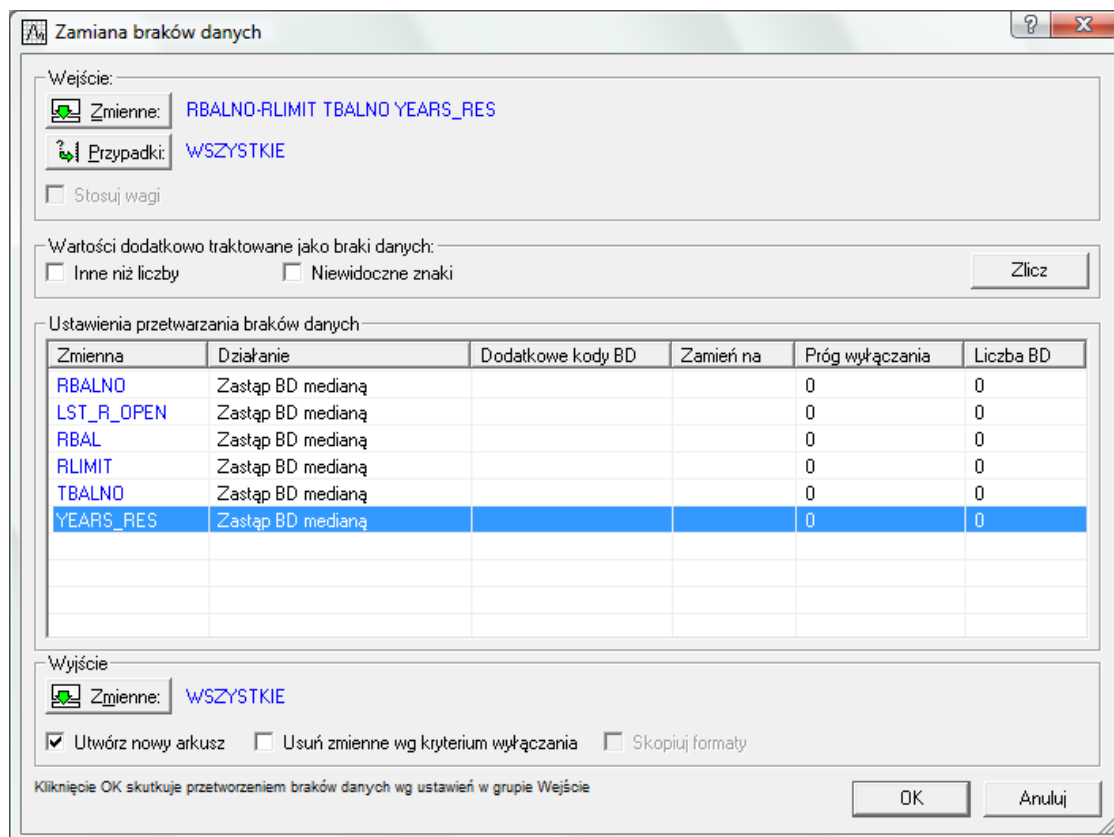
W przypadku pozostałych zmiennych, których wartości wypełnione są od 8,14% do 82,03%, ryzykownym byłoby zastępowanie braków danych średnią bądź medianą ze względu na ryzyko znaczącego zniekształcenia rozkładów analizowanych zmiennych. By zastąpić braki tych zmiennych, powinniśmy skorzystać z bardziej wyrafinowanych metod imputacji braków danych – na przykład wybierając metodę k-najbliższych sąsiadów bądź też przeprowadzić dyskretyzację tych zmiennych, definiując brak danych jako odrębną kategorię. Ponieważ w dalszej części analizy wykonamy dyskretyzację zmiennych, braki danych zastąpimy wartością -1, która jest wartością spoza zakresu zmienności wszystkich zmiennych.

Po usunięciu zmiennych, w których braki danych stanowiły ponad 95% przypadków, zajmujemy się grupą zmiennych o znikomym odsetku braków danych. Za pomocą opcji *Zamiana braków danych* z menu *Dane -> Czyszczenie danych* zamienimy braki danych

<sup>25</sup> W sytuacji, gdy nasz zbiór danych zawiera znaczną liczbę tego typu cech, możemy pokusić się o analizę tych zmiennych, przygotowując jedną bądź kilka zmiennych pochodnych, zawierających kombinację wartości zmiennych pierwotnych. Więcej na temat analizy tego typu danych można znaleźć w [3].



odpowiednich zmiennych medianą (w analogiczny sposób postąpimy ze zmiennymi o znacznym odsetku braków, które zamienimy stałą wartością).



Kolejne kroki analizy wykonamy w *Zestawie Skoringowym STATISTICA*, narzędziu przygotowanym specjalnie w celu optymalizacji procesu budowy, oceny i monitorowania modeli skoringowych.<sup>26</sup> W pierwszej kolejności użyjemy modułu *Wybór predyktorów*, który pozwoli nam wyróżnić w zbiorze danych wiązki zmiennych o podobnej zmienności, jednocześnie pozwalając wyeliminować ze zbioru danych zmienne nadmiernie skorelowane z innymi predyktorami. W kolejnym kroku wyeliminujemy zmienne nieistotnie wpływające na skłonność do zakupu karty.

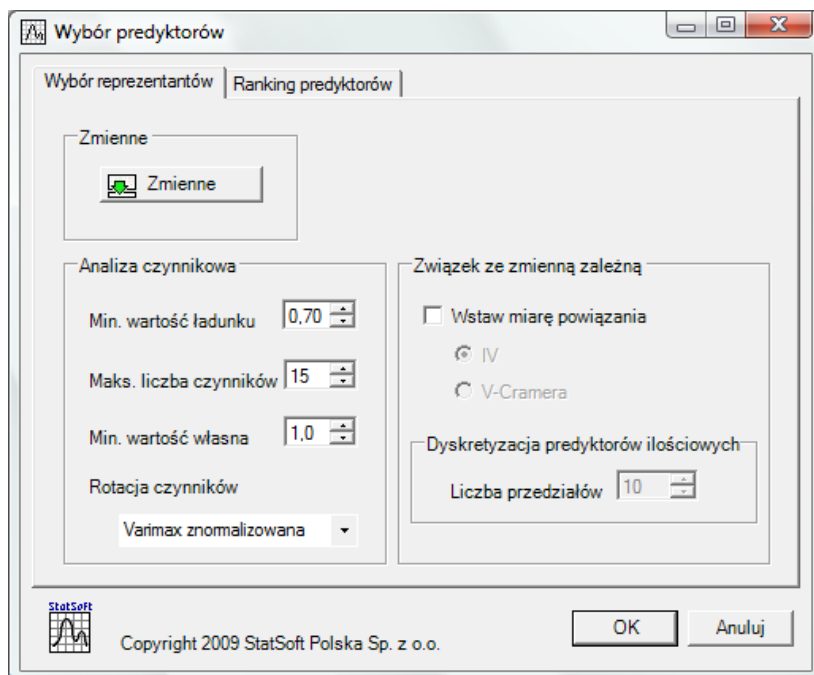
Z menu *Zestaw Skoringowy* wybieramy opcję *Wybór predyktorów*, a następnie na karcie *Wybór reprezentantów* klikamy *Zmienne*, aby wybrać zmienne do analizy i wybieramy wszystkie zmienne ilościowe.

Po zatwierdzeniu ustawień analizy wykonana zostanie analiza czynnikowa z rotacją czynników (*Varimax znormalizowana*). Analiza spowoduje wyodrębnienie niezależnych czynników (wymiarów) zmienności oraz przypisze do tych czynników te zmienne, które będą najmocniej z nimi korelowały. Dzięki temu analizowane zmienne pogrupowane zostaną w wiązki podobnych (w sensie korelacji) zmiennych, które zostaną przypisane do odpowiedniego czynnika. Korelację pomiędzy wyodrębnionym czynnikiem a pierwotną zmienną nazywamy ładunkiem, wartość ładunku pozostawiamy na poziomie 0,7. Jeśli dana

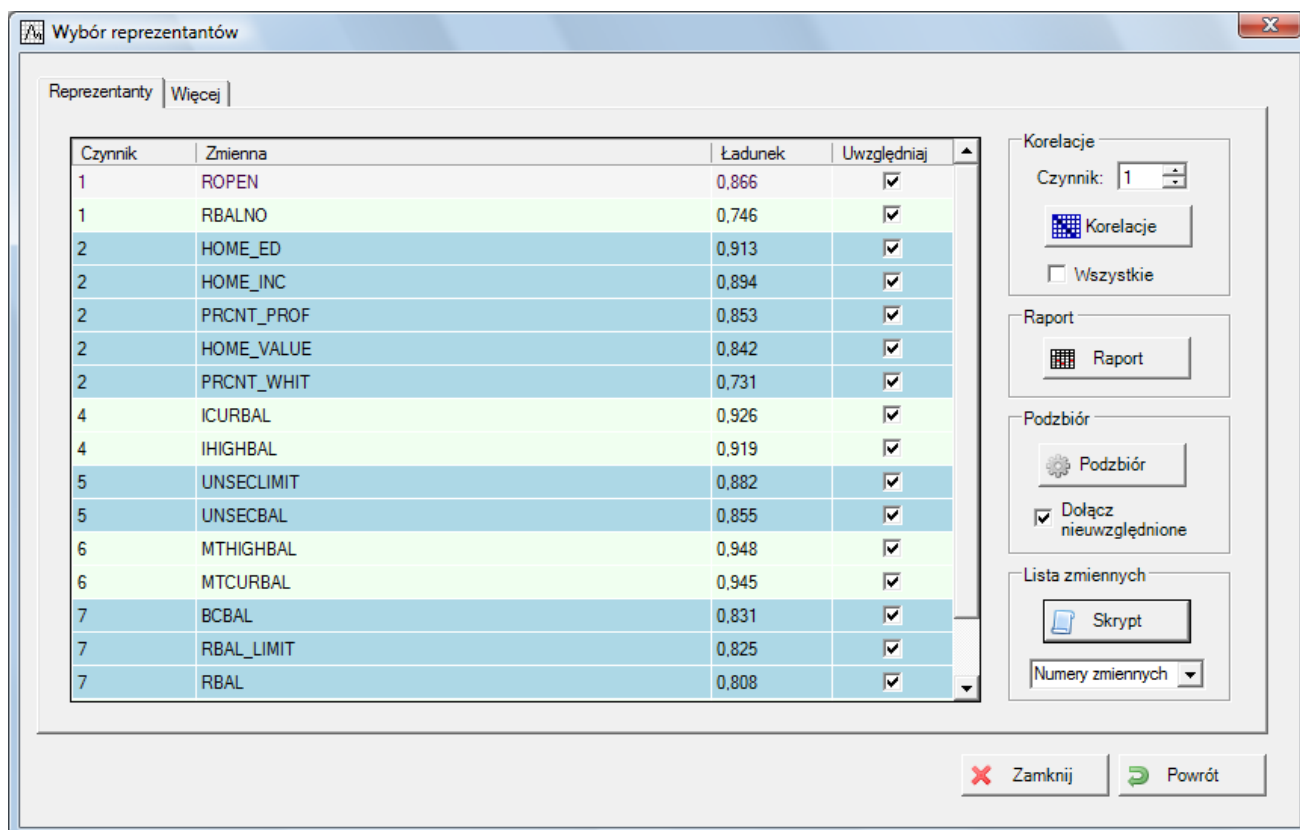
<sup>26</sup> Więcej informacji na temat *Zestawu Skoringowego* zamieszczono w końcowej części artykułu.



zmienna koreluje z wyodrębnionym czynnikiem mocniej niż określona wartość, traktowana będzie jako reprezentanta danego czynnika.



W poniższym oknie widzimy listę wyodrębnionych czynników oraz zmienne, jakie weszły do grupy reprezentantów danego czynnika (Ładunek powyżej 0,7).



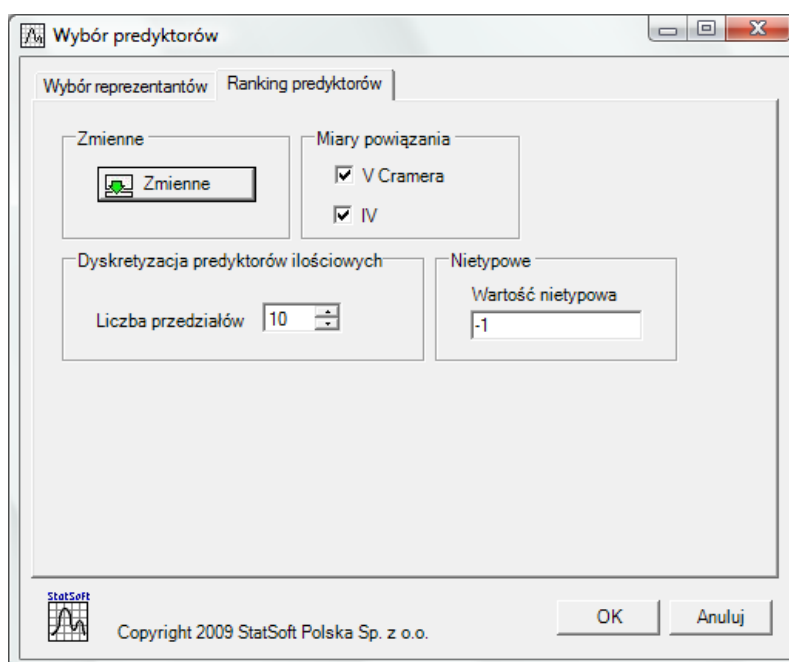


Następnie na podstawie korelacji pomiędzy poszczególnymi zmiennymi wchodzącymi w skład reprezentantów możemy usunąć niektóre zmienne bez ryzyka utraty informacji o badanym zjawisku. Przykładowo zobaczymy macierze korelacji zmiennych wchodzących w skład czynnika 4 i 6.

Zmienna	Czynnik 4		Zmienna	Czynnik 6	
	ICURBAL	IHIGHBAL		MTHIGHBAL	MTCURBAL
ICURBAL	1,00	0,88	MTHIGHBAL	1,00	0,99
IHIGHBAL	0,88	1,00	MTCURBAL	0,99	1,00

W obydwu przypadkach widzimy bardzo wysoką korelację pomiędzy zmiennymi, pozwalającą na bezpieczną eliminację po jednej zmiennej z obydwu par. Aby usunąć zmienne, odznaczamy pole *Uwzględnij* w wierszach odpowiadających tym zmiennym, a następnie klikamy *Podzbiór*, by wygenerować zbiór danych bez usuniętych zmiennych.<sup>27</sup> Procedura ta jest bardzo przydatna, zwłaszcza w sytuacji, gdy nasz zbiór danych zawiera bardzo dużą liczbę wskaźników na przykład finansowych, które są ze sobą mocno skorelowane, a ich liczba uniemożliwia efektywną analizę globalnej macierzy korelacji.

Kolejnym krokiem naszej analizy będzie eliminacja zmiennych, które nieistotnie wpływają na skłonność do zakupu karty. Do oceny siły wpływu poszczególnych predyktorów również użyjemy procedur zaimplementowanych w module *Wybór predyktorów* wchodzącym w skład *Zestawu Skoringowego*. Aby ocenić predyktory, przechodzimy na kartę *Ranking predyktorów*, a następnie wybieramy zmienne do analizy.



<sup>27</sup> Klikając przycisk *Skrypt*, możemy wygenerować makro selekcji zmiennych, którego uruchomienie wykona analogiczną czynność - *STATISTICA* zawiera zaimplementowany język makr oparty na Visual Basic – zgodny z językiem makr pakietu Office



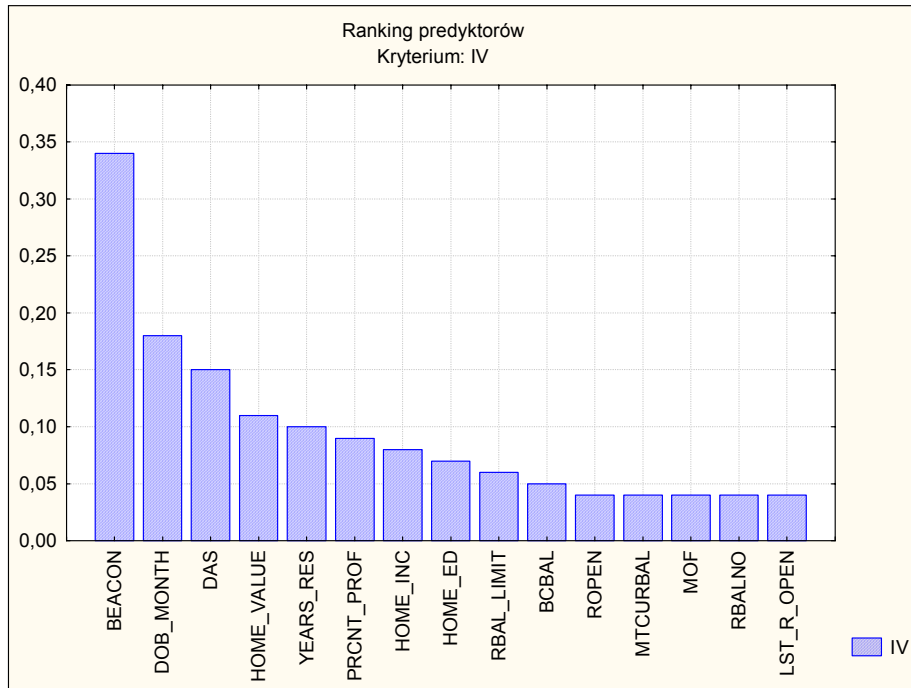
Zmienną zależną będzie zmienna *BUYER*, a pozostałe zmienne zmiennymi niezależnymi (wybieramy je na dwóch listach w zależności od skali pomiaru). Ranking predyktorów wykonany zostanie na podstawie miar IV (*Information Value*) oraz V Cramera.

Ponieważ braki danych pewnej grupy zmiennych zastąpiliśmy wartością  $-1$ , wskażemy ją teraz jako wartość nietypową, tak by uwzględnić również możliwość wpływu braku danych na skłonność do zakupu karty. Po zatwierdzeniu analizy otrzymujemy gotowy ranking predyktorów.

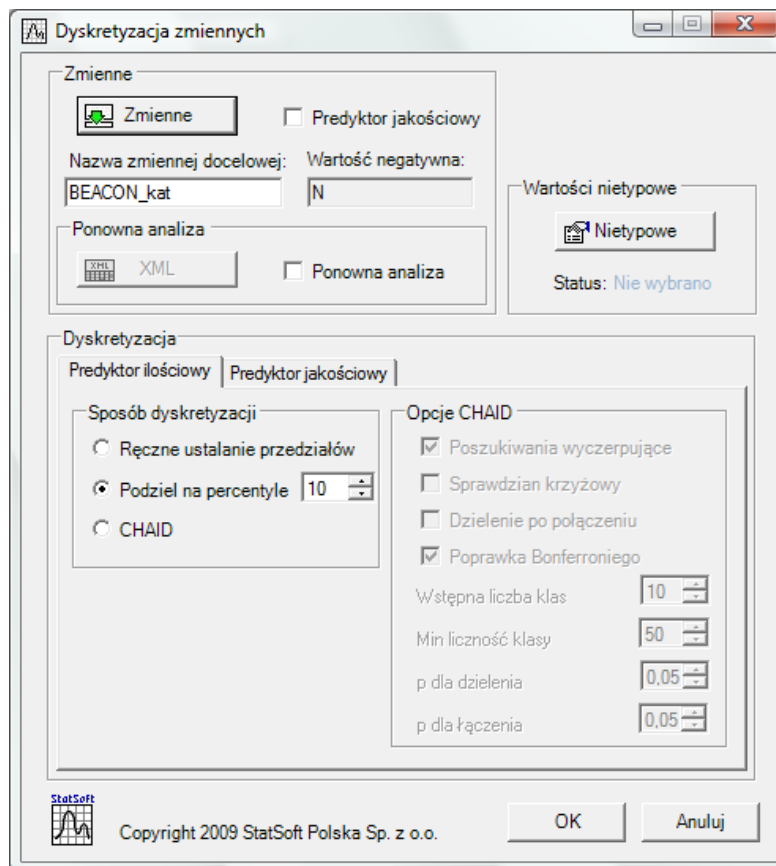
Nr	Nazwa	IV	V Cramera	Uwzględniaj
1	BEACON	0,34	0,25	<input checked="" type="checkbox"/>
2	DOB_MONTH	0,18	0,57	<input checked="" type="checkbox"/>
3	DAS	0,15	0,16	<input checked="" type="checkbox"/>
4	HOME_VALUE	0,11	0,13	<input checked="" type="checkbox"/>
5	YEARS_RES	0,10	0,13	<input checked="" type="checkbox"/>
6	PRCNT_PROF	0,09	0,12	<input checked="" type="checkbox"/>
7	HOME_INC	0,08	0,12	<input checked="" type="checkbox"/>
8	HOME_ED	0,07	0,10	<input checked="" type="checkbox"/>
9	RBAL_LIMIT	0,06	0,10	<input checked="" type="checkbox"/>
10	BCBAL	0,05	0,10	<input checked="" type="checkbox"/>
11	MOF	0,04	0,08	<input checked="" type="checkbox"/>
12	ROPEN	0,04	0,08	<input checked="" type="checkbox"/>
13	MTCURBAL	0,04	0,08	<input checked="" type="checkbox"/>
14	RBALNO	0,04	0,08	<input checked="" type="checkbox"/>
15	LST_R_OPEN	0,04	0,08	<input checked="" type="checkbox"/>
16	RBAL	0,03	0,08	<input checked="" type="checkbox"/>
17	DOB_YEAR	0,03	0,07	<input checked="" type="checkbox"/>
18	EST_INC	0,03	0,07	<input checked="" type="checkbox"/>
19	TBALNO	0,03	0,07	<input checked="" type="checkbox"/>
20	PRCNT_WHIT	0,02	0,05	<input checked="" type="checkbox"/>
21	BCLIMIT	0,02	0,05	<input checked="" type="checkbox"/>

Widzimy, że przy zastosowaniu kryterium IV zmienną, która najmocniej wpływa na skłonność do zakupu karty, jest zmienna *BEACON*, inne istotne zmienne to *DOB\_MONTH* oraz *DAS*. Kolejne zmienne wpływają na skłonność do zakupu karty w coraz mniejszym stopniu. Przyjmijmy kryterium odrzucenia zmiennych z dalszej analizy (tym samym uznania ich za nieistotne), gdy wskaźnik IV jest mniejszy od 0,4. Kryterium to określamy w obszarze *Nie uwzględniaj*, a następnie klikamy *Usuń*, co spowoduje odznaczenie opcji *Uwzględniaj* na liście predyktorów dla tych cech, które nie spełniają podanego warunku.

Usunięcie nieistotnych zmiennych zawęziło liczbę potencjalnych predyktorów do 15. Na ich podstawie w kolejnych etapach analizy będziemy budowali końcowy model. Aby ograniczyć zbiór danych tylko do istotnych predyktorów, klikamy przycisk *Podzbiór* podobnie jak w przypadku wyboru reprezentantów.



Ostatni krok wstępnej analizy danych to dyskretyzacja zmiennych. Naszym celem będzie wyróżnienie w każdej ze zmiennych pewnych grup jednorodnych ze względu na szansę zakupu karty kredytowej i na tej podstawie przygotowanie zmiennych pochodnych, które będą wykorzystane do finalnej analizy.





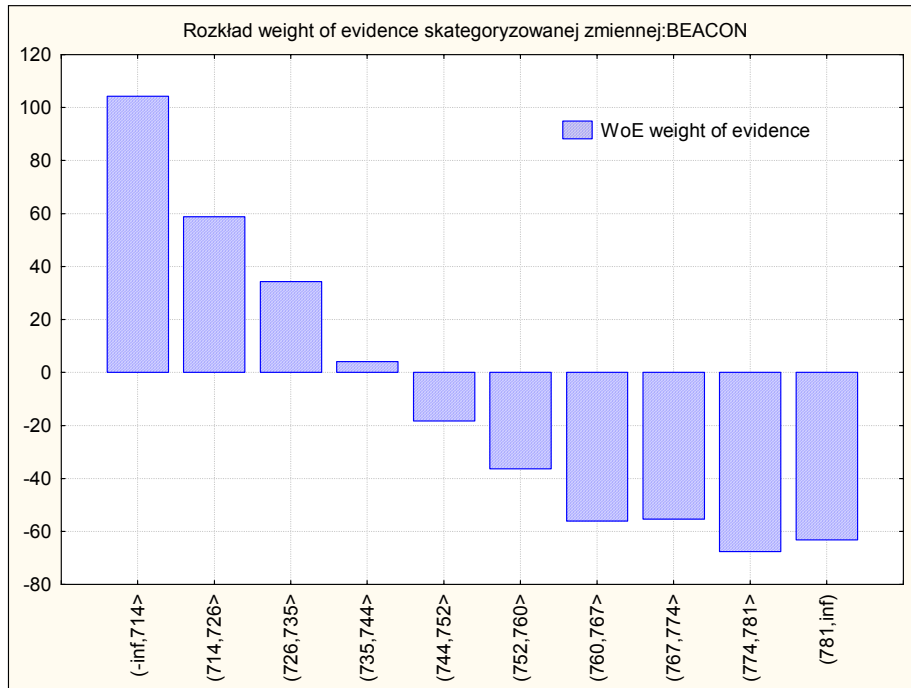
Analiza ta pozwoli nam lepiej zrozumieć charakter analizowanych zmiennych, wygładzić szumy, jakie występują w danych, a także wyeliminować negatywny wpływ obserwacji odstających. Co ważne, w sposób naturalny obsłużone zostaną braki danych.

Aby przygotować profile zmiennych, skorzystamy z modułu *Dyskretyzacja zmiennych* zawartego w *Zestawie Skoringowym*. W oknie *Dyskretyzacja zmiennych* wskazujemy zmienną *BUYER* jako zmienną stanu, natomiast dyskretyzację rozpoczniemy od zmiennej *BEACON*. Przed analizą określamy jeszcze klasę *N* zmiennej *BUYER* jako klasę negatywną (nie kupili karty kredytowej), a następnie dzielimy wartości zmiennej *BEACON* na percentyle.

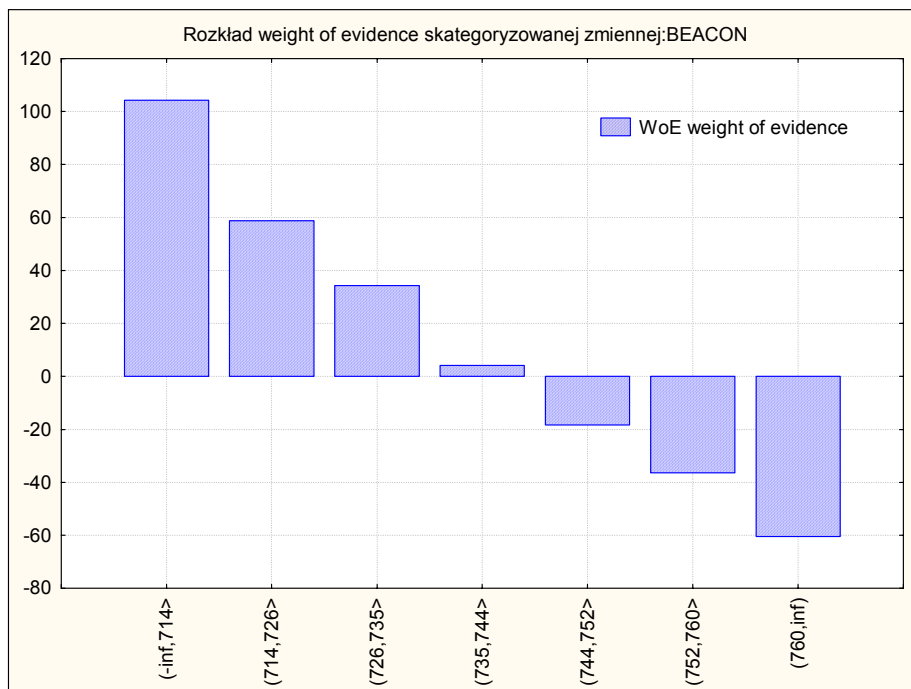
Od	Formuła	Do	Kategoria	Liczność	Scal
	< x <=	714	(-inf,714>	1406	<input type="checkbox"/>
714	< x <=	726	(714,726>	1526	<input type="checkbox"/>
726	< x <=	735	(726,735>	1316	<input type="checkbox"/>
735	< x <=	744	(735,744>	1492	<input type="checkbox"/>
744	< x <=	752	(744,752>	1331	<input type="checkbox"/>
752	< x <=	760	(752,760>	1463	<input type="checkbox"/>
760	< x <=	767	(760,767>	1359	<input type="checkbox"/>
767	< x <=	774	(767,774>	1389	<input type="checkbox"/>
774	< x <=	781	(774,781>	1339	<input type="checkbox"/>
781	< x <=		(781,inf)	1375	<input type="checkbox"/>

W oknie *Przekoduj ilościowe* klikamy przycisk *Przekoduj*, a następnie *Raport*, by wyświetlić raport dyskretyzacji.

Dla każdej kategorii zmiennej *BEACON* obliczono miarę siły wpływu na skłonność do zakupu karty kredytowej *Weight of Evidence* (w polskiej nomenklaturze spotyka się niekiedy termin waga dowodu). Wyższe wartości *WoE* informują o wyższej skłonności do zakupu karty kredytowej.

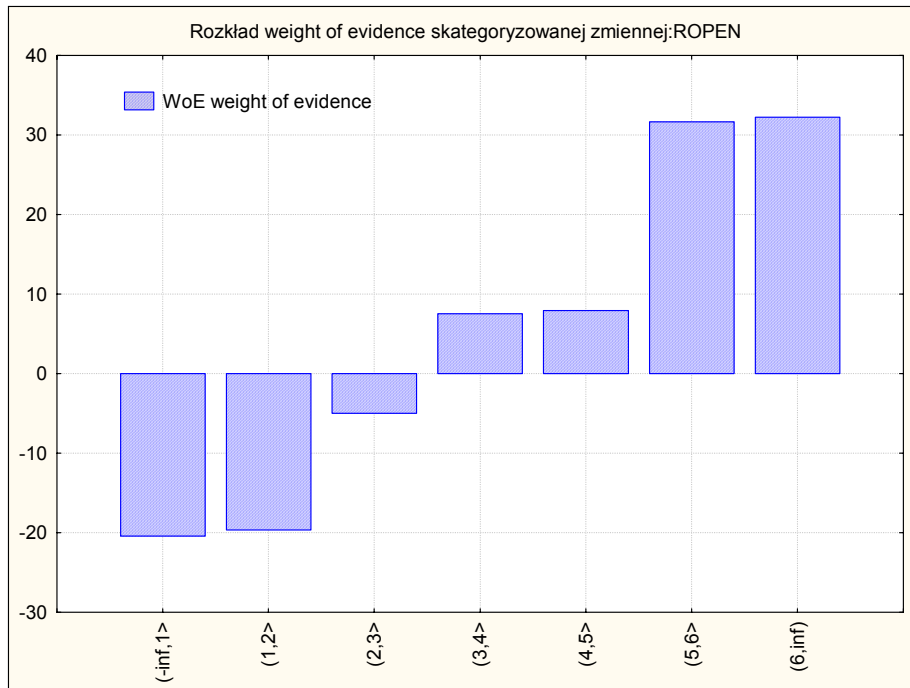


Przykładowo na podstawie wykresu widzimy, iż najwyższą skłonność do zakupu karty wykazują osoby, dla których zmienna *BEACON* jest mniejsza od 714. Skłonność ta stopniowo zmniejsza się wraz ze wzrostem wartości zmiennej *BEACON*. Ponieważ cztery ostatnie kategorie mają w zasadzie taką samą wartość *WoE*, scalimy je do wspólnej kategorii. W oknie *Przekoduj ilościowe* w odpowiednich kategoriach zmiennej zaznaczamy pola wyboru, a następnie klikamy przycisk *Scal*. Po scaleniu profil zmiennej *BEACON* wygląda następująco:

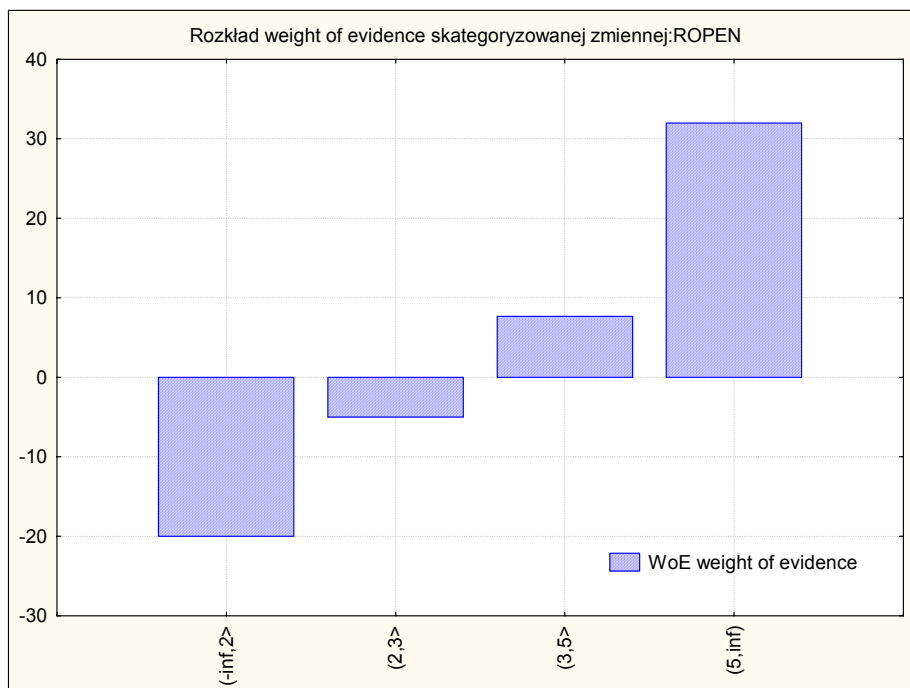




Przygotowany profil dyskretyzacji zapamiętujemy w pliku XML, który tworzymy za pomocą przycisku *Skrypt*. Podobne przekształcenia wykonujemy dla kolejnych zmiennych. Poniżej zamieszczono kilka przykładowych dyskretyzacji.



W przypadku zmiennej *ROPEN* widzimy, że niektóre wartości generują dokładnie taką samą skłonność do zakupu. Bez straty informacji możemy scalić klasy z taką samą wartością *WoE*, otrzymując poniższy profil:



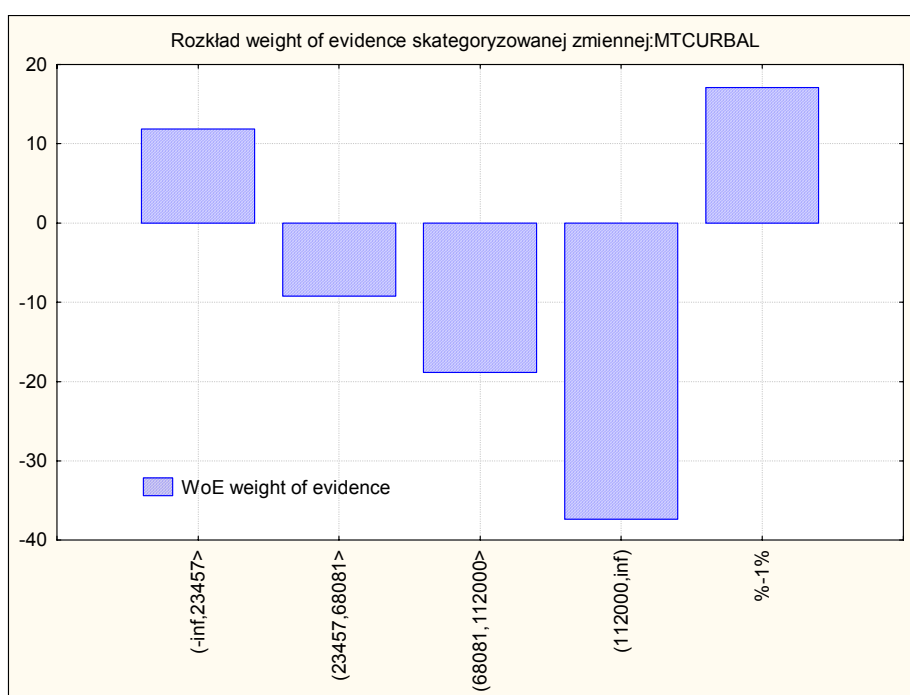
Po scaleniu każda z kategorii charakteryzuje się inną wartością *WoE*.



Dyskretyzacja zmiennych, choć może osłabić moc predykcyjną poszczególnych zmiennych, niesie ze sobą zdecydowanie więcej korzyści:

- ◆ modele zbudowane na podstawie tak przygotowanych zmiennych są bardziej stabilne,
- ◆ podczas estymacji parametrów wykazują mniejszą skłonność do przeuczenia,
- ◆ dyskretyzacja w naturalny sposób rozwiązuje problem danych odstających (skrajne wartości trafiają po prostu do odpowiednich przedziałów) oraz braków danych (braki danych stanowią osobną kategorię, co pozwala uwzględnić ich możliwy wpływ na badane zjawisko).

Dla przykładu poniżej widzimy profil dyskretyzacji zmiennej *MCTURBAL*, w którym brak danych (kategoria %-1%) wiąże się z największą dla tej zmiennej skłonnością do zakupu karty.



Dyskretyzacja zmiennych pozwala również wychwycić wiele błędów i sprzeczności występujących w danych oraz zidentyfikować zmienne anachroniczne, czyli zmienne, których wartości zostały określone już po fakcie zakupu karty. W naszym przykładzie taką zmienną okazała się być zmienna *DOB\_MONTH*

Kategoryzowana zmienna: DOB_MONTH					
DOB_MONTH	Kupił	Nie kupił	Suma	IV	WoE weight of evidence
Brak	1858	10999	12857	0,18	-47,8
Podano	1139	0	1139		
Ogół grp	2997	10999	13996	0,18	

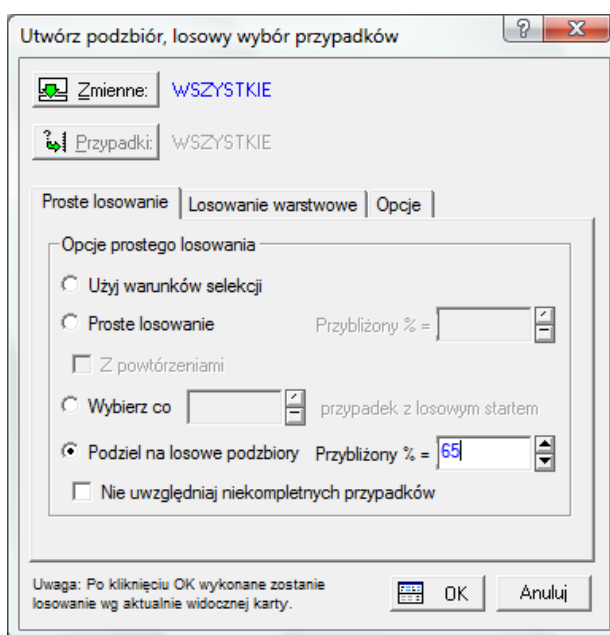
Zmienna ta wydaje się być bardzo neutralną zmienną, ponieważ określa miesiąc urodzenia posiadacza karty. Problemem jest jednak fakt, że miesiąc ten został uzupełniony po zakupie karty i wpis o nim mają jedynie posiadacze karty. Gdybyśmy chcieli uwzględnić tę



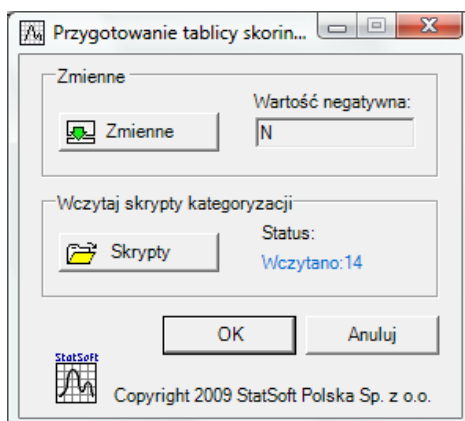
zmienną w naszym modelu, dla wszystkich osób, wobec których model byłby stosowany, wartość *DOB\_MONTH* byłaby pusta, a model byłby bezużyteczny.

### ***Szacowanie parametrów modelu logitowego***

Po przygotowaniu zmiennych do analizy przechodzimy do fazy modelowania. Metodą, jakiej użyjemy w pierwszej kolejności, będzie regresja logistyczna. Dodatkowo dla celów porównawczych zbudujemy model za pomocą drzew wzmacnianych. Aby być zgodnym z zasadami budowy modeli predykcyjnych, podzielimy nasz zbiór danych na dwa podzbiory: uczący (*Uczacy.sta*), na którym oszacujemy parametry modelu, oraz testowy (*Testowy.sta*), na podstawie którego ocenimy dobroć dopasowania do zadanego problemu. Najwygodniej będzie nam to zrobić za pomocą opcji *Podzbiór*, znajdującej się w menu *Dane*.



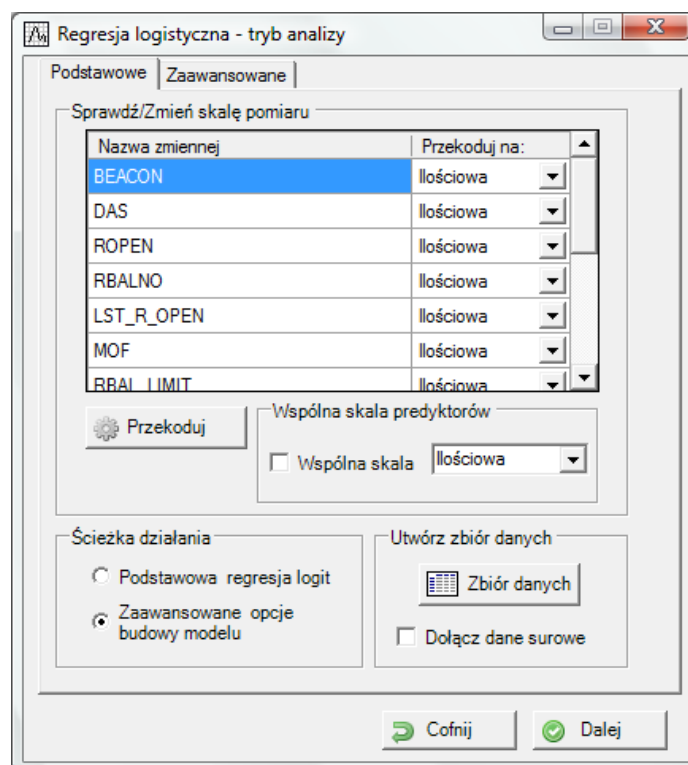
W oknie *Utwórz podzbiór, losowy wybór przypadków* zaznaczamy opcję *Podziel na losowe podzbiory* i określamy, by zbiór uczący zawierał 65% przypadków. Po zatwierdzeniu analizy nasz zbiór zostanie podzielony na dwa losowo określone podzbiory. Mniejszy z nich (około 5000 przypadków) odłożymy do celów testowych, natomiast większy (około 9000 przypadków) posłużymy nam do oszacowania parametrów modelu.





By zbudować model logistyczny, z menu *Zestaw Skoringowy* wybieramy opcję *Budowa tablicy skoringowej*, a następnie wybieramy zmienne do analizy. Ponieważ będziemy chcieli zbudować model na podstawie dyskretyzowanych zmiennych, za pomocą przycisku *Skrypty* wczytujemy definicje dyskretyzacji zapisane w plikach XML.

Po zatwierdzeniu wyboru zmiennych oraz profili dyskretyzacji przechodzimy do szczegółowych ustawień analizy, klikając *OK*.



W oknie *Regresja logistyczna – tryb analizy* klikamy *Przekoduj*, aby przygotować dyskretyzację (poszczególne wartości zostaną zamienione na odpowiadające im wartości *WoE*).<sup>28</sup> Po przekodowaniu zmiennych przechodzimy na kartę *Zaawansowane* i wybieramy opcję *Krokowa wsteczna* jako sposób budowy modelu, co pozwoli nam wykonać finalną eliminację zmiennych (z modelu odrzucone będą te zmienne, których oceny parametrów będą nieistotnie różnić się od 0).

By oszacować parametry regresji logistycznej, klikamy przycisk *dalej*, po czym w oknie *Wyniki regresji i parametry skali* możemy przejrzeć uzyskane wyniki. W tabeli poniżej możemy zaobserwować wartości ocen parametrów regresji uzyskane w wyniku analizy.

Raport *Budowanie modelu* umożliwi prześledzenie procesu doboru parametrów. Proces zakończył się już w drugiej iteracji, po odrzuceniu z modelu zmiennej *RBALNO*.

<sup>28</sup> Klikając przycisk *Zbiór danych*, możemy wygenerować przekodowany zbiór danych, którego możemy użyć do budowy modeli skoringowych za pomocą innych metod (np. drzew klasyfikacyjnych, drzew wzmacnianych czy sieci neuronowych).



	Ocena	Standard Błąd	Walda Stat.	p
Wyraz wolny	-1,31570	0,027732	2250,877	0,000000
HOME_ED_kat	0,00537	0,001200	20,026	0,000008
HOME_INC_kat	0,00295	0,001208	5,945	0,014763
HOME_VALUE_kat	0,00404	0,001033	15,322	0,000091
YEARS_RES_kat	0,01020	0,001366	55,815	0,000000
MTCURBAL_kat	0,00978	0,001529	40,895	0,000000
BCBAL_kat	0,00299	0,001487	4,043	0,044342
RBAL_LIMIT_kat	-0,00686	0,001492	21,159	0,000004
MOF_kat	0,01017	0,001352	56,601	0,000000
LST_R_OPEN_kat	0,00360	0,001647	4,783	0,028745
PRCNT_PROF_kat	0,00321	0,001210	7,030	0,008016
ROPEN_kat	0,00813	0,001669	23,711	0,000001
DAS_kat	0,00316	0,000854	13,672	0,000218
BEACON_kat	0,00968	0,000627	238,659	0,000000
Skala	1,00000	0,000000		

Możemy tak zbudowany model zapisać teraz do pliku PMML, by móc go stosować dla nowych danych za pomocą opcji *Data Mining - Szybkie wdrażanie modeli predykcyjnych PMML*. My jednak przekształcimy parametry modelu logistycznego do postaci karty skoringowej.

Wyniki regresji i parametry skali

Parametry skali | Parametry modelu | Podsumowanie regresji

Parametry skali

Punkty podwajające szansę (pdo): 20

Szansa 50 do 1 dla 600 punktów

Mnoznik: 28,8539008177 Przesunięcie: 487,122876204

Przelicz

Korekta skali dla próby zbalansowanej

Próba zbalansowana

Prawdop. losowania warstw

stan pozytywny: 0,05

stan negatywny: 1,00

Cofnij Dalej

W tym celu na karcie *Parametry skali* klikamy przycisk *Przelicz*, a następnie przycisk *Dalej*.

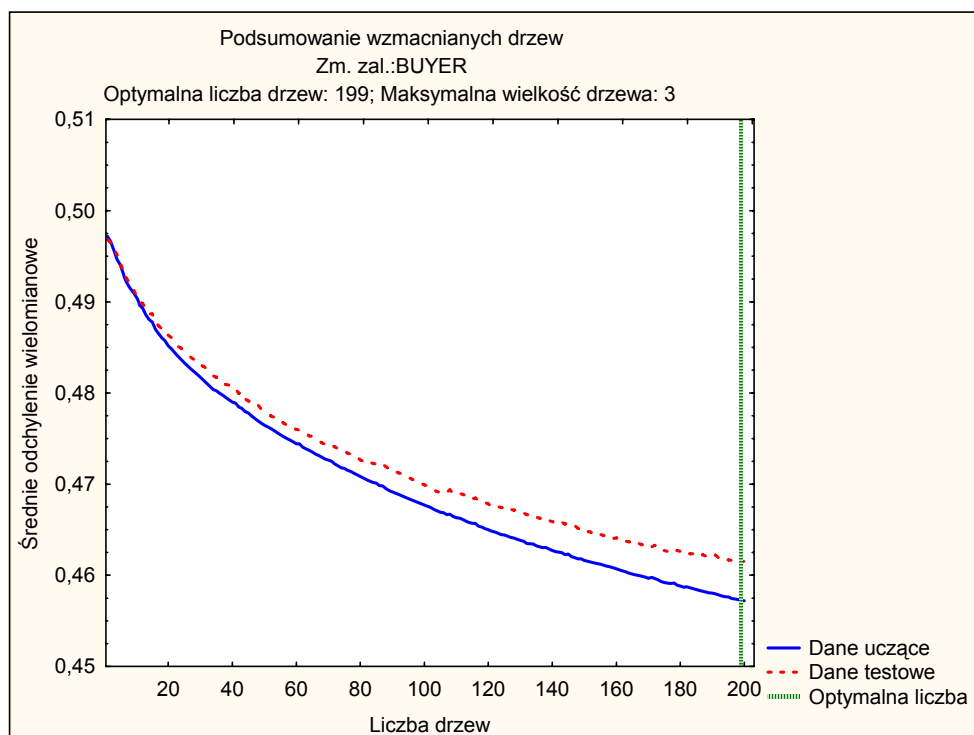
W wyniku przekształcenia ocen parametrów regresji logistycznej otrzymujemy tablicę skoringową, w której poszczególnym kategoriom zmiennych modelu przypisano określoną liczbę punktów.



Zmienna	Zakres	WoE	Ocena	s. Walda	p	Skoring	Skoring zaokr.
PRCNT_P...	(48;inf)	18,779	0,00321	7,02999	0,00802	36,290	36
PRCNT_P...	Wartość n...	-	-			34,315	34
ROPEN	(-inf;2>	-19,983	0,00813	23,71119	0,00000	29,863	30
ROPEN	(2;3>	-5,006	0,00813	23,71119	0,00000	33,376	33
ROPEN	(3;5>	7,696	0,00813	23,71119	0,00000	36,356	36
ROPEN	(5;inf)	31,996	0,00813	23,71119	0,00000	42,057	42
ROPEN	Wartość n...	-	-			34,345	34
DAS	(-inf;206>	-62,243	0,00316	13,67173	0,00022	28,876	29
DAS	(206;250>	-53,714	0,00316	13,67173	0,00022	29,653	30
DAS	(250;286>	-43,159	0,00316	13,67173	0,00022	30,616	31
DAS	(286;420>	-3,436	0,00316	13,67173	0,00022	34,237	34
DAS	(420;462>	24,523	0,00316	13,67173	0,00022	36,787	37
DAS	(462;517>	49,004	0,00316	13,67173	0,00022	39,019	39
DAS	(517;inf)	58,423	0,00316	13,67173	0,00022	39,878	40
DAS	Wartość n...	-	-			34,154	34
BEACON	(-inf;714>	104,273	0,00968	238,65945	0,00000	63,675	64
BEACON	(714;726>	58,732	0,00968	238,65945	0,00000	50,955	51
BEACON	(726;735>	34,258	0,00968	238,65945	0,00000	44,119	44
BEACON	(735;744>	4,139	0,00968	238,65945	0,00000	35,707	36
BEACON	(744;752>	-18,381	0,00968	238,65945	0,00000	29,417	29
BEACON	(752;760>	-36,354	0,00968	238,65945	0,00000	24,397	24
BEACON	(760;inf)	-60,410	0,00968	238,65945	0,00000	17,678	18
BEACON	Wartość n...	-	-			32,168	32

## Budowa modelu drzew wzmocnianych

Jako drugiej, konkurencyjnej metody budowy modelu skoringowego użyjemy modułu drzew wzmocnianych. Aby uruchomić moduł z menu *Data mining*, wybieramy opcję *Wzmocniane drzewa klasyfikacyjne i regresyjne*, następnie wybieramy typ analizy jako *Zadanie klasyfikacyjne*.

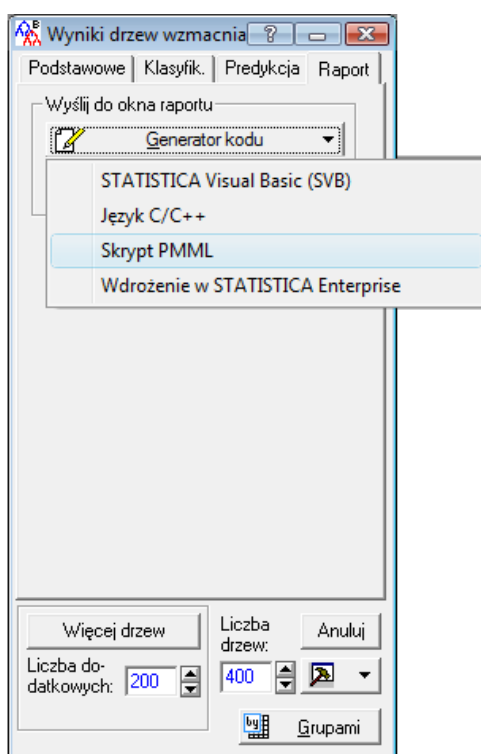




W kolejnym kroku w oknie *Ustawienia drzew wzmocnianych* wskazujemy zmienne do analizy – zmienna *BUYER* będzie podobnie jak w poprzednim przypadku zmienną zależną, pozostałe zmienne określamy jako predyktory ilościowe. Pozostałe parametry metody pozostawiamy na poziomie domyślnym i zatwierdzamy wykonanie analizy.

Zbudowany model składa się z zespołu 199 prostych drzew klasyfikacyjnych. Analizując wykres przebiegu uczenia, widzimy, że chociaż błąd na danych testowych zaczął się stabilizować, kształt krzywej sugeruje, że zwiększenie liczby drzew może spowodować poprawę zdolności predykcyjnej modelu. Klikamy więc opcję *Więcej drzew*, w wyniku czego zbudowany model został powiększony do 380 drzew.

Po wykonaniu modelu w oknie *Wyniki drzew wzmocnianych* przechodzimy na kartę *Raport* i klikamy przycisk *Generator kodu* i zapisujemy zbudowany model w postaci pliku *PMML*, który będziemy mogli stosować dla nowych danych. Po wygenerowaniu modelu zamykamy moduł drzew wzmocnianych.



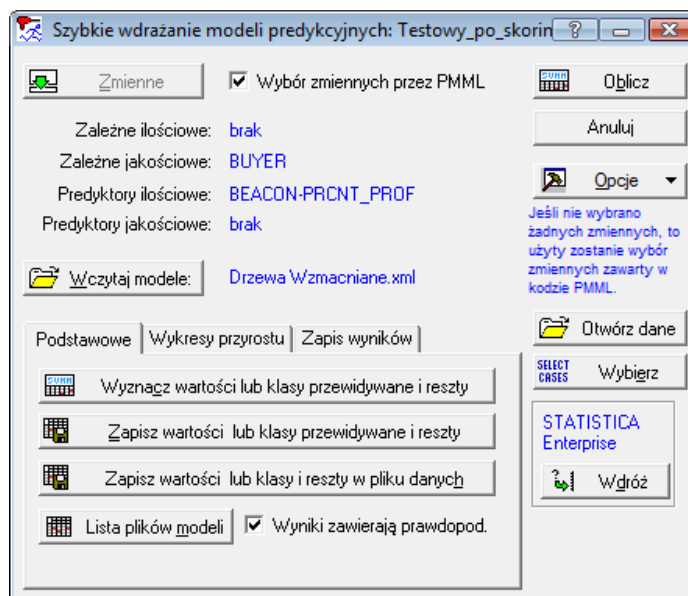
## ***Ocena i porównanie modeli***

W celu oceny zbudowanych modeli otwieramy plik *Testowy.sta*, dla którego zastosujemy zbudowany model, generując odpowiedź modelu w postaci prawdopodobieństwa przynależności do grupy osób, które kupiły kartę.

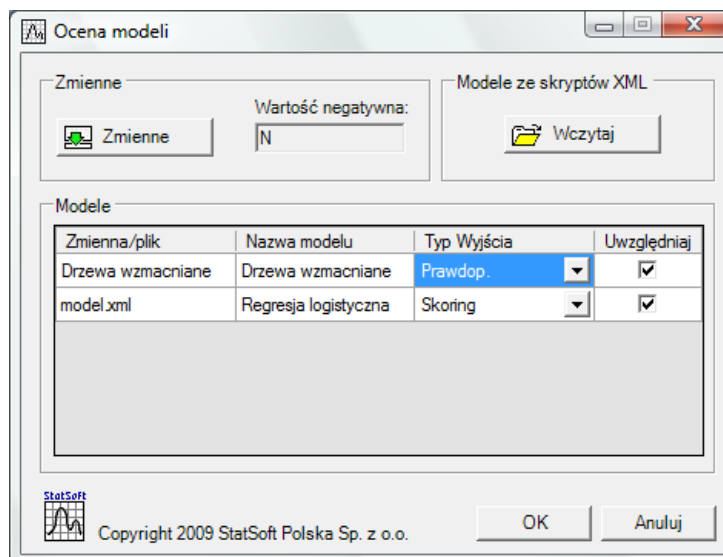
Aby zastosować model drzew wzmocnianych dla zbioru testowego, z menu *Data Mining* wybieramy opcję *Szybkie wdrażanie modeli predykcyjnych PMML* i wczytujemy skrypt PMML za pomocą polecenia *Wczytaj modele*. Następnie generujemy przewidywania



modelu za pomocą przycisku *Zapisz wartości lub klasy przewidywane i reszty*. Po jego naciśnięciu otrzymujemy arkusz *STATISTICA* zawierający przewidywania modelu.<sup>29</sup>

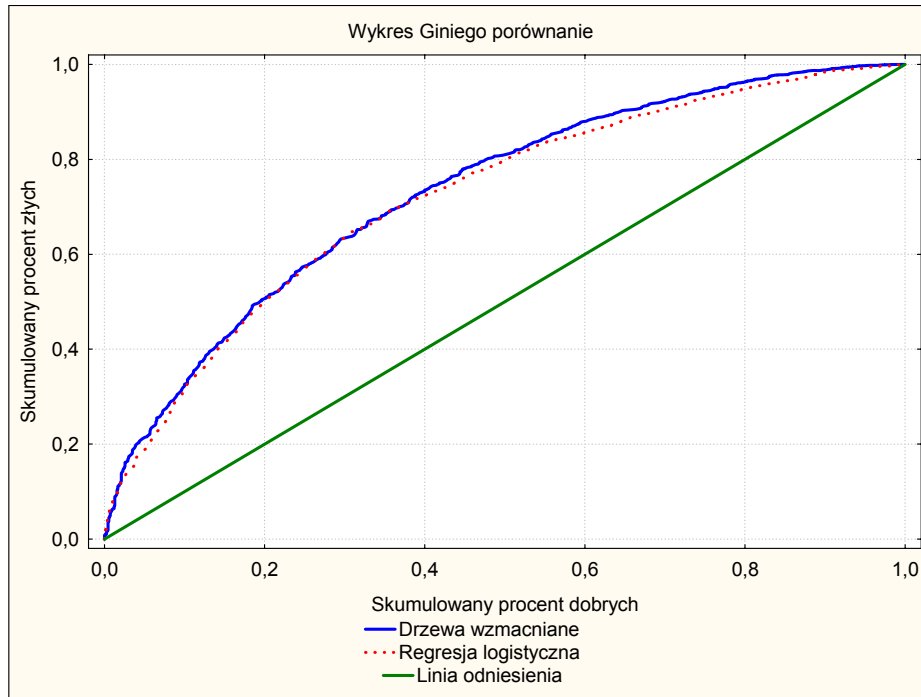


W kolejnym kroku z menu *Zestaw Skoringowy* wybieramy opcję *Ocena modeli*, a następnie wybieramy zmienną *BUYER* jako zmienną zależną oraz zmienną *Drzewa wzmacniane* zawierającą wynik modelu drzew. Za pomocą opcji *Wczytaj* wczytujemy dodatkowo model regresji logistycznej. Następnie na liście *Modele* zmieniamy *Typ Wyjścia* dla modelu drzew wzmacnianych na *Prawdop.*, aby uwzględnić fakt, że wyniki działania modelu drzew zapisane są w postaci prawdopodobieństwa.



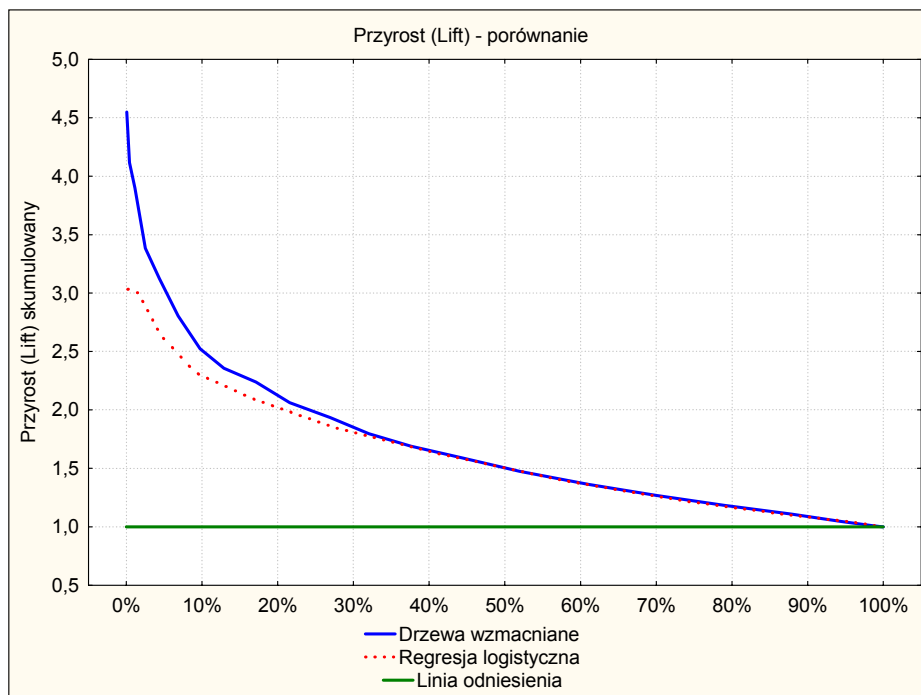
Po zatwierdzeniu analizy w oknie *Ocena modeli* – wyniki klikamy przycisk *Wskaźniki*, aby otrzymać podsumowanie jakości modeli.

<sup>29</sup> Istnieje także możliwość zapisywania przewidywań modelu bezpośrednio do bazy danych za pomocą tabeli zdalnego przetwarzania (IDP).



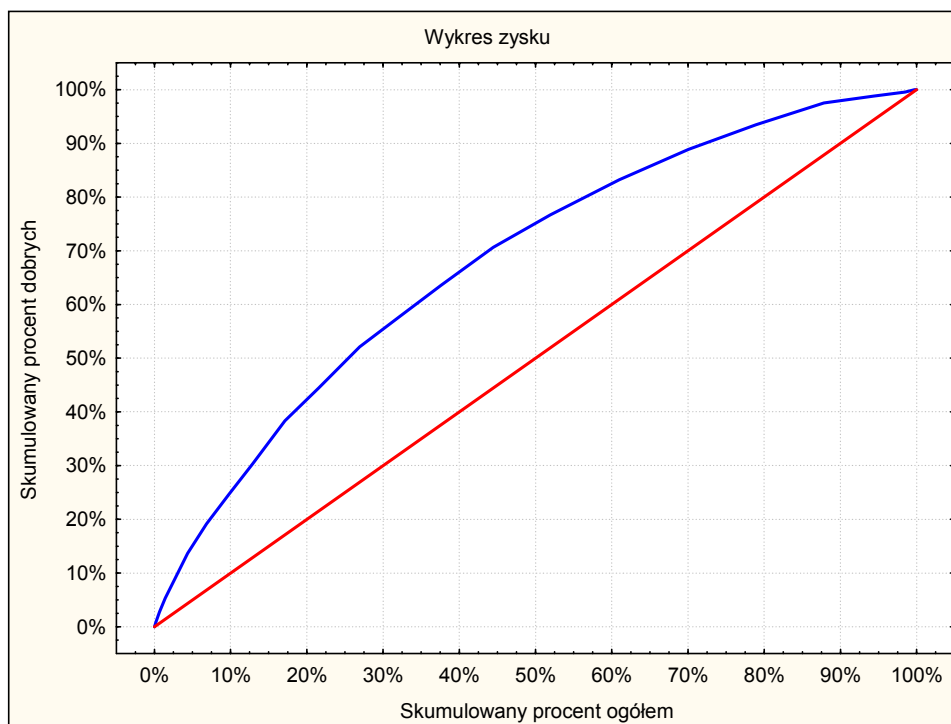
	IV	KS	Gini	Dywergencja	ROC
Drzewa wzmocniane	0,696	0,340	0,464	0,779	0,732
Regresja logistyczna	0,657	0,340	0,442	0,683	0,721

Na podstawie wyliczonych wskaźników jakości stwierdzamy, że oba modele mają porównywalną jakość, jednak nieznacznie lepiej sprawdza się model drzew wzmocnianych. Poza statystyką KS (Kolmogorowa-Smirnowa) ma wyższe wszystkie wskaźniki dobroci dopasowania.





Jego lepszą jakość potwierdza również wykres przyrostu (*Lift*), na którym widzimy większą wartość przyrostu dla pierwszych 30% osób z najwyższym prawdopodobieństwem zakupu. Dodatkowo dla modelu drzew wygenerujemy wykres zysku. Na jego podstawie możemy stwierdzić, że wysyłając ofertę do 45% naszych klientów dotrzemy do około 70% osób, które byłyby skłonne na nią odpowiedzieć.

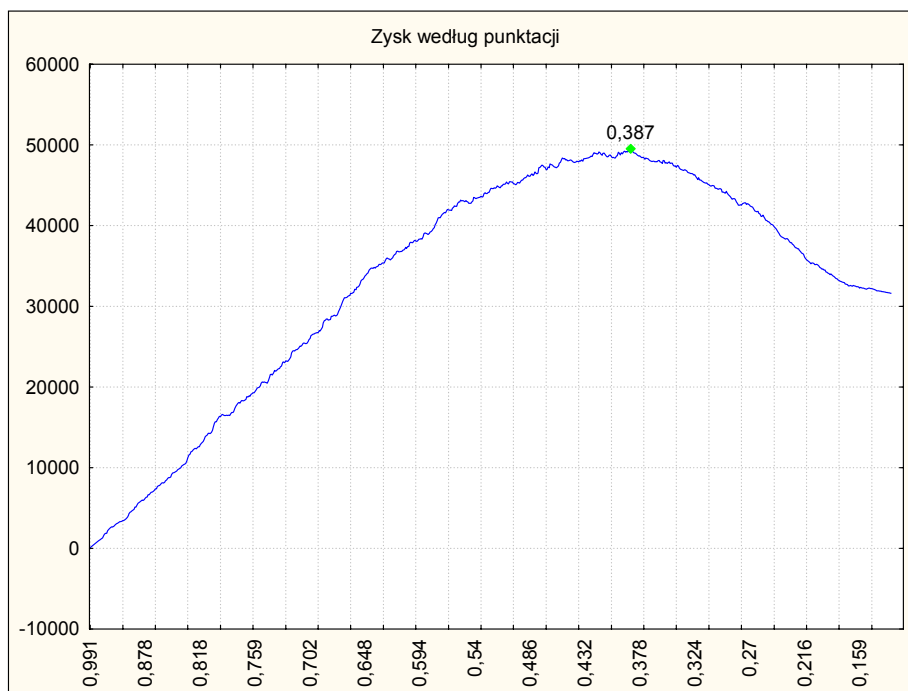
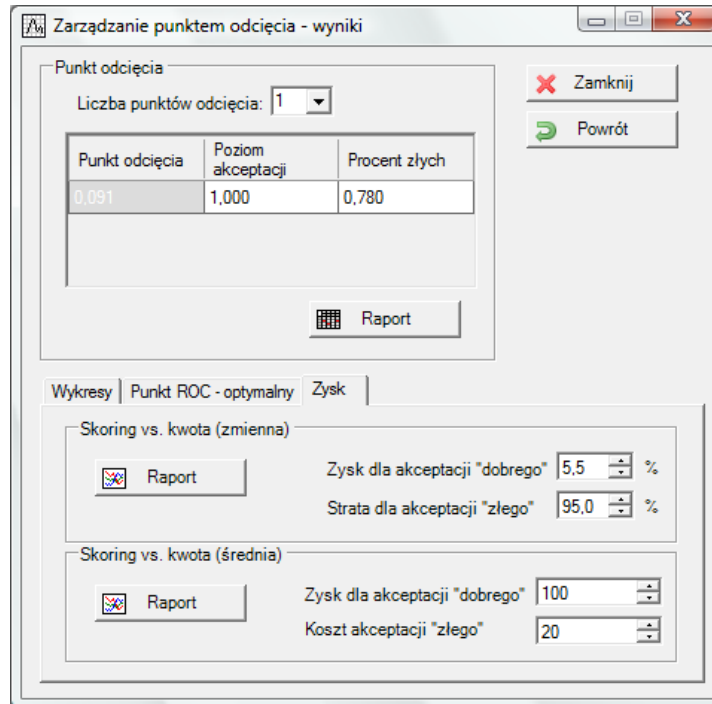


### Wybór punktu odcięcia

Ostatnim elementem związanym z oceną modelu jest wybór optymalnego punktu odcięcia, czyli wskazanie wartości progowej skoringu określonego przez model, powyżej której będziemy podejmować działanie. Osoby ze skoringiem poniżej tego punktu będą wyłączone z planowanej kampanii. Aby określić optymalny punkt odcięcia, użyjemy modułu *Zarządzanie punktem odcięcia* i wyznaczymy optymalny punkt odcięcia dla modelu drzew wzmacnianych.

W oknie *Zarządzanie punktem odcięcia – wyniki* przechodzimy na kartę *Zysk*, a następnie w obszarze *Skoring vs kwota (średnia)* określamy koszt dotarcia do klienta, który odrzuci naszą ofertę (20), oraz zysk, jaki spodziewamy się uzyskać dla osób, które ją zaakceptują (100). Po określeniu powyższych parametrów klikamy *Raport*, by wyświetlić podsumowanie analizy.

Poniższy wykres przedstawia spodziewany zysk z kampanii w zależności od przyjętego punktu odcięcia. Analizując przebieg krzywej na wykresie, możemy stwierdzić, że najwyższy zysk z kampanii sprzedażowej osiągniemy, jeśli zastosujemy punkt odcięcia modelu na poziomie 0,387. Osoby ze skoringiem poniżej tej wartości nie powinny uczestniczyć w kampanii.



## Zestaw Skoringowy

Zestaw Skoringowy STATISTICA jest dedykowanym zestawem narzędzi wspierających proces przygotowania i oceny modeli skoringowych, będącym dodatkiem do systemu STATISTICA Data Miner. Został zaprojektowany w oparciu o sprawdzone standardy przygotowania i oceny modeli skoringowych. Dzięki prostemu interfejsowi i logicznemu ukła-



dowi modułów pozwala szybko i intuicyjnie przejść przez cały proces przygotowania modelu skoringowego. Za jego pomocą użytkownicy mają możliwość budowania modeli na potrzeby skoringu marketingowego, kredytowego, wyłudzeń czy medycznego. *Zestaw Skoringowy* zawiera moduły umożliwiające:

- ◆ wybór zmiennych istotnie wpływających na badane zjawisko,
- ◆ narzędzia do dyskretyzacji zmiennych ilościowych i rekatoryzacji zmiennych jakościowych,
- ◆ budowy i oceny modeli skoringowych,
- ◆ wyboru optymalnego punktu odcięcia.

Więcej informacji na temat *Zestawu Skoringowego STATISTICA* można znaleźć na stronie [WWW.statsoft.pl/industries/skoring.html](http://WWW.statsoft.pl/industries/skoring.html).

## Literatura

1. Demski T. *Model data mining przewidujący odpowiedź klientów na ofertę* [w:] Data mining: poznaj siebie i swoich klientów, Materiały z seminariów StatSoft Polska, 2005, [http://www.statsoft.pl/czytelnia/8\\_2007/Demski05-1.pdf](http://www.statsoft.pl/czytelnia/8_2007/Demski05-1.pdf).
2. Migut G. *Wspomaganie kampanii sprzedaży krzyżowej (cross-selling) na przykładzie oferty banku* [w:] Data mining: poznaj siebie i swoich klientów, Materiały z seminariów StatSoft Polska, 2005, [http://www.statsoft.pl/czytelnia/8\\_2007/Migut05-1.pdf](http://www.statsoft.pl/czytelnia/8_2007/Migut05-1.pdf).
3. Pyle D., *Data Preparation for Data Mining*, Morgan Kaufmann, 1999.