



MODELE SKORINGOWE W BIZNESIE I NAUCE

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

Jednym z szerzej wykorzystywanych typów modeli statystycznych są modele klasyfikacyjne, gdzie modelowana zmienna zależna przyjmuje dwa stany. Modele tego typu określamy mianem modeli skoringowych, ponieważ rezultatem ich działania jest ocena (*scoring*) wyrażająca prawdopodobieństwo lub szansę zajścia modelowanego zdarzenia. Zakres zastosowań modeli skoringowych jest bardzo rozległy, począwszy od szeroko pojętego biznesu, poprzez zagadnienia technologiczne, po różnorakie zagadnienia naukowe.

Modele skoringowe są już obecnie standardowym narzędziem wsparcia procesu oceny wiarygodności kredytowej klientów indywidualnych oraz małych i średnich przedsiębiorstw. Coraz popularniejsze stają się także w obszarze ryzyka operacyjnego, wspierając proces wykrywania nadużyć.

Kolejnym zastosowaniem biznesowym modeli skoringowych jest obszar analitycznego CRM. Modele skoringowe pozwalają na lepszą identyfikację grup klientów z najwyższym potencjałem zakupowym, wskazywanie grup docelowych do kampanii sprzedażowych oraz identyfikację klientów najmocniej zagrożonych odejściem.

W przypadku zastosowań technologicznych modele skoringowe wspierają proces identyfikacji wadliwych produktów oraz identyfikują ryzyko wystąpienia awarii maszyn.

Bardzo ważnym obszarem wykorzystania narzędzi skoringowych jest medycyna. Modele skoringowe stosowane w tym obszarze umożliwiają między innymi klasyfikację pacjentów do grupy chorych lub zdrowych na podstawie wyników badań diagnostycznych bądź też określenie optymalnego sposobu terapii.

Charakterystyka metod skoringowych

Osoby pragnące modelować zależności za pomocą metod skoringowych stoją przed wyborem szeregu metod analitycznych umożliwiających budowę tego typu modeli.

Metody skoringowe możemy podzielić na dwie grupy. Pierwsza grupa to metody klasyczne – statystyczne, do których możemy zaliczyć: regresję logistyczną, regresję probit, analizę dyskryminacyjną czy też model proporcjonalnego hazardu Coxa. Druga grupa to metody typu data mining, spośród których najpopularniejsze to sieci neuronowe, drzewa



wzmacniane (*boosted trees*), losowy las (*random forests*) czy metoda wektorów nośnych (*support vector machines*).

Wyczerpujące porównanie obydwóch grup metod wymagałoby stosownej rozprawy naukowej. W niniejszym artykule skoncentrujemy się na charakterystycznych cechach dwóch metod: regresji logistycznej oraz drzewach wzmacnianych oraz ich konsekwencjach dla możliwości wykorzystania ich w biznesie i nauce.

Regresja logistyczna jest niewątpliwie najbardziej popularną metodą budowy modeli skoringowych. Podobnie jak pozostałe metody zaliczane do grupy metod statystycznych pozwala na przedstawienie modelowanych zależności w postaci wzoru matematycznego. Dodatkowo na podstawie oceny parametru regresji łatwo jest uzyskać iloraz szans¹⁴.

Bardzo ważną zaletą regresji logistycznej jest możliwość przekształcenia uzyskanego wzoru do formatu karty skoringowej, umożliwiając zrozumienie i interpretację zbudowanego modelu nawet osobom niezajmującym się analizą danych.

Drzewa wzmacniane to metoda data mining polegająca na budowie sekwencji modeli, przy czym każdy kolejny model przykłada większą „wagę” do tych obserwacji, które zostały błędnie zaklasyfikowane przez modele poprzednie. W kolejnych modelach błędnie zaklasyfikowane obserwacje otrzymują większe wagi, przez co każdy kolejny model uczy się lepiej rozróżniać obserwacje wcześniej źle zaklasyfikowane. Finalny model często składa się więc z kilkuset składowych modeli, które głosując dają końcowy wynik.

Już sama podstawowa charakterystyka obydwóch metod pozwala na wychwycenie zasadniczych różnic w tych podejściach. Z jednej strony metody statystyczne pozwalają uzyskać modele o czytelnej i prostej w interpretacji postaci, w których możemy w łatwy sposób ocenić kierunek i siłę wpływu poszczególnych cech na model, z drugiej strony modele data mining są modelami relatywnie dużo bardziej skomplikowanymi, przez co praktyczna interpretacja ich parametrów jest niemożliwa. Często mówi się, że działają one na zasadzie czarnej skrzynki.

Budując modele data mining, tracimy więc możliwość interpretacji, jednak wraz z komplikacją postaci modelu rośnie także jego siła predykcyjna. O ile modele statystyczne są w stanie modelować jedynie główne efekty, bez uwzględnienia interakcji (uwzględnienie interakcji w modelach liniowych jest możliwe za pomocą dodatkowych zmiennych, które reprezentują zidentyfikowane interakcje), modele data mining dają możliwość modelowania nieliniowych i często ulotnych wzorców, co zwykle pozwala zbudować lepsze modele, o wyższym wskaźniku dopasowania oraz zwykle szerszym obszarze punktacji, w którym są one w stanie dobrze dyskryminować analizowane grupy.

Bardzo ważnym etapem budowy statystycznych modeli skoringowych jest przygotowanie danych. Należy zidentyfikować zmienne, które są ze sobą nadmiernie skorelowane, i usunąć je ze zbioru danych. Kolejnym krokiem jest eliminacja zmiennych, które nie wpływają w istotnym stopniu na modelowane zjawisko – tego typu zmienne mogłyby działać

¹⁴ Jeżeli b to ocena parametru regresji logistycznej, to iloraz szans = $\exp\{b\}$.



destabilizująco na model, nie wnosząc do niego żadnych istotnych informacji. Metodyka budowy kart skoringowych wymaga, by każda analizowana zmienna podlegała dyskretyzacji, czyli podziałowi wartości analizowanych zmiennych na przedziały jednorodne ze względu na natężenie występowania w nich analizowanego zjawiska. Tego typu operacja umożliwia lepsze zrozumienie charakteru związku cechy z analizowanym zjawiskiem, pozwala także na wychwycenie błędów i pomyłek w danych. Dyskretyzacja w naturalny sposób rozwiązuje problem wartości odstających oraz braków danych, które możemy traktować jako odrębne klasy, pozwalając analitykowi uwzględnić fakt, iż brak danych może być istotną informacją. Pozwala także na modelowanie niemonotonicznych zmian wpływu poszczególnych cech na zjawisko co jest dużą zaletą tego podejścia.

Kolejnym istotnym elementem przygotowania danych jest przygotowanie próby uczącej. Przyjętą praktyką budowy modeli statystycznych jest równoważenie zbioru uczącego, tak aby proporcja modelowanych przypadków należących do klasy pozytywnej i negatywnej była zbliżona.

Drzewa wzmocnione nie wymagają od analityków aż tak dużego zaangażowania w proces czyszczenia danych. W sposób naturalny obsługują dane zarówno jakościowe, jak i ilościowe oraz są w dużym stopniu niewrażliwe na występowanie wartości odstających i rzadkich klas. Są również odporne na występowanie danych bez znaczenia oraz cech nadmiernie ze sobą skorelowanych. Tak więc budując model za pomocą tej metody, możemy zaniedbać szereg czasochłonnych i pracochłonnych etapów analizy danych, z drugiej jednak strony podejmujemy ryzyko, że szумы i nieprawidłowości zawarte w danych (niezidentyfikowane na etapie ich czyszczenia) spowodują, że model dawał będzie obciążone wyniki. Metoda ta jest również dosyć odporna na występowanie w zbiorze danych niezrównoważonych proporcji przypadków dobrych i złych (czego nie można jednak powiedzieć o innych metodach data mining, takich jak na przykład sieci neuronowe). Oczywiście nic nie zastąpi naszej pomysłowości i inwencji w procesie opracowania zmiennych pochodnych, jednak efekty będące interakcjami pomiędzy poszczególnymi cechami metoda ta identyfikuje w sposób automatyczny.

Przykład skoringu medycznego

Opisane na wstępie zastosowanie modeli skoringowych w medycynie zostanie zaprezentowane na przykładzie zbioru danych *WDBC.sta* (*Wisconsin Diagnostic Breast Cancer*)¹⁵ zawierającego informacje na temat wycinka obrazu tkanki pobranej z piersi kobiet za pomocą biopsji cienkoigłowej. Plik danych zawiera informacje o 569 badaniach, przedstawione za pomocą 31 zmiennych. Zmienna *Typ nowotworu* zawiera informację o diagnozie i będzie pełniła w analizie rolę zmiennej zależnej. Zmienna ta przyjmuje dwie wartości: wartość M (*malignant*) informuje o wystąpieniu zmian złośliwych, natomiast B (*benign*) informuje o wystąpieniu zmian łagodnych. Kolejne 30 zmiennych to informacje uzyskane

¹⁵ Pochodzące ze strony UCI Machine Learning Repository <http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>



na podstawie analizowanych jąder komórkowych widocznych w wycinkach. Są to charakterystyki dotyczące obwodu, średnicy, gładkości i tym podobnych parametrów jądra (zmierzono 10 charakterystyk jąder komórkowych). Parametry te przedstawiono za pomocą trzech miar: średniej wartości, odchylenia standardowego oraz średniej z trzech „najgorszych” wartości, co daje łącznie 30 cech diagnostycznych. Na podstawie tych cech przygotowujemy model, którego zadaniem będzie wsparcie procesu diagnozy pacjentów.

W pierwszym kroku musimy podjąć decyzję o wyborze metody analitycznej. Ze względu na łatwość interpretacji oraz możliwość oceny siły wpływu poszczególnych cech na uzyskany wynik zdecydujemy się na model oparty na regresji logistycznej, dodatkowo do jego przygotowania wykorzystamy metodologię budowy kart skoringowych, aby jeszcze bardziej uprościć postać modelu i ułatwić jego interpretację.

Konsekwencją wyboru regresji logistycznej jako metody modelowania jest konieczność przeprowadzenia pewnych czynności związanych ze wstępną analizą danych. Pierwszym z nich będzie eliminacja tych cech diagnostycznych, które są nieistotne z punktu widzenia wpływu na analizowaną zmienną zależną. Aby dokonać tego wyboru skorzystamy z modułu *Wybór predyktorów* zawartego w *STATISTICA Zestaw Skoringowy* – dedykowanym programie wspierającym budowę, ocenę i monitorowanie modeli skoringowych. Z menu *Zestaw skoringowy* wybieramy opcję *Wybór predyktorów*. Aby ocenić siłę wpływu poszczególnych cech, w oknie o tej samej nazwie przechodzimy na kartę *Ranking predyktorów*, a następnie wybieramy zmienne do analizy. Zmienną zależną będzie zmienna *Typ nowotworu*, a pozostałe zmienne będą zmiennymi niezależnymi (wybieramy je na dwóch listach w zależności od skali pomiaru). Ranking predyktorów wykonany zostanie na podstawie miar IV (*Information Value*) [1] oraz V Cramera. Po zatwierdzeniu analizy otrzymujemy gotowy ranking predyktorów.

Widzimy, że przy zastosowaniu kryterium IV zmienną, która najmocniej wpływa na możliwość dyskryminacji typu nowotworu, jest zmienna *największa wklęsłość*. Większość zmiennych znajdujących się w zbiorze danych również bardzo mocno wpływa na zmienną zależną. Przyjmijmy kryterium odrzucenia zmiennych z dalszej analizy (tym samym uznania ich za nieistotne), gdy wskaźnik IV jest mniejszy od 0,08. 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.

Wprowadzenie tego kryterium spowodowało usunięcie z listy predyktorów zmiennych *SE Struktury* oraz *SE Symetria*, czyli jedynie w niewielkim stopniu pozwoliło nam ograniczyć liczbę zmiennych wykorzystywanych w analizie. Aby ograniczyć zbiór danych tylko do istotnych predyktorów, klikamy przycisk *Podzbiór*, otrzymując nowy arkusz danych do analizy.

Kolejnym problemem, jaki należy zbadać, jest kwestia występowania nadmiernej korelacji analizowanych cech. Budując model, musimy pamiętać, że żadna ze zmiennych niezależnych nie może być liniową funkcją pozostałych zmiennych. Powinniśmy również unikać uwzględnienia w modelu zmiennych mocno ze sobą skorelowanych. Prowadzi to do zawyżenia błędów standardowych, a więc fałszywej istotności analizowanych zmiennych [2].

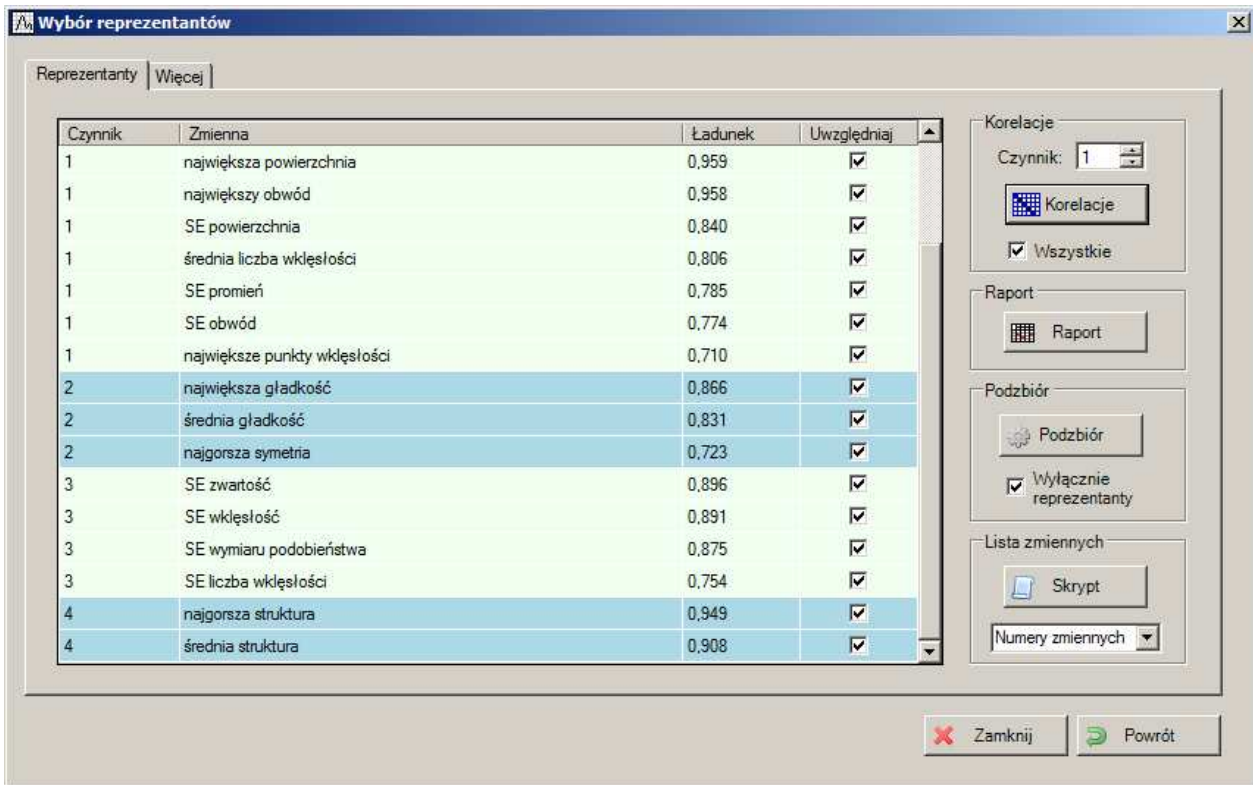


| Nr | Nazwa | IV | V Cramera | Uwzględnij |
|----|------------------------------|------|-----------|-------------------------------------|
| 1 | największa wklęsłość | 3,57 | 0,74 | <input checked="" type="checkbox"/> |
| 2 | średnia wklęsłość | 3,41 | 0,79 | <input checked="" type="checkbox"/> |
| 3 | średni obwód | 2,88 | 0,81 | <input checked="" type="checkbox"/> |
| 4 | średnia liczba wklęsłości | 2,75 | 0,84 | <input checked="" type="checkbox"/> |
| 5 | średnia powierzchnia | 2,69 | 0,80 | <input checked="" type="checkbox"/> |
| 6 | średni promień | 2,65 | 0,80 | <input checked="" type="checkbox"/> |
| 7 | SE powierzchnia | 2,39 | 0,77 | <input checked="" type="checkbox"/> |
| 8 | największe punkty wklęsłości | 2,34 | 0,85 | <input checked="" type="checkbox"/> |
| 9 | najgorsza zawartość | 2,31 | 0,64 | <input checked="" type="checkbox"/> |
| 10 | SE promień | 2,21 | 0,68 | <input checked="" type="checkbox"/> |
| 11 | średnia zawartość | 2,21 | 0,63 | <input checked="" type="checkbox"/> |
| 12 | największy obwód | 2,16 | 0,88 | <input checked="" type="checkbox"/> |
| 13 | największa powierzchnia | 1,85 | 0,86 | <input checked="" type="checkbox"/> |
| 14 | największy promień | 1,84 | 0,86 | <input checked="" type="checkbox"/> |
| 15 | SE liczba wklęsłości | 1,35 | 0,51 | <input checked="" type="checkbox"/> |
| 16 | średnia struktura | 1,28 | 0,50 | <input checked="" type="checkbox"/> |
| 17 | najgorsza struktura | 1,18 | 0,48 | <input checked="" type="checkbox"/> |
| 18 | SE obwód | 1,01 | 0,66 | <input checked="" type="checkbox"/> |
| 19 | najgorsza symetria | 1,01 | 0,46 | <input checked="" type="checkbox"/> |

Aby zidentyfikować cechy nadmiernie ze sobą skorelowane, możemy oprzeć się na analizie macierzy korelacji lub skorzystać z dedykowanego modułu *Zestawu skoringowego* służącego do wyboru reprezentantów. My skorzystamy z tej drugiej możliwości. Z menu *Zestaw Skoringowy* wybieramy opcję *Wybór predyktorów*. Następnie na karcie *Wybór reprezentantów* klikamy *Zmienne*, aby wybrać zmienne do analizy i wybieramy wszystkie predyktory.

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 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 fragment macierzy korelacji zmiennych wchodzących w skład czynnika pierwszego.

| Zmienna | średnia powierzchnia | średni promień | średni obwód | największy promień | największa powierzchnia | największy obwód |
|-------------------------|----------------------|----------------|--------------|--------------------|-------------------------|------------------|
| średnia powierzchnia | 1,00 | 0,99 | 0,99 | 0,96 | 0,96 | 0,96 |
| średni promień | 0,99 | 1,00 | 1,00 | 0,97 | 0,94 | 0,97 |
| średni obwód | 0,99 | 1,00 | 1,00 | 0,97 | 0,94 | 0,97 |
| największy promień | 0,96 | 0,97 | 0,97 | 1,00 | 0,98 | 0,99 |
| największa powierzchnia | 0,96 | 0,94 | 0,94 | 0,98 | 1,00 | 0,98 |
| największy obwód | 0,96 | 0,97 | 0,97 | 0,99 | 0,98 | 1,00 |

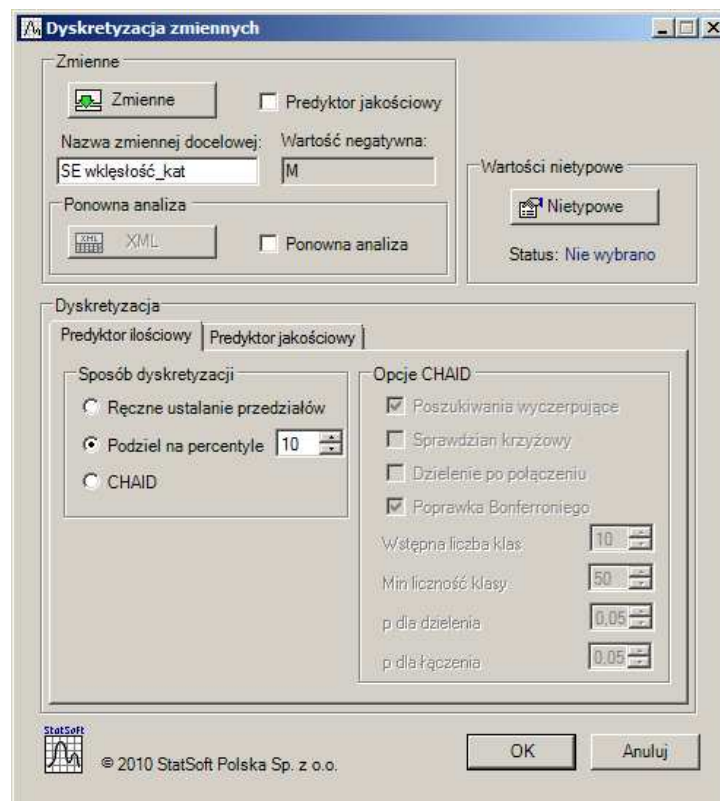
Widzimy bardzo wysoką korelację pomiędzy zmiennymi, pozwalającą na ich bezpieczną eliminację i pozostawienie jedynie jednej z nich. Aby usunąć zmienne, odznaczamy pole *Uwzględnij* w wierszach odpowiadających tym zmiennym, a następnie klikamy *Podzbiór*,



by wygenerować zbiór danych bez usuniętych zmiennych.¹⁶ Procedura ta jest bardzo przydatna zwłaszcza w sytuacji, gdy nasz zbiór danych zawiera bardzo dużą liczbę wskaźników, które są ze sobą mocno skorelowane, a ich liczba uniemożliwia efektywną analizę globalnej macierzy korelacji.

Usunięcie nieistotnych oraz nadmiernie skorelowanych zmiennych zawężyło liczbę potencjalnych predyktorów do 12. Na ich podstawie w kolejnych etapach analizy będziemy budowali końcowy model. Przed rozpoczęciem jego budowy wykonamy ostatni krok wstępnej analizy danych, czyli dyskretyzację zmiennych. Naszym celem będzie wyróżnienie w każdej ze zmiennych grup jednorodnych ze względu na ryzyko wystąpienia nowotworu i na tej podstawie przygotowanie zmiennych pochodnych, które będą wykorzystane do finalnej analizy. Analiza ta pozwoli nam lepiej zrozumieć charakter analizowanych zmiennych, wygładzić szумы, jakie występują w danych, a także wyeliminować negatywny wpływ obserwacji odstających. Co ważne, podejście to w sposób naturalny pozwala obsłużyć braki danych.

Aby przygotować profile ryzyka predyktorów, skorzystamy z modułu *Dyskretyzacja zmiennych* zawartego w *Zestawie Skoringowym*.



W oknie *Dyskretyzacja zmiennych* wskazujemy zmienną *Typ Nowotworu* jako zmienną stanu, natomiast dyskretyzację rozpoczniemy od zmiennej *SE_wkleslosc*. Przed analizą

¹⁶ Klikając przycisk *Skrypt*, możemy wygenerować makro selekcji zmiennych, którego uruchomienie spowoduje analogiczne działanie - *STATISTICA* zawiera zaimplementowany język makr oparty na Visual Basic – zgodny z językiem makr pakietu Office.



określamy jeszcze klasę M zmiennej *Typ nowotworu* jako klasę negatywną (wystąpił nowotwór złośliwy), a następnie dzielimy wartości zmiennej *SE_wklęsłość* na percentyle.

Statystyki Opisowe
N Waznych: 569 Minimum: 0,0000 Maksimum: 0,3960

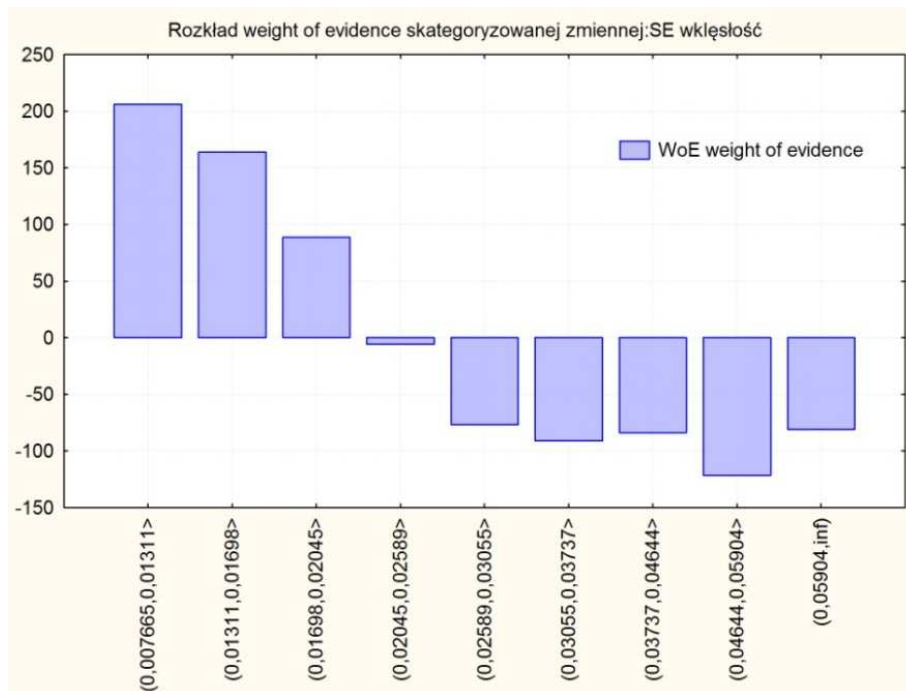
| | Od | Formuła | Do | Kategoria | Liczność | Scal |
|---|----------|----------|----------|--------------------|----------|--------------------------|
| ▶ | | $x \leq$ | 0,007665 | {-inf,0,007665} | | <input type="checkbox"/> |
| | 0,007665 | $x \leq$ | 0,01311 | {0,007665,0,01311} | | <input type="checkbox"/> |
| | 0,01311 | $x \leq$ | 0,01698 | {0,01311,0,01698} | | <input type="checkbox"/> |
| | 0,01698 | $x \leq$ | 0,02045 | {0,01698,0,02045} | | <input type="checkbox"/> |
| | 0,02045 | $x \leq$ | 0,02589 | {0,02045,0,02589} | | <input type="checkbox"/> |
| | 0,02589 | $x \leq$ | 0,03055 | {0,02589,0,03055} | | <input type="checkbox"/> |
| | 0,03055 | $x \leq$ | 0,03737 | {0,03055,0,03737} | | <input type="checkbox"/> |
| | 0,03737 | $x \leq$ | 0,04644 | {0,03737,0,04644} | | <input type="checkbox"/> |
| | 0,04644 | $x \leq$ | 0,05904 | {0,04644,0,05904} | | <input type="checkbox"/> |
| | 0,05904 | $x \leq$ | | {0,05904,inf} | | <input type="checkbox"/> |

Wartości nietypowe

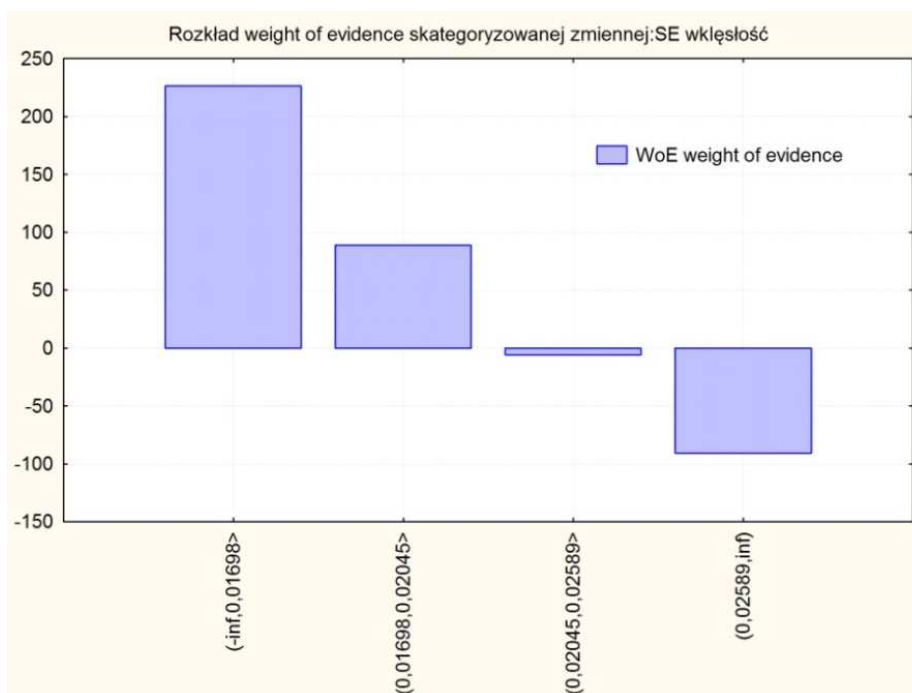
| Wartość nietypowa | Kategoria | Liczność |
|-------------------|-----------|----------|
|-------------------|-----------|----------|

Przekoduj

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



Dla każdej kategorii zmiennej *SE wklęsłość* obliczono miarę szansy, że przypadki danej kategorii są zmianami łagodnymi - *Weight of Evidence* (w polskiej nomenklaturze spotyka się niekiedy termin „waga dowodu”). Wyższe wartości *WoE* informują o większym prawdopodobieństwie łagodnych zmian. Przykładowo na podstawie wykresu widzimy, iż najmniejsze ryzyko wystąpienia nowotworu złośliwego występuje u osób, dla których wartość zmiennej *SE wklęsłość* jest mniejsza niż 0,013. Ryzyko to stopniowo się zwiększa wraz ze wzrostem wartości zmiennej *SE wklęsłość*. Ponieważ pięć ostatnich kategorii ma bardzo zbliżoną wartość *WoE*, przyjmujemy założenie, że są to różnice wynikające z niedoskonałości analizowanej próby, a nie z rzeczywistych zmian wpływu na ryzyko, stąd też scalimy je do wspólnej kategorii. W oknie *Przekoduj ilościowe* w odpowiednich kategoriach zmiennej zaznaczamy pola wyboru, a następnie klikamy przycisk *Scal*. W analogiczny sposób scalimy też początkowe kategorie. Po scaleniu profil zmiennej *SE wklęsłość* wygląda następująco:



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

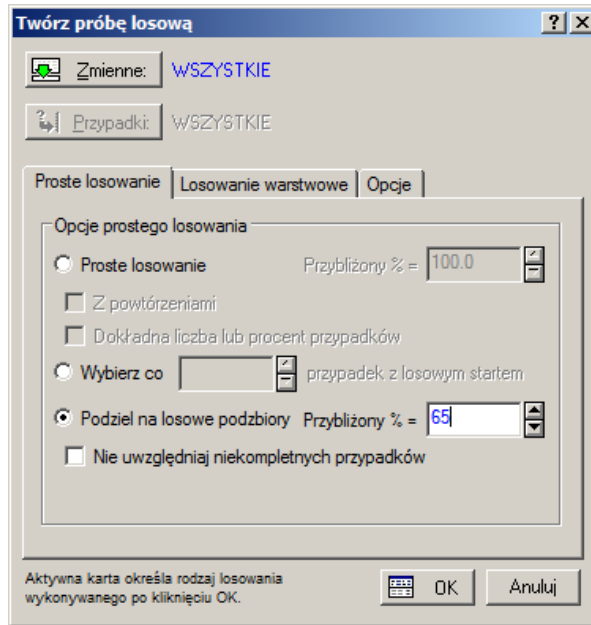
Dyskretyzacja zmiennych, chociaż może osłabić moc predykcyjną poszczególnych zmiennych, niesie ze sobą zdecydowanie więcej korzyści:

- ◆ 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),

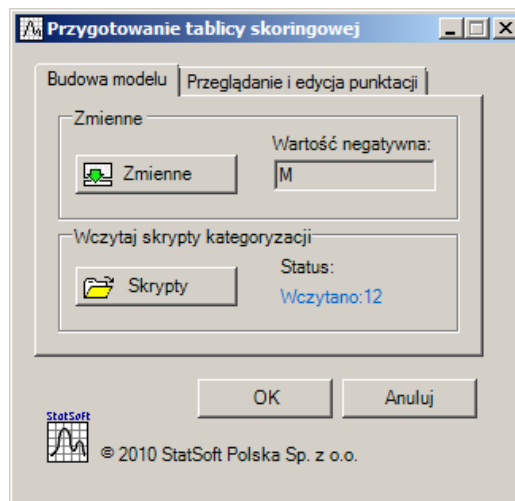


- ◆ dyskretyzacja zmiennych pozwala również wychwycić wiele błędów i sprzeczności występujących w danych.

Po przygotowaniu zmiennych do analizy przechodzimy do fazy modelowania. Aby zachować zgodność 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 zrobić nam to za pomocą opcji *Próbkowanie losowe* znajdującej się w menu *Dane*.



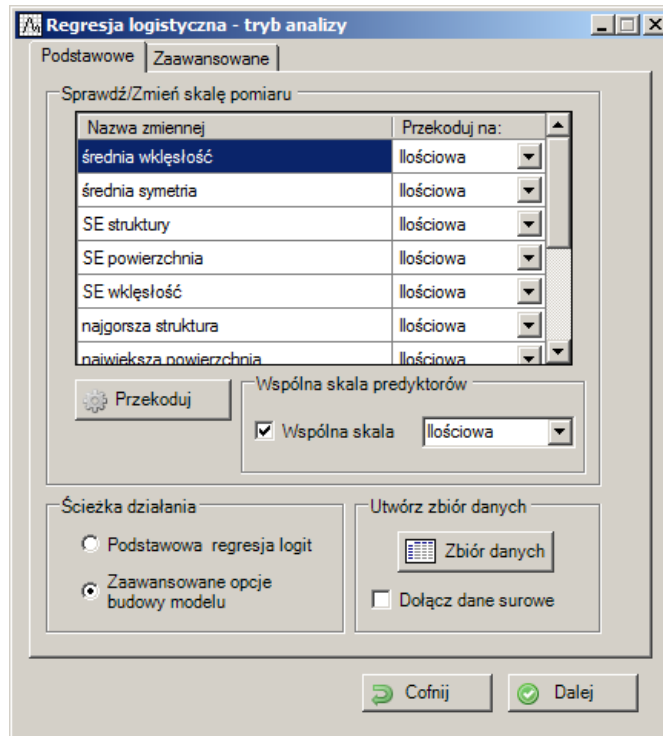
W oknie *Twórz próbę losową* zaznaczamy opcję *Podziel na losowe podzbiory* i określamy, by zbiór uczący zawierał 65% przypadków. Po zatwierdzeniu analizy nasz zbiór zostanie podzielony na dwa losowo określone podzbiory. Mniejszy z nich (prawie 200 przypadków) odłożymy do celów testowych, natomiast większy (około 370 przypadków) posłuży 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*.

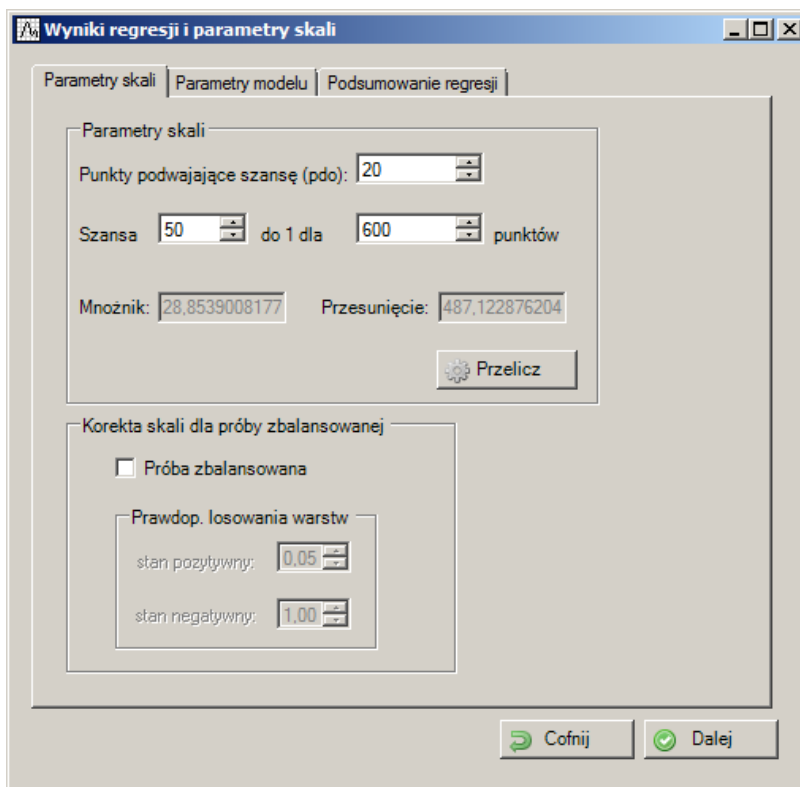


W oknie *Regresja logistyczna – tryb analizy* klikamy *Przekoduj*, aby przygotować dyskretyzację (poszczególne wartości zostaną zamienione odpowiadającym im wartościami *WoE*). Po przekodowaniu zmiennych przechodzimy na kartę *Zaawansowane* i wybieramy opcję *Krokowa wsteczna* jako sposób budowy modelu,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

| Dane: Oceny_parametrów | | | | | | |
|---|--------|---------------|-------------|----------------|----------------|--------|
| Typ Nowotworu - Oceny parametrów (Arkuszt905) | | | | | | |
| Rozkład: DWUMIANOWY , F. wiążąca: LOGIT | | | | | | |
| Modelowane prawdopodobieństwo Typ Nowotworu = B | | | | | | |
| Efekt | Ocena | Standard Błąd | Walda Stat. | GU górna 95, % | GU dolna 95, % | p |
| Wyraz wolny | 0,2365 | 0,3190 | 0,5496 | -0,3888 | 0,8618 | 0,4585 |
| najgorsza struktura_kat | 0,0133 | 0,0034 | 15,3059 | 0,0066 | 0,0200 | 0,0001 |
| największa powierzchnia_kat | 0,0079 | 0,0017 | 20,5406 | 0,0045 | 0,0113 | 0,0000 |
| SE powierzchnia_kat | 0,0058 | 0,0022 | 6,7025 | 0,0014 | 0,0102 | 0,0096 |
| największe punkty wklęsłości_kat | 0,0045 | 0,0016 | 7,3893 | 0,0012 | 0,0077 | 0,0066 |
| największa gładkość_kat | 0,0131 | 0,0042 | 9,7041 | 0,0048 | 0,0213 | 0,0018 |
| średnia wklęsłość_kat | 0,0055 | 0,0017 | 10,2973 | 0,0021 | 0,0088 | 0,0013 |
| Skala | 1,0000 | 0,0000 | | 1,0000 | 1,0000 | |

By oszacować parametry regresji logistycznej, klikamy przycisk dalej, po czym w oknie *Wyniki regresji i parametry skali* możemy przejrzeć uzyskane wyniki. Na przykład wartości ocen parametrów regresji uzyskane w wyniku analizy widoczne są powyżej.

Raport *Budowanie modelu* umożliwia prześledzenie procesu doboru parametrów. Proces zakończył się w siódmej iteracji po odrzuceniu z modelu zmiennych *SE wklęstość*, *średnia symetria*, *SE struktury*, *najgorszy wymiar podobieństwa*, *najgorsza symetria* oraz *największa wklęstość*. 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.



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

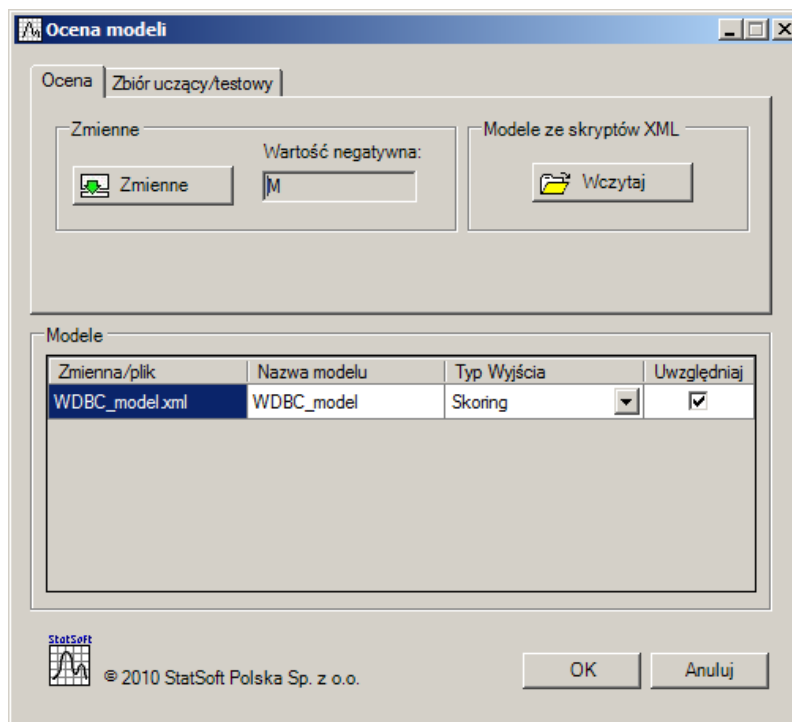
W wyniku przekształcenia ocen parametrów regresji logistycznej otrzymujemy tablicę skoringową, w której poszczególnym kategoriom zmiennych modelu przypisano określoną liczbę punktów, które interpretujemy w kategoriach szans (*ODDS*) wystąpienia nowotworu łagodnego zgodnie z podanymi parametrami skali. Zbudowany model zapisujemy do pliku *WDBC_model.xml*, który wykorzystamy do oceny i kalibracji modelu.

Zauważmy, że postać modelu jest bardzo łatwa w interpretacji i stosowaniu. Każdej kategorii zmiennych w modelu przypisana została określona liczba punktów. Wynikiem modelu jest suma punktów dla poszczególnych kategorii, do jakich trafiły wyniki badań analizowanego pacjenta.



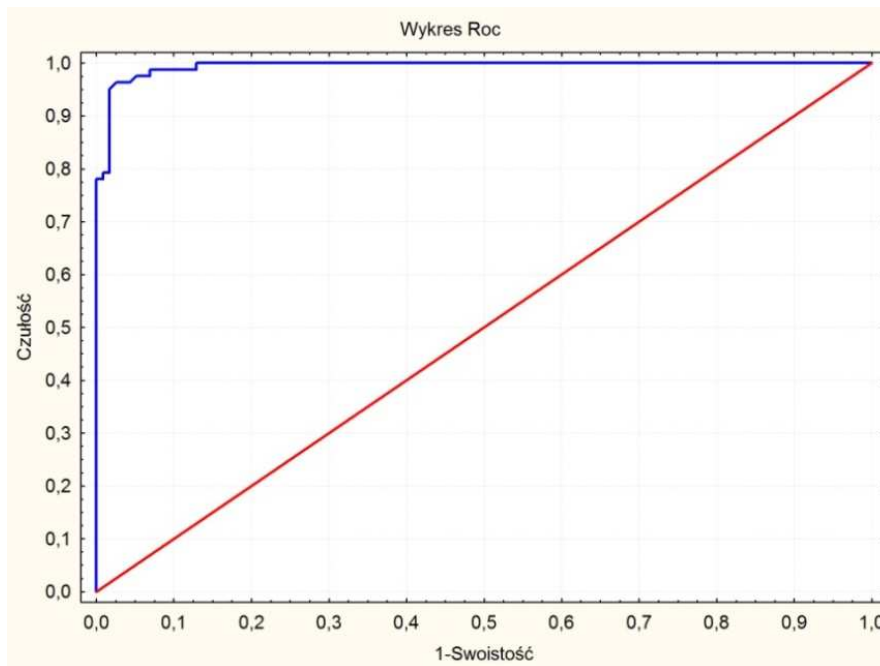
| Zmienna | Zakres | WoE | Ocena | s. Walda | p | Skoring | Skoring zaokr. |
|------------------|----------------|----------|---------|----------|---------|---------|----------------|
| najgorsza str... | {-inf;23,58> | 152,902 | 0,01332 | 15,30594 | 0,00664 | 141,090 | 141 |
| najgorsza str... | (23,58;28,46> | -29,801 | 0,01332 | 15,30594 | 0,00664 | 70,871 | 71 |
| najgorsza str... | (28,46;inf) | -112,729 | 0,01332 | 15,30594 | 0,00664 | 38,999 | 39 |
| najgorsza str... | Wartość ne... | - | - | | | 90,276 | 90 |
| największa ... | {-inf;728,3> | 279,469 | 0,00785 | 20,54056 | 0,00446 | 145,625 | 146 |
| największa ... | (728,3;830,9> | 1,785 | 0,00785 | 20,54056 | 0,00446 | 82,729 | 83 |
| największa ... | (830,9;inf) | -282,373 | 0,00785 | 20,54056 | 0,00446 | 18,366 | 18 |
| największa ... | Wartość ne... | - | - | | | 97,062 | 97 |
| SE powierzc... | {-inf;29,44> | 168,583 | 0,00578 | 6,70249 | 0,00140 | 110,440 | 110 |
| SE powierzc... | (29,44;inf) | -183,173 | 0,00578 | 6,70249 | 0,00140 | 51,776 | 52 |
| SE powierzc... | Wartość ne... | - | - | | | 87,038 | 87 |
| największe ... | {-inf;0,1218> | 220,888 | 0,00446 | 7,38929 | 0,00124 | 110,750 | 111 |
| największe ... | (0,1218;inf) | -221,807 | 0,00446 | 7,38929 | 0,00124 | 53,781 | 54 |
| największe ... | Wartość ne... | - | - | | | 89,406 | 89 |
| największa ... | {-inf;0,1124> | 161,892 | 0,01307 | 9,70411 | 0,00485 | 143,377 | 143 |
| największa ... | (0,1124;0,1... | 46,971 | 0,01307 | 9,70411 | 0,00485 | 100,038 | 100 |
| największa ... | (0,1377;inf) | -97,107 | 0,01307 | 9,70411 | 0,00485 | 45,704 | 46 |
| największa ... | Wartość ne... | - | - | | | 86,638 | 87 |
| średnia wkłę... | {-inf;0,0862> | 197,333 | 0,00549 | 10,29727 | 0,00214 | 113,584 | 114 |
| średnia wkłę... | (0,0862;0,1... | -72,879 | 0,00549 | 10,29727 | 0,00214 | 70,780 | 71 |
| średnia wkłę... | (0,1122;inf) | -278,551 | 0,00549 | 10,29727 | 0,00214 | 38,200 | 38 |
| średnia wkłę... | Wartość ne... | - | - | | | 87,315 | 87 |

Kolejnym krokiem analizy będzie ocena zbudowanego modelu oraz określenie najbardziej odpowiedniego punktu bądź punktów odcięcia. Aby ocenić zbudowany model, wykorzystamy plik *Testowy.sta* zawierający przypadki, które nie brały udziału w procesie szacowania modelu.

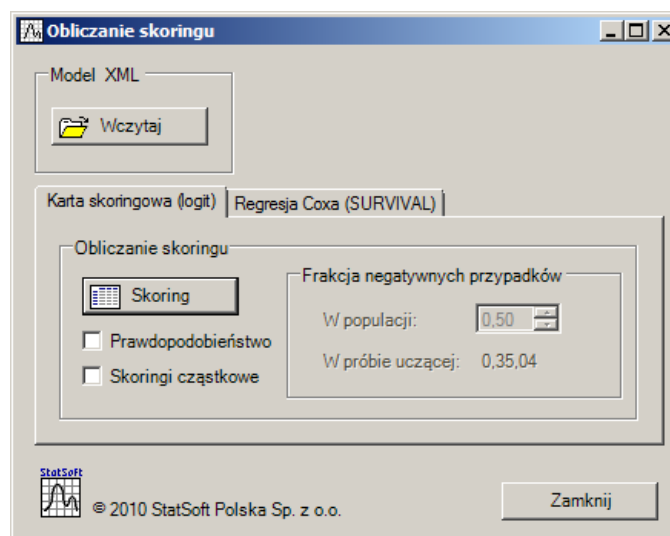


W tym celu z menu *Zestaw Skoringowy* wybieramy opcję *Ocena modeli*, następnie wybieramy zmienną *Typ nowotworu* jako zmienną zależną oraz wczytujemy plik *WDBC_model.xml* zawierający specyfikację zbudowanego modelu.

Po zatwierdzeniu analizy w oknie *Ocena modeli – wyniki* klikamy przycisk *Wskaźniki*, aby otrzymać podsumowanie jakości modeli. Analizując przebieg krzywej ROC, możemy zauważyć, że model niemal idealnie dyskryminuje dobre i złe przypadki. Pole powierzchni pod krzywą jest bliskie 1.



Zbudowany model jest więc bardzo dobrej jakości i może być przydatny jako narzędzie wspierające diagnozę lekarską. W ostatnim kroku przeprowadzimy jeszcze analizę punktu odcięcia. Będziemy szukać takiego miejsca lub miejsc w punktacji, które podzielią analizowanych pacjentów na grupy zdrowych i chorych, bądź częściej na grupy zdrowych, chorych i niemożliwych do zdiagnozowania.





Aby określić optymalny punkt odcięcia, w pierwszej kolejności obliczymy wartość skoringu dla każdego przypadku ze zbioru testowego. Z menu *Zestaw skoringowy* wybieramy opcję *Obliczanie skoringu* i po wczytaniu pliku *WDBC_model.xml* klikamy przycisk *Skoring*, co spowoduje obliczenie odpowiedzi modelu dla każdego z analizowanych przypadków w dodatkowej zmiennej o takiej samej nazwie.

Na podstawie wygenerowanej zmiennej *Skoring* oraz zmiennej zależnej *Typ nowotworu* dokonamy analizy punktu odcięcia, korzystając z modułu *Dyskretyzacja zmiennych* w sposób analogiczny do przedstawionego w części dotyczącej przygotowanie danych. Na podstawie uzyskanych wyników analizy wyróżniono trzy punkty odcięcia dzielące pacjentów na cztery grupy ryzyka.

| WDBC Skoring | Dobry | Zły | Suma | Procent ogółem |
|--------------|-------|-----|------|----------------|
| (-inf,349> | 0 | 60 | 60 | 30,30% |
| (349,483> | 2 | 18 | 20 | 10,10% |
| (483,589> | 16 | 4 | 20 | 10,10% |
| (589,765> | 98 | 0 | 98 | 49,49% |
| Ogół | 116 | 82 | 198 | 100,00% |

Osoby, które uzyskały do 349 punktów, możemy z całą pewnością zaklasyfikować do grupy osób z nowotworem złośliwym, osoby, które uzyskały powyżej 589 punktów, klasyfikujemy do grupy osób ze zmianami łagodnymi. Pozostałe osoby w zależności od punktacji możemy zaklasyfikować do grup wysokiego i niskiego ryzyka, jednak z całą pewnością nie możemy postawić dla nich jednoznacznej diagnozy.

Analizując uzyskane wyniki możemy stwierdzić, że zbudowany model może być użyteczny w procesie diagnozy nowotworu piersi, podaje bezbłędną odpowiedź dla prawie 80% przypadków. Pozostałe przypadki wymagają dodatkowych czynności diagnostycznych.

Przykład skoringu marketingowego

Modele skoringowe są jednym z najczęściej stosowanych podejść analitycznych pozwalających na optymalizację kampanii marketingowych. Model taki umożliwia wskazanie klientów najbardziej skłonnych odpowiedzieć na naszą ofertę. Aby taki model zbudować, musimy mieć dostęp do danych na temat naszych klientów – mogą to być zarówno cechy demograficzne, jak i opisujące zachowanie klientów, oraz produkty, z jakich korzystają itp. Wykorzystując te dane oraz zmienną informującą o zakupie interesującego nas produktu, która w modelowaniu pełni rolę zmiennej zależnej, model określa wzorce zachowań klientów. Jeśli wzorce wychwycone przez model okażą się wartościowe, możemy je następnie zastosować dla nowych klientów. Celem budowy modelu jest określenie, jaki produkt powinien zostać zaoferowany konkretnemu klientowi i jaki jest najlepszy kanał komunikacji z klientem.



Przykład budowy modelu na potrzeby kampanii marketingowej zaprezentujemy na podstawie zbioru danych *SkoringMarketingowy.sta* zawierającego informację o około 6500 osobach. Zbiór ten zawiera zmienną *Odp* informującą o przyjęciu lub odrzuceniu przez klienta oferty na pewien produkt. Zmienna ta przyjmuje wartości *Tak* w 16% przypadków, pozostali klienci odrzucili zaproponowaną ofertę. Dodatkowo analizowany zbiór jest opisany przez ponad 200 zmiennych charakteryzujących każdego z klientów. Zmienne opisujące każdego z klientów możemy podzielić na trzy grupy:

- ◆ informacje o wcześniejszych zakupach (np. kiedy ostatnio został zakupiony określony produkt) i inne informacje uzyskane przy ich realizacji (np. płeć),
- ◆ informacje spisowe dotyczące obszaru, w jakim zamieszkuje dana osoba (np. liczba rodzin z jednym rodzicem),
- ◆ informacje podatkowe dotyczące obszaru, w jakim zamieszkuje dana osoba (np. liczba podatników o dochodzie większym niż zero).

Ponieważ nie interesuje nas wyjaśnienie kierunku oraz siły wpływu poszczególnych zmiennych na model ani zrozumienie ogólnego charakteru analizowanego zjawiska, a naszym głównym zadaniem jest zbudowanie skutecznego modelu, model skorigowy zbudujemy za pomocą metod data mining.

Ze względu na dużą liczbę zmiennych nie będziemy brali pod uwagę metod czułych na tę kwestię, czyli sieci neuronowych oraz metody wektorów wspierających (SVM). Wydaje się, że najbardziej odpowiednią metodą data mining, jaką powinniśmy zastosować, są opisywane we wcześniejszej części artykułu drzewa wzmacniane. Głównym argumentem przemawiającym za tą metodą jest jej odporność na wartości odstające, zmienne nadmierne ze sobą skorelowane oraz zmienne bez znaczenia. Metoda ta jest równocześnie bardzo mocna predykcyjnie i znakomicie nadaje się do rozwiązywania problemów klasyfikacyjnych, gdzie wzorce zawarte w danych są rozmyte. Musimy pamiętać, iż w naszym zbiorze danych zawarty jest szereg cech spisowych i dotyczących obszaru, z którego pochodzą klienci. Siłą rzeczy informacje te będą mocno obciążone i często wprost niezgodne z rzeczywistą przynależnością danej osoby do danej religii czy grupy zawodowej.

Podobnymi charakterystykami cechuje się także inna metoda data mining – losowy las (*random forests*), jednak drzewa wzmacniane są dodatkowo bardziej odporne na nierówne proporcje klas zmiennej zależnej i właśnie tę metodę wykorzystamy do budowy modelu.

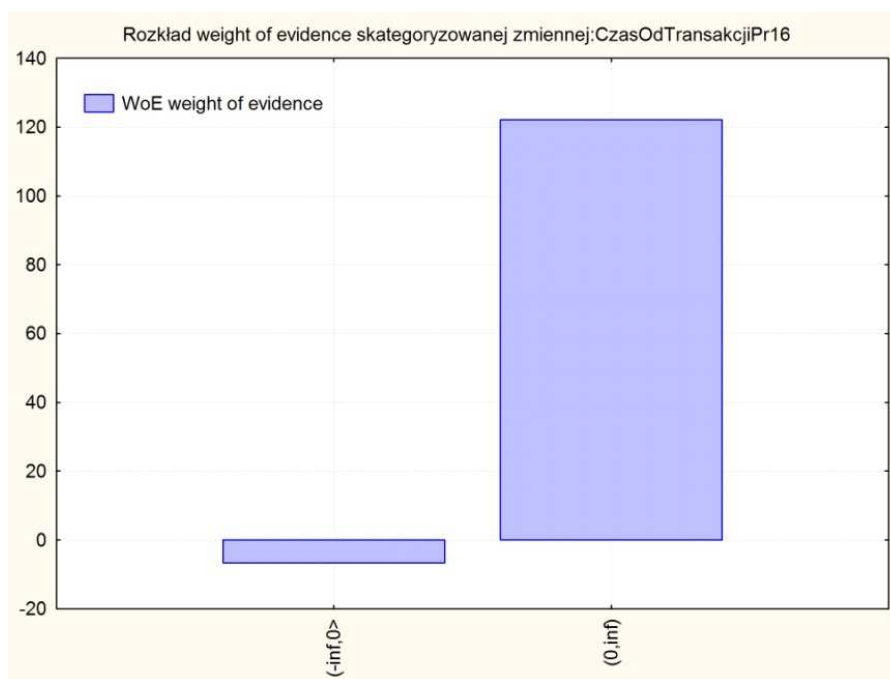
Choć wybrana metoda analityczna uwalnia nas od szeregu zadań związanych z czyszczeniem i przygotowaniem danych, jednak nie rozwiązuje automatycznie wszystkich problemów, jakie są w nich zawarte. Analizę danych rozpoczniemy więc od oceny rozkładów analizowanych zmiennych oraz zbadania poprawności ich zakresów. Aby wykonać wstępną analizę, możemy skorzystać z modułu *Statystyk opisowych* i dostępnych w nim szeregu opcji. Analizując uzyskane wyniki, stwierdzono, że wszystkie zmienne mają prawdopodobny zakres, w zbiorze danych nie występują braki danych i inne problemy mogące wpłynąć negatywnie na sam proces modelowania.

Jedynym poważniejszym problemem występującym w zbiorze danych, jaki mógłby wpłynąć negatywnie na proces modelowania, jest rozkład zmiennych informujących o czasie,

jaki upłynął od ostatniej transakcji zakupu innych produktów. W zbiorze danych zawartych zostało 17 tego typu zmiennych. Bliższa analiza siły wpływu poszczególnych predyktorów na skłonność do zakupu wykazała, że zdecydowanie większy wpływ niż sam czas zakupu ma sam fakt dokonania transakcji. Niestety wartości tych zmiennych w przeważającej części przypadków (powyżej 95%) wynoszą 0, co należy rozumieć jako brak zakupu.

Wzorce zawarte w tego typu zmiennych bardzo trudno jest wykorzystać w klasycznej analizie. Nawet jeżeli wzorec jest silny, to rzadkość jego występowania powoduje, że może on zostać pominięty w trakcie analizy.

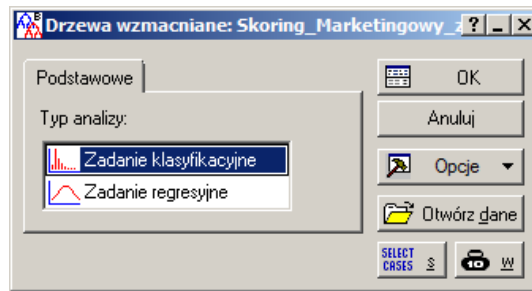
Korzystając z modułu *Dyskretyzacja zmiennych* programu *STATISTICA Zestaw Skoringowy*, oceniono siłę i kierunek wpływu poszczególnych zmiennych z tej grupy na zakup analizowanego produktu.



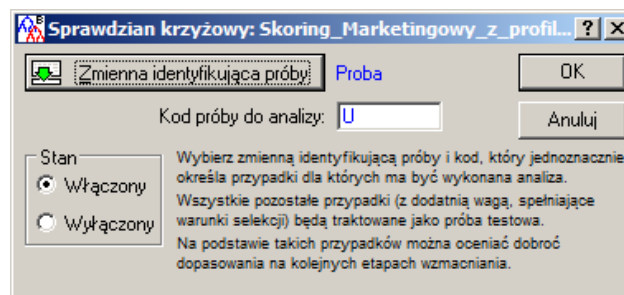
Przykładem takiej zmiennej jest zmienna *CzasOdTransakcjiPr16*. Jeżeli wartość tej zmiennej jest większa od zera, to skłonność klientów do zakupu analizowanego produktu mierzona za pomocą miary *Weight of Evidence* jest wysoka. Podobne wzorce występują w pozostałych zmiennych, dlatego też aby w pełni wykorzystać tkwiący w nich potencjał konieczne będzie przekształcenie tych zmiennych umożliwiające połączenie informacji z wielu zmiennych w jedną.

Scalenie informacji pochodzących ze zmiennych rzadkich przeprowadzono w programie *STATISTICA Zestaw Skoringowy*, dla każdej ze zmiennych przygotowano profil skłonności do zakupu, analogiczny do przedstawionego na powyższym rysunku. Następnie wartości analizowanych zmiennych przekodowano zgodnie z przygotowanymi profilami i zsumowano do jednej zmiennej, której nadano nazwę *CzasOdTransakcjiScalony*. Utworzona zmienna miała zadowalający rozkład i okazała się bardzo mocna predykcyjnie. Pierwotne zmienne, na podstawie których została utworzona, pominięto w analizie.

Moduł drzew wzmocnianych uruchamiamy za pomocą polecenia *Data Mining* -> *Wzmocniane drzewa klasyfikacyjne i regresyjne*.



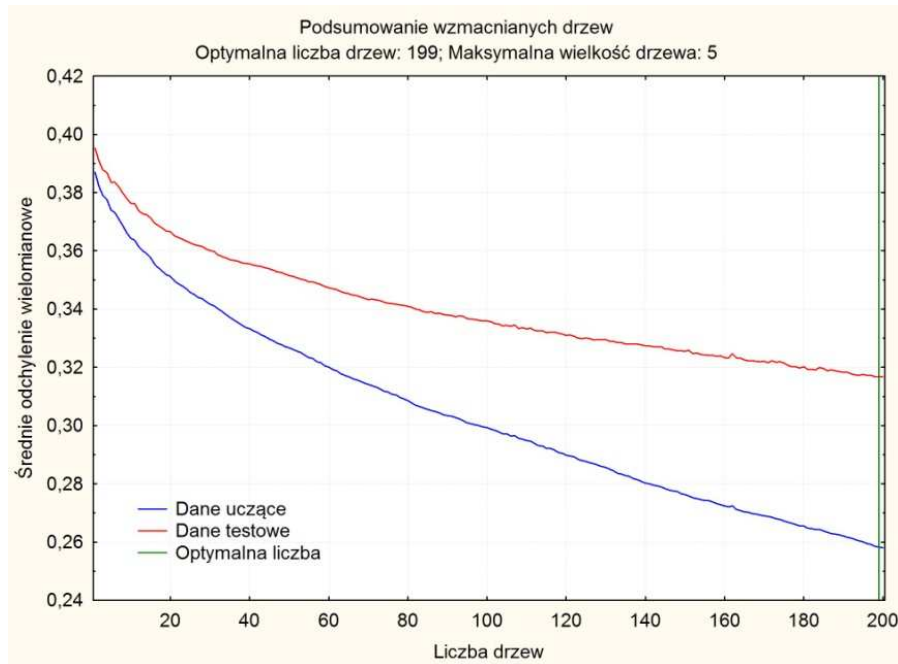
Następnie wybieramy typ analizy jako *Zadanie klasyfikacyjne*. W kolejnym kroku w oknie *Ustawienia drzew wzmocnianych* wskazujemy zmienne do analizy – zmienna *Odp* będzie pełniła rolę zmiennej zależnej, zmienną *Płeć* określamy jako predyktor jakościowy, natomiast pozostałe zmienne określamy jako predyktory ilościowe. Na karcie *Więcej* w tym samym oknie klikamy przycisk *Sprawdzian krzyżowy* i wskazujemy zmienną *Próba*, jako zmienną identyfikującą próbę uczącą i testową.



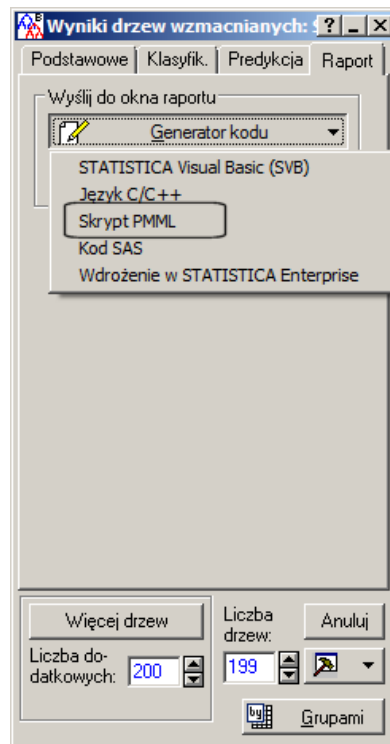
Zastosujemy więc tutaj klasyczne podejście do budowy modeli, w którym parametry modelu szacujemy na próbie uczącej, natomiast błąd oceniamy na podstawie próby testowej, która nie bierze udziału w procesie estymacji.

Również na karcie *więcej* zwiększamy parametr *Maksymalna liczba węzłów* na 5, aby model łatwiej wychwycił występujące w danych interakcje. Zatwierdzamy wykonanie analizy, klikając *OK*.

Z pośród szeregu dostępnych wyników zwróćmy uwagę na wykres obrazujący przebieg procesu estymacji. Widzimy, że dodanie kolejnych drzew do budowanego modelu spowodowało obniżanie się błędu zarówno na próbie uczącej, jak i na próbie testowej. Nachylenie obydwóch krzywych sugeruje potencjalną poprawę modelu po dołożeniu kolejnych elementów. Możemy więc w opcji wyników wskazać liczbę dodatkowych drzew, o które rozbudujemy model. Proces budowy modelu powinien zostać zatrzymany w momencie, w którym kolejna rozbudowa modelu powoduje wzrost błędu na próbie testowej.

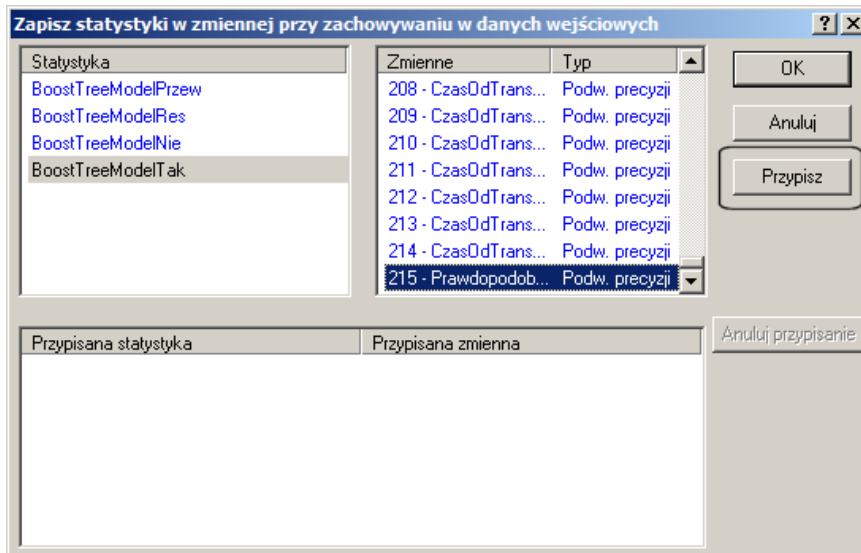


Po zbudowaniu modelu przechodzimy do etapu jego oceny. W pierwszej kolejności zapiszemy zbudowany model w formacie PMML specjalnym języku opartym na standardzie XML przeznaczonym do zapisu modeli statystycznych i data mining. Na karcie *Raport* okna *Wyniki drzew wzmacnianych* wybieramy opcję *Skrypt PMML*, a następnie zapisujemy model w wybranym formacie do pliku *Skoring.xml*. Zwróćmy uwagę na szereg innych formatów, do których możemy zapisać zbudowany model.





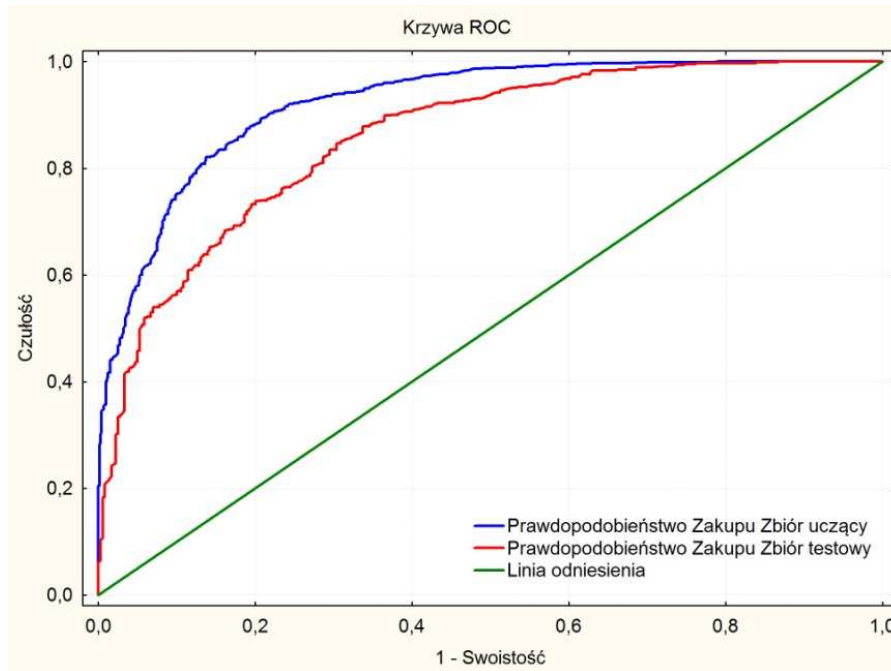
Za pomocą przygotowanego pliku obliczymy odpowiedzi modelu dla poszczególnych przypadków. Przed wykonaniem obliczeń do zbioru danych dodamy kolejną zmienną *Prawdopodobieństwo zakupu*, do której trafią uzyskane wyniki. Następnie z menu *Data mining* wybieramy opcję *Szybkie wdrażanie modeli predykcyjnych*. W wyświetlonym oknie wczytujemy model *PMML*, wskazujemy opcję *Wyniki zawierają prawdopodobieństwo*, a następnie za pomocą przycisku *Zapisz wartości lub klasy i reszty w pliku danych* przypisujemy odpowiedzi modelu do zmiennej *Prawdopodobieństwo zakupu* i wykonujemy obliczenia.



Po przeprowadzeniu obliczeń każdemu klientowi zawartemu w zbiorze danych została przypisana miara skłonności do zakupu oferowanego produktu. Miarę tę możemy interpretować jako prawdopodobieństwo zakupu: im wyższa jej wartość, tym większa skłonność danego klienta do zakupu.

Na jej podstawie przeprowadzimy ocenę jakości modelu w module *Ocena modeli* programu *STATISTICA Zestaw Skoringowy*. Do oceny wykorzystamy, podobnie jak w poprzednim przykładzie, popularną miarę jakości modeli skoringowych opartą na krzywej ROC. Obliczenia wykonujemy w sposób analogiczny do poprzedniego przykładu, z tą różnicą, że zamiast pliku z modelem wskazujemy zmienną *Prawdopodobieństwo zakupu*.

Analizując wartości pola powierzchni pod krzywą, możemy zaobserwować zauważalną różnicę w działaniu modelu pomiędzy zbiorem uczącym i testowym. Może to być sygnał o nadmiernym dopasowaniu modelu do danych, sugerujący konieczność uproszczenia modelu bądź w zakresie liczby drzew, bądź liczby węzłów poszczególnych modeli składowych. Z drugiej strony pole powierzchni pod krzywą ROC dla zbioru testowego wynosi 0,85, co świadczy o dobrym działaniu modelu. Możemy więc zaakceptować zbudowany model i przejść do fazy określenia najbardziej odpowiedniego sposobu jego wykorzystania.



Za pomocą modułu *Zarządzanie punktem odcięcia* zawartego w *Zestawie Skoringowym STATISTICA* określimy podział klientów na osoby, które powinny otrzymać ofertę na analizowany produkt, oraz te, do których nie należy jej kierować. Punkt ten nazywamy punktem odcięcia, a osoby, które cechują się skłonnością do zakupu poniżej wyznaczonej granicy zostaną pominięte w planowanej kampanii. Aby ten podział był możliwy, musimy określić, jaki jest koszt wysłania oferty do osoby, która jej nie przyjmie, oraz średni zysk, jaki spodziewamy się uzyskać od osób, które zaakceptują naszą ofertę.

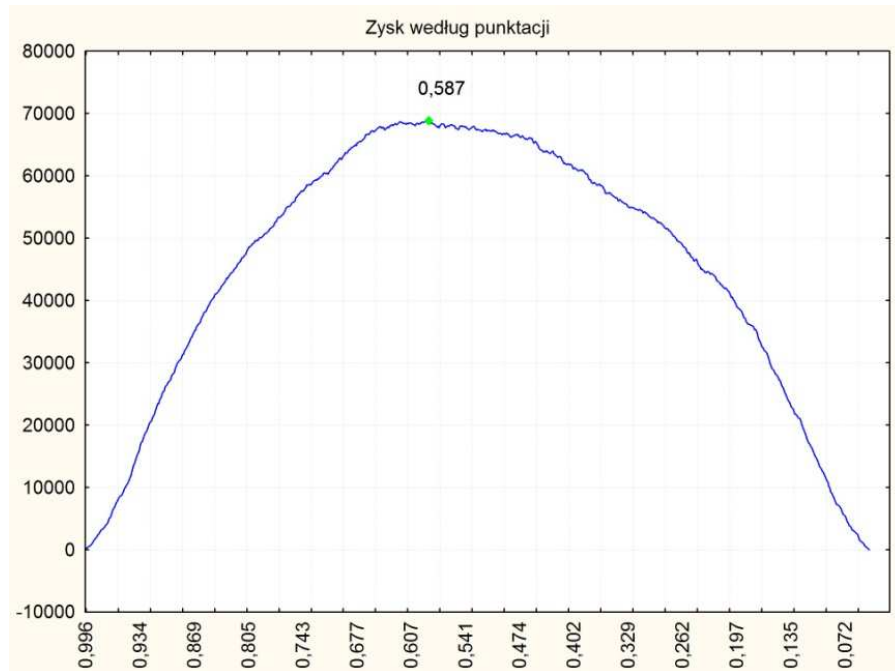
The screenshot shows the "Zarządzanie punktem odcięcia - wyniki" (Threshold Management - Results) window. It features a table with the following data:

| Punkt odcięcia | Poziom akceptacji | Procent złych |
|----------------|-------------------|---------------|
| 0,587 | 0,239 | 0,465 |

Below the table, there are two sections for scoring vs. amount:

- Skoring vs. kwota (zmienna)**:
 - Zysk dla akceptacji "dobrego": 5,5 %
 - Strata dla akceptacji "złego": 95,0 %
- Skoring vs. kwota (średnia)**:
 - Zysk dla akceptacji "dobrego": 100,0 %
 - Koszt akceptacji "złego": 20

Przyjmijmy, że koszt wysyłania oferty wyniesie 20 złotych, a zysk 100 złotych. Na podstawie tych założeń przeprowadzimy symulację działania modelu. Podane powyżej parametry wprowadzamy na karcie *Zysk* w obszarze *Skoring vs. kwota (średnia)*, a następnie klikamy przycisk *Raport*. Uzyskany w ten sposób wykres przedstawia symulację zysku w zależności od przyjętego punktu odcięcia.



Analizując przebieg krzywej, możemy stwierdzić, że najwyższy zysk z kampanii sprzedażowej osiągniemy, jeżeli zastosujemy punkt odcięcia na poziomie 0,587. Osoby, dla których model wyznaczył prawdopodobieństwo poniżej tej granicy, nie powinny uczestniczyć w kampanii. Stosując się do odpowiedzi modelu, możemy spodziewać się zysku na poziomie 70 tysięcy PLN. Wysłanie oferty do całej bazy klientów spowodowałoby, że zyski byłyby zerowe (wartość krzywej po prawej części wykresu).

Zauważmy, że zyski już z jednej kampanii z nawiązką pokrywają koszt narzędzia analitycznego, natomiast niezastosowanie modelu powoduje, że tego typu kampania nie byłaby opłacalna.

Zakończenie

Modele data mining (a zwłaszcza opisane powyżej drzewa wzmacniane) są godne polecenia, jeżeli pragniemy relatywnie szybko zbudować model dobrej jakości i nie zależy nam na jego czytelności oraz możliwości interpretacji uzyskanych reguł. Jednak musimy pamiętać, iż w przypadku modeli data mining potęguje się problem przeuczenia modeli ze względu na większe możliwości dopasowania się ich do analizowanych danych. Dlatego też szczególną uwagę należy zwrócić na proces walidacji modelu oraz zapewnienia odpowiednio licznego zbioru danych do analizy.



Modele statystyczne są idealne, gdy pragniemy, by model był prosty w interpretacji oraz zrozumiały dla szerokiej grupy osób nieposiadających zaawansowanej wiedzy z dziedziny statystyki i data mining. Są także zalecane w przypadku, gdy chcemy rozumieć przyczyny, dla których analizowany przypadek otrzymał niską lub wysoką ocenę, bądź objaśnić odpowiedzi modelu innym osobom w prosty i przejrzysty sposób. Proces przygotowania karty skoringowej oraz jej postać nie są czarną skrzynką i są dobrze opisane i wystandaryzowane. Karta skoringowa jest prosta w ocenie i monitorowaniu. Jej prostota sprawia, że można ją wykorzystywać bez zaawansowanej wiedzy statystycznej czy informatycznej.

Jaki więc model wybrać? Decyzja nigdy nie jest oczywista. Przed dokonaniem wyboru zawsze musimy odpowiedzieć sobie na pytanie, jakie właściwości modelu są dla nas najważniejsze. Warto również próbować rozwiązać dany problem za pomocą kilku konkurencyjnych metod, a następnie wybrać rozwiązanie, które nas w największym stopniu satysfakcjonuje.

Literatura

1. Siddiqui Naeem, *Credit Risk Scorecards Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons, Inc. 2006.
2. Stanisław Andrzej, *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 2. Modele liniowe i nieliniowe*, StatSoft Polska Sp. z o.o. Kraków 2007.