StatSoft Polska

Zestaw skoringowy 6.0

instrukcja instalacji oraz podstawowe informacje o systemie

Spis Treści

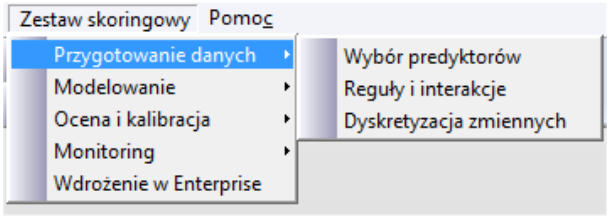
1. Instrukcja instalacji i dezinstalacji programu

1.1. Instalacja wersji jednostanowiskowej i sieciowej

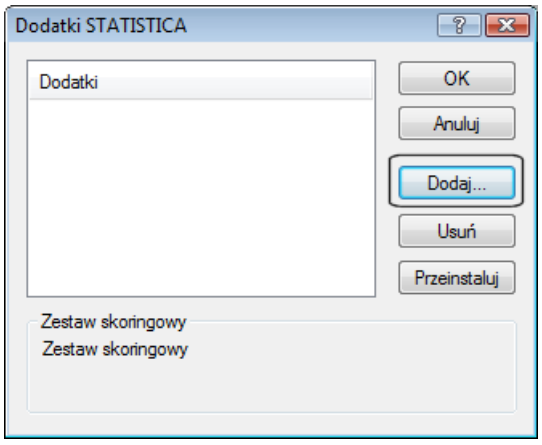
W celu zainstalowania programu należy:

a.) Zainstalować odpowiednią wersję *STATISTICA* (wymagana jest wersja 13.x z odpowiednim zestawem modułów analitycznych).
b.) Uruchomić instalator programu *Zestawy Analityczne.exe*, a następnie zatwierdzać kolejne kroki instalacji.
c.) W trakcie instalacji wskazać plik – *License.xml* – zostanie on skopiowany do katalogu z programem (domyślnie c:\Program Files\StatSoft\Zestawy Analityczne).

Po zainstalowaniu dodatku utworzone zostanie menu umożliwiające dostęp do modułów *Zestawu skoringowego*.



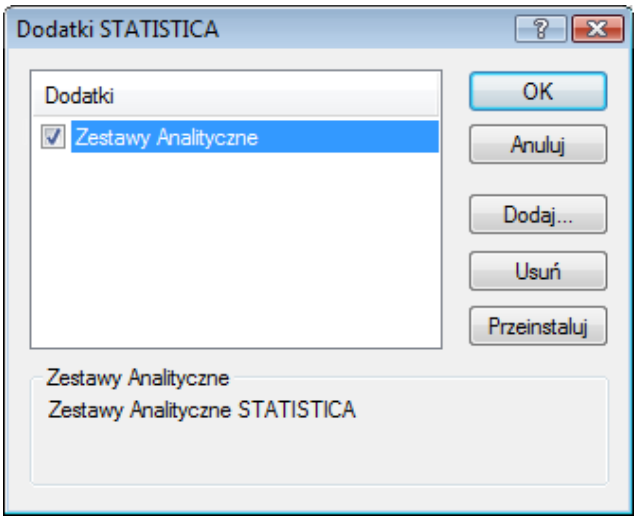
W przypadku niepojawienia się menu ***Zestaw Skoringowy*** po instalacji, należy w menu ***Narzędzia*** | ***Makro*** | ***Dodatki*** otworzyć okno ***Dodatki STATISTICA***.



W wyświetlonym oknie kliknąć przycisk ***Dodaj…*** a następnie wprowadzić napis `ZestawyAnalityczne`.

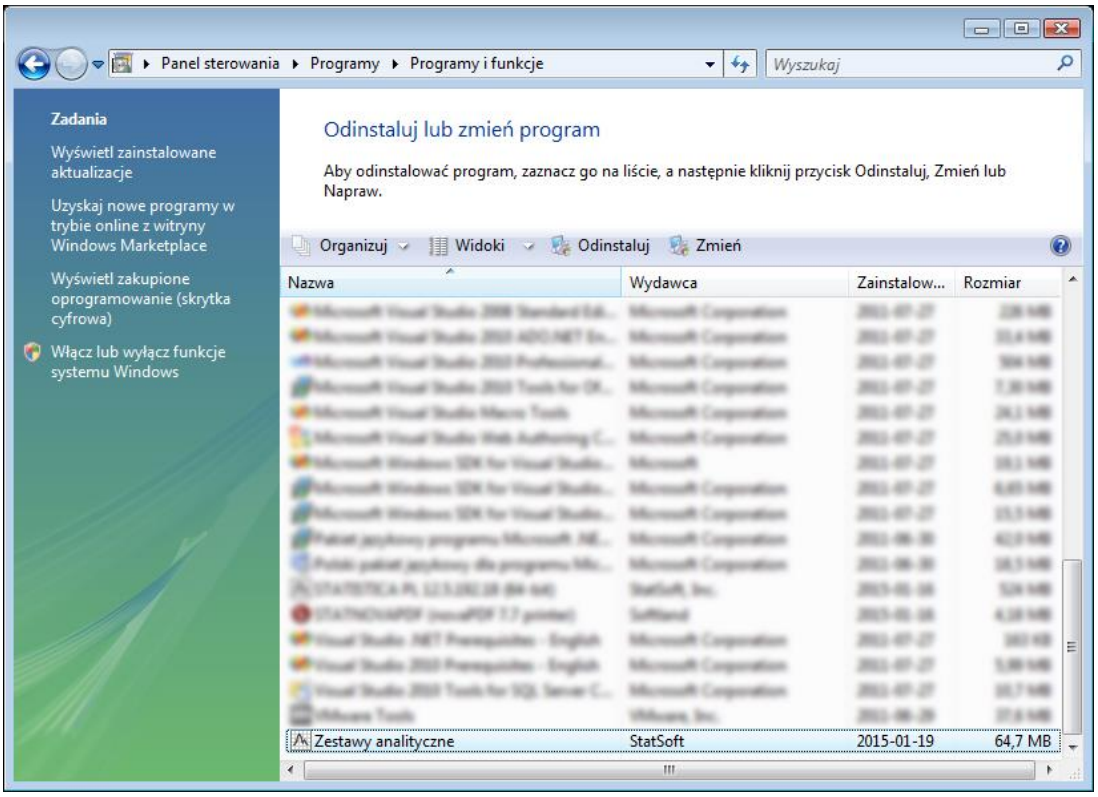
1.2. Odinstalowanie

W celu odinstalowania programu należy z menu ***Narzędzia*** wybrać ***Makro*** | ***Dodatki***



Pojawi się okno ***Dodatki STATISTICA***, należy zaznaczyć ***Zestawy analityczne***, a następnie kliknąć przycisk ***Usuń***.

Następnie należy wejść do ***Panelu sterowania*** systemu Windows, a następnie wybrać grupę ***Programy*** -> ***Programy i funkcje***.



W wyświetlonym oknie należy odnaleźć program ***Zestawy analityczne*** i kliknąć na niego dwukrotnie myszą lub wybrać opcję ***Odinstaluj*** w celu rozpoczęcia procesu dezinstalacji.

W przypadku trudności z odinstalowaniem programu ***Zestawy analityczne*** zalecane jest skorzystanie z narzędzia Microsoft Fix it dla problemów z instalacją i usuwaniem programów dostępnego na stronie firmy Microsoft.

2. Ogólne założenia programu

2.1. Przegląd modułów programu

Zestaw skoringowy jest programem pozwalającym w intuicyjny i wygodny sposób przeprowadzić proces budowy, oceny oraz pielęgnacji karty skoringowej. Program składa się z jedenastu podstawowych oraz dwóch dodatkowych modułów podzielonych na grupy odpowiadające kolejnym etapom analizy:

- **Przygotowanie danych**
  - Wybór predyktorów
  - Reguły i interakcje
  - Dyskretyzacja zmiennych
- **Modelowanie**
  - Budowa tablicy skoringowej
  - Analiza przeżycia (SURVIVAL)
  - Wnioski odrzucone (Reject inference)
- **Ocena i kalibracja**
  - Ocena modeli
  - Zarządzanie punktem odcięcia
  - Wyliczanie skoringu
  - Testy kalibracji
- **Monitoring**
  - Stabilność populacji
  - Analiza Vintage
  - Macierze migracji

➢ Przygotowanie danych

▪ ***Wybór predyktorów***. Moduł do wyboru predyktorów pozwala wykonać ranking ważności predyktorów na podstawie miar *IV* (*Information Value*), *Gini* oraz *V Cramera*, a następnie ograniczyć zbiór danych jedynie do zmiennych istotnie powiązanych ze zmienną zależną. Moduł pozwala również na wybranie reprezentantów zmiennych spośród zbioru skorelowanych ze sobą cech ilościowych. Wyboru reprezentantów dokonujemy na podstawie wyników analizy czynnikowej z rotacją czynników, analizując macierz ładunków czynnikowych.

▪ Moduł ***Reguły i interakcje*** umożliwia wyszukanie zestawu reguł pozwalających na identyfikację podgrup o wysokim prawdopodobieństwie przynależności do jednej z modelowanych klas. Proces identyfikacji reguł odbywa się za pomocą metody *Losowy Las* (*Random Forest*). Jakość utworzonych reguł możemy ocenić za pomocą przyrostu (*lift*) dla obydwóch klas oraz liczności i odsetka negatywnych elementów w klasie. Wybrane reguły możemy przedstawić w postaci drzewa decyzyjnego oraz zapisać do raportu. Na podstawie wybranych reguł mamy możliwość przygotowania dwustanowych zmiennych pochodnych. Dodatkowo moduł pozwala na utworzenie rankingu interakcji pomiędzy parami zmiennych w oparciu o model regresji logistycznej. Moduł dla każdej możliwej pary predyktorów buduje model logistyczny zawierający parę zmiennych oraz model zawierający tę samą parę zmiennych i ich interakcję. Użytkownik ma możliwość oceny siły interakcji za pomocą testu LR.

■ Moduł ***Dyskretyzacja zmiennych*** umożliwia określenie optymalnego sposobu podziału wartości każdego z predyktorów na przedziały. Za pomocą metod automatycznych opartych na drzewach CHAID lub C&RT oraz metodach ręcznych (liczba percentyli lub minimalna liczność) możemy podzielić zmienne na klasy reprezentujące jednorodne klasy ryzyka. Jakość podziału oceniana jest za pomocą miar:

- *WoE* (*Weight of Evidence*), informującej o sile predykcyjnej poszczególnych poziomów danej zmiennej oraz
- *IV* (*Information Value*), informującej o sile predykcyjnej całej zmiennej.

Dodatkowo bardzo ważnym elementem oceny jakości podziału jest ocena wykresu *WoE* (sprawdzenie czy zmiana wartości *WoE* układa się w logiczny trend).

Wszystkie początkowe granice podziału można regulować w sposób ręczny, zmieniając ich granice lub łącząc klasy o podobnym poziomie ryzyka. Po wykonaniu przekształcenia, przepis kategoryzacji zapisujemy w pliku XML lub pliku Reguł (SRX). Plik ten wykorzystany zostanie w module do budowy tablicy skoringowej.

➢ Modelowanie

■ W module ***Budowa tablicy skoringowej*** zmienne wybrane do budowy modelu są przekształcane za pomocą skryptów XML utworzonych w module ***Dyskretyzacja zmiennych***, a następnie na podstawie przekształconych danych budowany jest model regresji logistycznej. Podczas budowy modelu logistycznego mamy do wyboru różne strategie doboru cech. Domyślnie uwzględniane są wszystkie efekty. Istnieje również możliwość wyboru bardziej zaawansowanych strategii doboru cech, począwszy od *wprowadzania postępującego* i *eliminacji wstecznej*, aż po metody typu *„step-wise”*, czyli *metodę krokową postępującą* i *krokową wsteczną*. Kolejną opcją jest budowa modelu z wykorzystaniem strategii *bootstrap*. Po zbudowaniu modelu możemy ocenić jakość jego dopasowania (między innymi za pomocą miar *AIC Akaike Information Criterion* oraz *BIC – Bayesian Information Criterion*), zbadać poziom korelacji i kowariancji parametrów regresji oraz wyświetlić wartości ocen parametrów regresji. W celu utworzenia tablicy skoringowej z modelu logistycznego, należy w kolejnym kroku podać parametry skali, na podstawie których konstruowana jest tablica. Dodatkowo możliwe jest wprowadzenie korekty parametrów regresji (wyrazu wolnego) uwzględniającej fakt budowy modelu na próbie niezbalansowanej. Zbudowaną tablicę można następnie zapisać w dowolnej postaci (zaimplementowane formaty to XML oraz *STATISTICA* Visual Basic, Excel oraz Reguły (SRX) jednak możliwe jest przygotowanie mechanizmu zapisującego tablicę skoringową do formatu wskazanego przez użytkownika).

■ Moduł ***Analiza przeżycia (SURVIVAL)*** umożliwia budowę modeli skoringowych za pomocą proporcjonalnego hazardu Coxa, który poza czynnikami wpływającymi na zmienną zależną pozwala uwzględnić również czas zajścia analizowanego zdarzenia. Modele te pozwalają nie tylko określić prawdopodobieństwo zajścia danego zdarzenia, ale również czas w jakim osiągnie ono dany (krytyczny) poziom - np. kiedy dana osoba przestanie spłacać kredyt bądź kiedy odejdzie do konkurencji.

■ Moduł ***Wnioski odrzucone*** pozwala na uwzględnienie w zbiorze danych będącym podstawą modelowania także tych klientów, którzy zostali odrzuceni na etapie weryfikacji wniosków. Do wyboru są dwie metody przypisania wniosków odrzuconych do odpowiedniej klasy klientów:

- paczkowanie (*parceling*),
- k - najbliższych sąsiadów.

➢ Ocena i kalibracja

■ Moduł *Ocena modeli* umożliwia ocenę i porównanie zbudowanych modeli skoringowych. W module zawarto szereg miar pozwalających ocenić jakość modelu. Należą do nich:

- Wskaźnik *IV* (*Information Value*)
- *KS* – współczynnik Kołmogorowa-Smirnowa
- Wskaźnik *Gini*
- Dywergencja
- Wskaźnik Hosmera-Lemeshowa
- *AUC* - pole powierzchni pod krzywą *ROC*
- Wykres przyrostu *(Lift)*
- Wykres zysku (*Gains*, krzywa *CAP*)

Dodatkowo dla każdego wskaźnika tworzony jest szczegółowy raport, a dla wskaźnika Kołmogorowa-Smirnowa, *Gini*, *ROC*, przyrostu oraz zysku generowane są także odpowiednie wykresy.

Moduł ten dodatkowo daje możliwość:

- Generowania raportu punktacji końcowej (*Final score)* wraz z wykresami szans (*Odds)* oraz zdarzeń niepożądanych *Bad rate,*
- Generowania raportu cech (*Characteristic Report).*

■ Moduł ***Obliczanie skoringu*** pozwala wyznaczyć wartości skoringu wraz ze skoringami cząstkowymi dla każdej z cech. Istnieje także możliwość obliczenia wartości prawdopodobieństwa wystąpienia zdarzenia niepożądanego (np. w skoringu kredytowym prawdopodobieństwa niespłacenia kredytu - *default probability*). Obliczoną wartość prawdopodobieństwa można dodatkowo skorygować o wartość prawdopodobieństwa *a priori* (podawaną przez użytkownika) wystąpienia zdarzenia niepożądanego w badanej populacji. Moduł ten umożliwia także obliczenie prawdopodobieństwa wystąpienia modelowanego zdarzenia w określonym punkcie czasu na podstawie modelu zrównoważonego hazardu Coxa.

■ Moduł ***Zarządzanie punktem odcięcia*** pozwala na wskazanie od 1 do 3 punktów odcięcia. Po wyborze punktu (punktów) odcięcia generowane są odpowiednie raporty oceniające jakość podziału analizowanego zbioru. Optymalny punkt odcięcia może być dodatkowo wyznaczony przy pomocy krzywej *ROC*, przy uwzględnieniu podanych przez użytkownika kosztów błędnych klasyfikacji oraz rzeczywistej frakcji przypadków niepożądanych. Na zakładce ***Zysk*** po podaniu odpowiednich parametrów dotyczących kwotowych kosztów i zysków wynikających z działania modelu np. ile procent zarabiamy na każdym spłaconym kredycie oraz ile procent każdego niespłaconego kredytu nie udaje się nam odzyskać, możliwe jest wyznaczenie optymalnego punktu odcięcia oraz dokonania symulacji zyskowności wynikającej ze stosowania modelu. Dodatkowo możliwe jest przeprowadzenie symulacji zyskowności modelu dla skoringu marketingowego, skoringu *anty-churn* oraz *anty-fraud*. Zaś w bardziej skomplikowanych scenariuszach, gdzie zyski i koszty zależą od dodatkowych zmiennych np. grupujących, można zdefiniować własną *macierz zysków i strat*. Przykładowo, szacowana strata związana z niespłaconym kredytem może zależeć od jego celu lub zabezpieczania posiadanego przez wnioskującego.
W sytuacjach, gdy dysponujemy dwoma ocenami wniosków (np. z wewnętrznego modelu oraz z BIK), możemy przeprowadzić *ocenę macierzową* punktu odcięcia. Raporty generowane wtedy przez program ilustrują, jak zmienia się ryzyko kredytowe w zależności od obydwu ocen jednocześnie. W konsekwencji możliwe jest sformułowanie bardziej złożonych reguł decyzyjnych i skuteczniejsza identyfikację segmentów niskiego i podwyższonego ryzyka.

- Moduł ***Testy kalibracji*** umożliwia przetestowanie zgodności realizacji ryzyka w poszczególnych grupach ratingowych ze zdefiniowaną skalą wzorcową (masterskalą). W zależności od liczności kredytów w poszczególnych grupach ratingowych przeprowadzony jest test dwumianowy, bądź test normalny – wyboru testu można dokonać ręcznie bądź skorzystać z zaimplementowanych wytycznych. W celu ułatwienia interpretacji uzyskanych wyników wprowadzono strategię *traffic light approach*.

➢ Monitoring

- Moduł ***Stabilność populacji*** umożliwia porównanie dwóch zbiorów danych (np. zbioru aktualnych i historycznych kredytobiorców) pod kątem różnic w strukturze wartości poszczególnych cech oraz strukturze samego skoringu. Na podstawie wskaźników oraz wykresów użytkownik może ocenić poziom odchylenia bieżącego zbioru od zbioru bazowego. Znaczące zakłócenia w bieżącym zestawie danych mogą być sygnałem do ponownego oszacowania parametrów modelu.
- Moduł ***Analiza Vintage*** (wykorzystywana przede wszystkim w ryzyku kredytowym) jest dodatkowym modułem Zestawu Skoringowego (wymagającym dodatkowego wdrożenia). Pozwala na monitorowanie stanu portfela kredytów w kolejnych miesiącach spłaty. Umożliwia przygotowanie raportu w zależności od celu kredytów, ich statusu, liczby dni przeterminowania oraz wieku kredytobiorców. Raporty w postaci tabelarycznej są uzupełnione zestawem wykresów pozwalających na łatwiejszy monitoring portfela kredytów i interpretację zachodzących zmian.
- Moduł ***Macierze migracji*** jest dodatkowym modułem Zestawu Skoringowego (wymagającym dodatkowego wdrożenia). Pozwala na obliczenie raportów opisujących strukturę portfela oraz macierze migracji dla wskazanego punktu startowego. Raporty w postaci tabelarycznej są uzupełnione zestawem wykresów przedstawiających zmiany przeterminowania w zależności od miesiąca obserwacji oraz portfela.

2.2. Pliki przykładowe

Wszystkie przykłady przedstawione w tej dokumentacji bazują na zestawie plików znajdujących się w katalogu *Skoring kredytowy*.

- *Skoring kredytowy*
  - *Zbiory danych*
  - *Skrypty dyskretyzacji*
  - *Modele*
  - *Testy kalibracji*

W katalogu *Zbiory danych* znajdują się przykładowe arkusze *STATISTICA* używane podczas przykładów.

W katalogu *Skrypty dyskretyzacji* znajdują się przygotowane skrypty dyskretyzacji zmiennych gotowe do wykorzystania w module do budowy kart skoringowych.

W katalogu *Modele* znajdują się wszystkie modele konieczne do wykonania ćwiczeń zapisane w formie skryptów *XML* interpretowanych przez *Zestaw skoringowy.*

W katalogu *Testy kalibracji* znajdują się pliki pozwalające zademonstrować możliwości modułu ***Testy kalibracji***.

Dodatkowo niniejszą dokumentację uzupełniają artykuły opisujące wykorzystanie skoringu marketingowego oraz medycznego znajdujące się odpowiednio w katalogu *skoring marketingowy* oraz *skoring medyczny*.

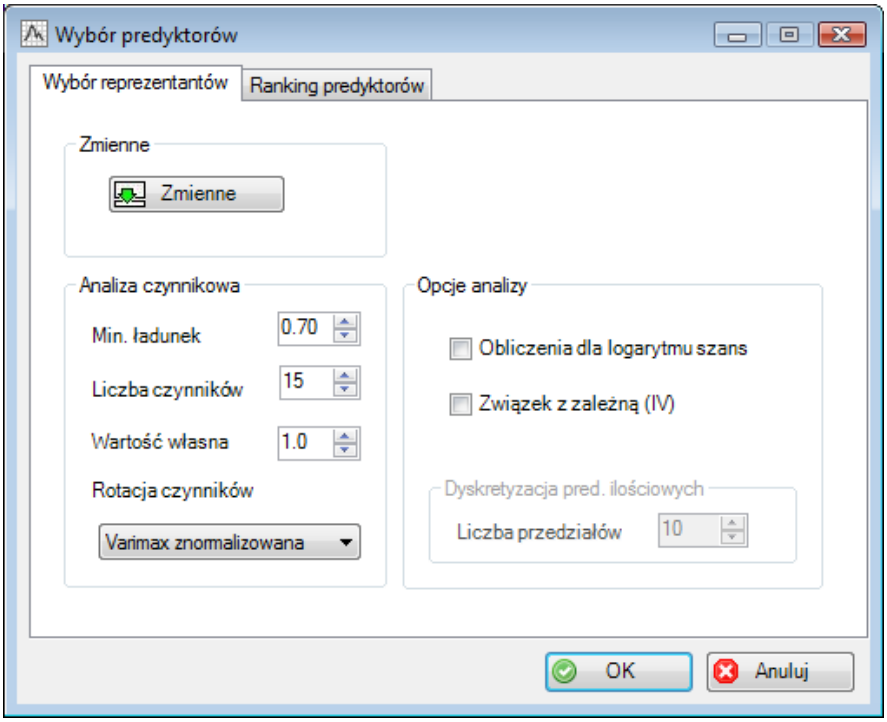
3. Przygotowanie danych

3.1. Wybór predyktorów

Bardzo często wejściowy zbiór danych zawiera nieistotne z punktu widzenia analizy zmienne, które dodatkowo są wzajemnie mocno skorelowane, powielając tym samym zawarte w nich informacje. Usunięcie ich ze zbioru danych takich zmiennych jest więc koniecznością, która nie tylko nie pogarsza jakości zbudowanego modelu, ale wręcz ją poprawia, wpływając dodatkowo na poprawę jego stabilności.

3.1.1. Wybór reprezentantów

Opcja ***Wybór reprezentantów*** modułu ***Wybór predyktorów*** pozwala na identyfikację nadmiarowych (w sensie korelacji) zmiennych bez konieczności tworzenia i analizowania macierzy korelacji dla wszystkich zmiennych. Moduł ten tworzy wiązki skorelowanych ze sobą zmiennych na podstawie analizy czynnikowej (metodą ekstrakcji czynników jest metoda głównych składowych) z rotacją czynników, który jest realizowany za pomocą standardowej procedury *STATISTICA*. Wiązki zmiennych są tworzone w oparciu o wartości ładunków czynnikowych (korelacji pomiędzy daną zmienną i konkretnym czynnikiem). Użytkownik może określić minimalną wartość bezwzględną ładunku powodującą, że dana zmienna będzie traktowana jako potencjalna reprezentanta danego czynnika. Liczba składowych jest określana na podstawie opcji ***Liczba czynników*** oraz ***Wartość własna***.
Jeżeli wybrano do analizy również predyktory jakościowe, przed wykonaniem analizy czynnikowej zmienne są przekodowywane za pomocą transformacji WoE (opartej na logarytmie szansy modelowanego zjawiska), w takiej sytuacji konieczne jest zatem wskazanie dwuwartościowej zmiennej zależnej.

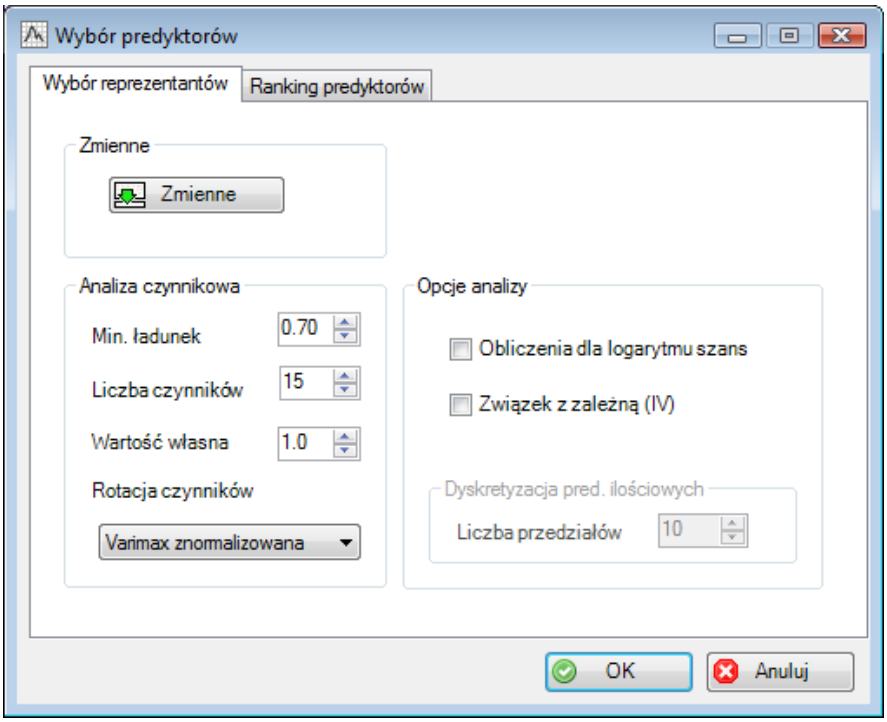


Po wykonanej analizie, zmienne znajdujące się w tej samej wiązce mocno (powyżej wartości określonej w opcji ***Min. ładunek***) korelują z przypisanym tej wiązce czynnikiem i co istotne bardzo często są mocno skorelowane także między sobą. Spośród tych zmiennych, na podstawie oceny ich

wzajemnej korelacji, stopnia powiązania ze zmienną zależną oraz wiedzy eksperckiej wybierane są zmienne do dalszej analizy. Możliwa jest także opcja automatycznego wyboru reprezentantów z wiązek na podstawie korelacji – wybieramy zmienne, które mają najwyższą średnią korelację z pozostałymi zmiennymi w wiązce. Drugą opcją automatycznego wyboru jest kryterium *IV*, wybieramy takich reprezentantów wiązki, którzy mają najmocniejsze powiązanie ze zmienną zależną.

Przykład 1. –Wybór reprezentantów

Aby zobrazować proces wyboru reprezentantów ze zbioru silnie skorelowanych zmiennych wykonamy analizę na podstawie zbioru danych *SelectRepresentatives.sta* znajdującego się w katalogu *Zbiory danych*. Po otwarciu zbioru danych z menu ***Zestaw skoringowy*** wybieramy polecenie ***Przygotowanie danych*** | ***Wybór predyktorów*** wyświetlając okno dialogowe o tej samej nazwie. W oknie tym wybieramy kartę ***Wybór reprezentantów***, aby uzyskać dostęp do następujących opcji analizy:



Min. ładunek określa kryterium przypisywania poszczególnych zmiennych do wiązek reprezentowanych przez uzyskane czynniki. Wartość oznacza minimalną wartość bezwzględną ładunku (korelacji pomiędzy zmienną a czynnikiem), jaka kwalifikuje zmienną do wiązki.

Liczba czynników określa, jaka liczba czynników zostanie wyodrębniona. Opcja ta jest powiązana z opcją ***Wartość własna***. Program wyodrębni taką liczbę czynników, jaka wynika z obydwóch kryteriów łącznie – liczba czynników będzie nie większa niż wartość parametru ***Liczba czynników***, natomiast wartość własna każdego z wyodrębnionych czynników będzie nie mniejsza niż wartość parametru ***Wartość własna***.

Rotacja czynników umożliwia wybór odpowiedniej opcji rotacji – szczegóły patrz: Strategie rotacji czynników

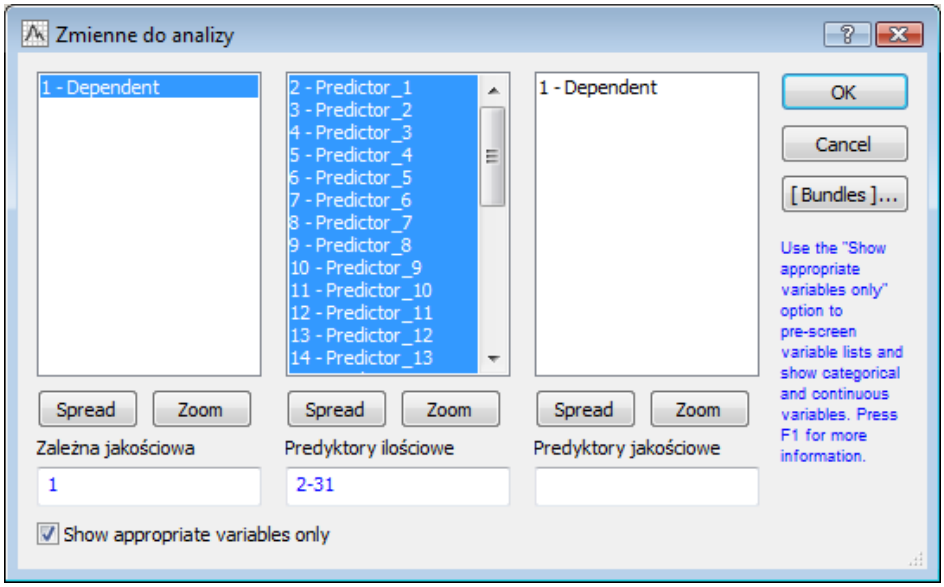
Obliczenia dla logarytmu szans transformuje predyktory na wartości WoE (szczegóły obliczeniowe patrz: Scorecard Formula Guide). Dzięki tej transformacji w analizie można wykorzystać zmienne jakościowe. Zmienne ilościowe przed transformacją zostaną podzielone na równoliczne kategorie.

Liczba kategorii określona jest parametrem ***Liczba przedziałów***. Zwróćmy uwagę, że transformacja ta jest możliwa jedynie w sytuacji, gdy w analizie uwzględniono dychotomiczną zmienną zależną.

Związek z zależną (IV) umożliwia obliczenie miary IV (szczegóły obliczeniowe patrz: Scorecard Formula Guide) oceniającej związek predykorów ze zmienną zależną. Przed wykonaniem obliczeń zmienne ilościowe zostaną podzielone na równoliczne kategorie. Liczba kategorii określona jest parametrem ***Liczba przedziałów***. Zwróćmy uwagę, że obliczenie wskaźnika IV jest możliwe jedynie w sytuacji, gdy w analizie uwzględniono dychotomiczną zmienną zależną.

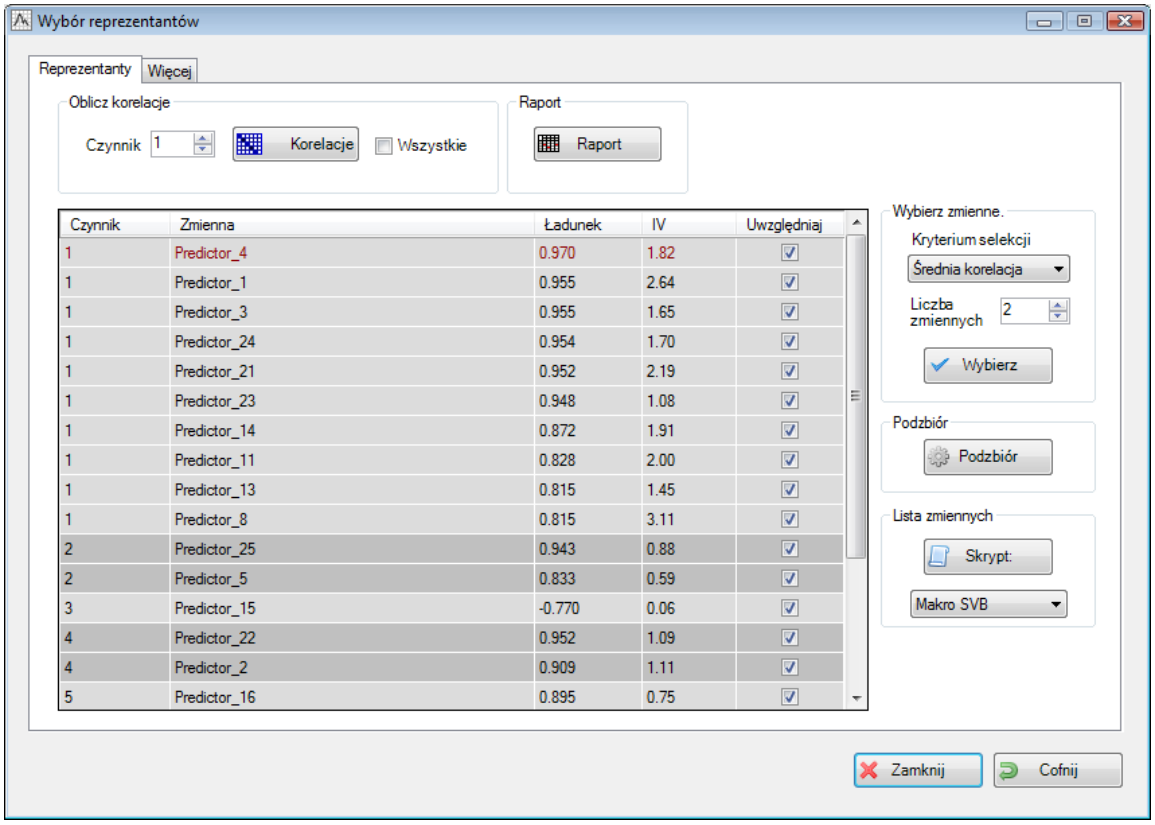
Liczba przedziałów określa liczbę równolicznych kategorii, na jakie podzielone zostaną predyktory ilościowe przed transformacją do logarytmu szans lub obliczeniem IV.

Aby wybrać zmienne do analizy, klikamy przycisk ***Zmienne***.



Wybieramy zmienną *Dependent* jako zmienną ***Zależną jakościową*** oraz zmienne od 2-31 jako ***Predyktory ilościowe***. Zaznaczamy opcję ***Związek z zależną (IV)***, pozostałe opcje pozostawiamy na domyślnych poziomach.

Klikamy ***OK*** aby wykonać obliczenia i przejść do okna ***Wybór reprezentantów***.



Czynnik	Zmienna	Ładunek	IV	Uwzględniaj
1	Predictor_4	0.970	1.82	☑
1	Predictor_1	0.955	2.64	☑
1	Predictor_3	0.955	1.65	☑
1	Predictor_24	0.954	1.70	☑
1	Predictor_21	0.952	2.19	☑
1	Predictor_23	0.948	1.08	☑
1	Predictor_14	0.872	1.91	☑
1	Predictor_11	0.828	2.00	☑
1	Predictor_13	0.815	1.45	☑
1	Predictor_8	0.815	3.11	☑
2	Predictor_25	0.943	0.88	☑
2	Predictor_5	0.833	0.59	☑
3	Predictor_15	-0.770	0.06	☑
4	Predictor_22	0.952	1.09	☑
4	Predictor_2	0.909	1.11	☑
5	Predictor_16	0.895	0.75	☑

W wyświetlonym oknie kolumna ***Czynnik*** informuje o numerze wiązki, do której trafiła dana ***Zmienna***. Kolumna ***Ładunek*** zawiera wartość korelacji pomiędzy zmienną a danym czynnikiem, ***IV*** wyświetla siłę predykcyjną danego predyktora. Kolumna ***Uwzględniaj*** pozwala użytkownikowi na wybór/usuwanie zmiennych z wejściowego zbioru danych.

Opcje ***Oblicz korelacje*** umożliwiają użytkownikowi wyświetlenie macierzy korelacji dla zmiennych, jakie znalazły się w poszczególnych wiązkach (czynnikach).

> ***Czynnik*** określa numer czynnika, dla którego ma zostać wyświetlona macierz korelacji.
>
> ***Wszystkie*** powoduje wyświetlenie osobnych macierzy korelacji dla wszystkich czynników.

Raport wyświetla arkusz zawierający informacje przedstawiane w tabeli.

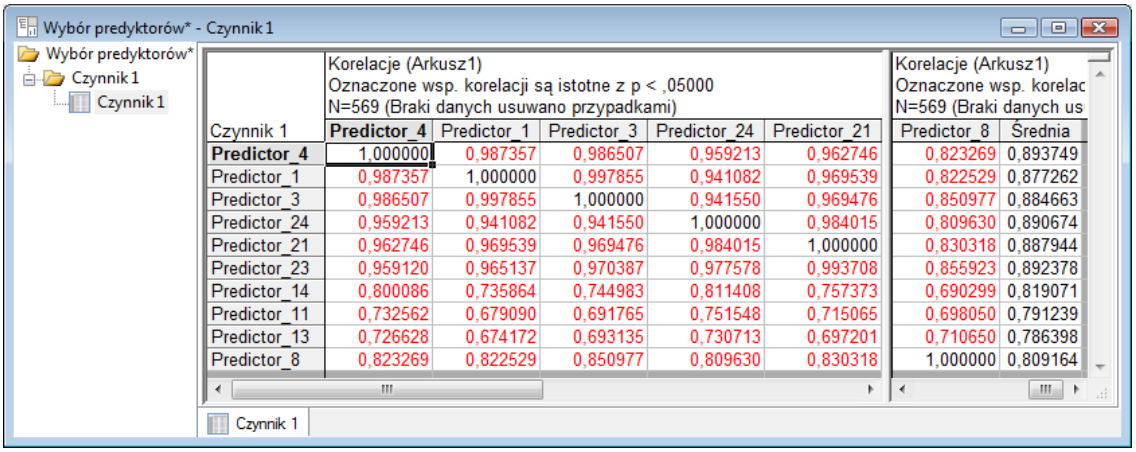
Wybierz zmienne pozwala użytkownikowi na automatyczny wybór reprezentantów spośród zmiennych zawartych w wyodrębnionych wiązkach.

> ***Kryterium selekcji*** pozwala użytkownikowi wybrać metodę automatycznej selekcji reprezentantów. Jeśli wybrana jest opcja *Średnia korelacja* program wybierze zmienne z największą średnią korelacją z innymi zmiennymi znajdującymi się w wiązce. Wybór opcji *Information Value* pozwoli na wybór reprezentantów najmocniej powiązanych ze zmienną zależną.
>
> ***Liczba zmiennych*** określa liczbę zmiennych z każdej wiązki, jaka ma zostać wybrana do analizy.

Przycisk ***Podzbiór*** umożliwia wygenerowanie nowego arkusza danych niezawierającego zmiennych mających odznaczoną opcję ***Uwzględniaj*** w tabeli.

Przycisk ***Skrypt*** umożliwia wygenerowanie makra (opcja *Makro SVB*) dokonującego selekcji zmiennych lub raportu z numerami wybranych zmiennych.

W wyświetlonym oknie upewniamy się, że pole ***Czynnik*** w obszarze ***Oblicz korelacje*** ma wartość 1, a następnie klikamy przycisk ***Korelacje***, aby wyświetlić macierz korelacji dla zmiennych, które znalazły się w pierwszej wiązce. Możemy zauważyć, że znalazły się tam bardzo mocno skorelowane zmienne.



Ostatnia kolumna powyższego arkusza przedstawia średnią korelację danej zmiennej z pozostałymi z wiązki.

W obszarze ***Wybierz zmienne*** na liście wyboru wskazujemy kryterium ***Information Value***, aby wybrać reprezentanty na podstawie tego kryterium. Opcję ***Liczba zmiennych*** pozostawiamy na niezmienionym poziomie (pozostawiamy zatem po dwie zmienne z każdej wiązki – ich dalszą selekcją będziemy mogli zająć się w kolejnych krokach analizy, na przykład na etapie budowy modelu), a następnie klikamy przycisk ***Wybierz***, aby przeprowadzić automatyczną selekcję zmiennych. Za pomocą przycisku ***Podzbiór*** tworzymy nowy zbiór danych z usuniętymi nadmiarowymi zmiennymi.

Opcje zawarte na karcie ***Więcej*** pozwalają ocenić model czynnikowy. Więcej informacji zobacz: Wyniki analizy czynnikowej.

3.1.2. Wybór predyktorów

Druga z opcji modułu ***Wybór predyktorów*** – ***Ranking predyktorów*** umożliwia wykonanie rankingu zmiennych na podstawie miar *Information Value*, *V Cramera* oraz *Gini* a następnie ograniczenie zbioru danych jedynie do zmiennych istotnie wpływających na badane zjawisko.

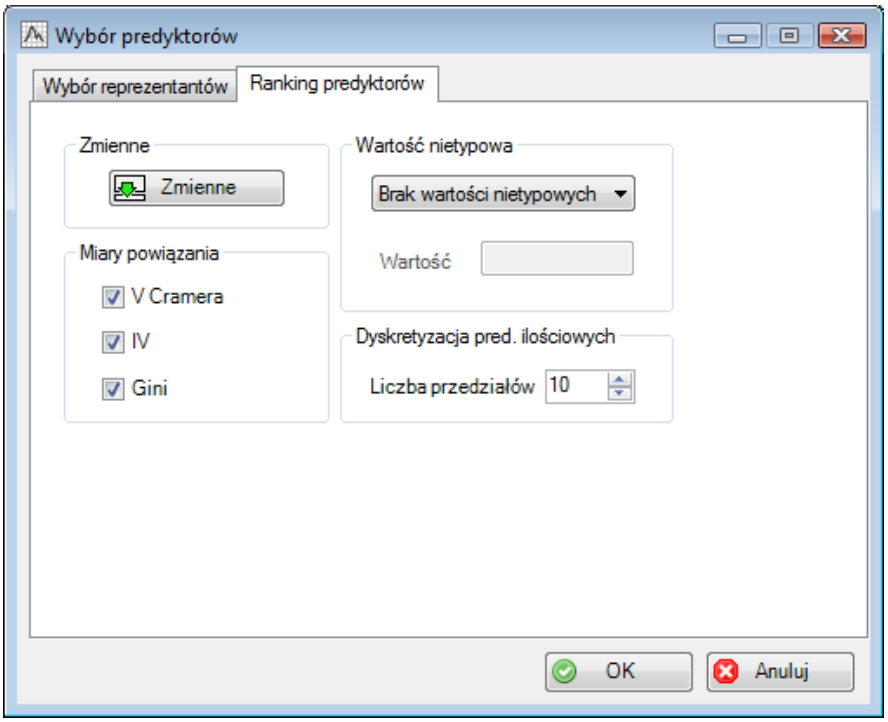
Przykład 2. –Wybór zmiennych do analizy

Podstawą poniższego przykładu będzie plik danych *CreditScoring.sta* znajdujący się w katalogu *Datasets*. Plik ten zawiera 1000 przypadków opisujących historycznych kredytobiorców - każdy przypadek reprezentuje jeden kredyt. Załóżmy, że zmienna *Ocena* jest zmienną informującą o statusie kredytu – czy był on spłacany poprawnie (należy do klasy **dobry**), czy też nie (klasyfikujemy go jako **zły**). Zmienne od 2 do 18 to zmienne informujące o poszczególnych cechach kredytobiorcy, które, jak sądzimy, mają wpływ na jego wiarygodność kredytową. W zbiorze danych występują zarówno zmienne ilościowe (np. *Wiek*, *Kwota*) jak i jakościowe (np. *Cel* czy *Historia*).

Dane: CreditScoring.sta (19 zmn. * 1000 prz.)

	1 Ocena	2 Stan konta	3 Okres	4 Historia	5 Cel	6 Kwota	7 Suma aktywów	8 Zatrudnie
1	zły	brak	36	bez problemów	rehabilitacja	3003	brak	5 do 8 lat
2	dobry	debet	48	opóźnienia	rehabilitacja	17085,6	>1400	1 do 5 lat
3	zły	>$300	36	brak	używany samochód	15363,6	brak	bezrobotny
4	dobry	brak	24	spłacone	nowy samochód	8986,6	brak	Ponad 8
5	dobry	>$300	24	brak	rehabilitacja	1761,2	brak	5 do 8 lat
6	dobry	debet	12	brak	rehabilitacja	1451,8	<140	5 do 8 lat
7	zły	brak	30	brak	używany samochód	4351,2	brak	do roku
8	dobry	debet	15	spłacone	meble	2151,8	>1400	Ponad 8
9	dobry	>$300	15	spłacone	meble	2059,4	brak	1 do 5 lat
10	zły	debet	27	spłacone	meble	3528	140-700	1 do 5 lat
11	zły	debet	24	brak	używany samochód	5679,8	brak	5 do 8 lat
12	zły	brak	18	brak	remont	1050	brak	bezrobotny
13	zły	debet	36	bez problemów	rehabilitacja	6237	brak	1 do 5 lat
14	dobry	>$300	6	bez problemów	rehabilitacja	2440,2	<140	1 do 5 lat
15	dobry	brak	12	brak	inny	2650,2	brak	1 do 5 lat
16	dobry	>$300	42	brak	meble	10032,4	>1400	5 do 8 lat

Po otwarciu pliku *CreditScoring.sta* z menu ***Zestaw skoringowy*** wybieramy polecenie ***Przygotowanie danych*** | ***Wybór predyktorów***, a następnie w wyświetlonym oknie wybieramy kartę ***Ranking predyktorów***.



Okno to udostępnia następujący zestaw parametrów:

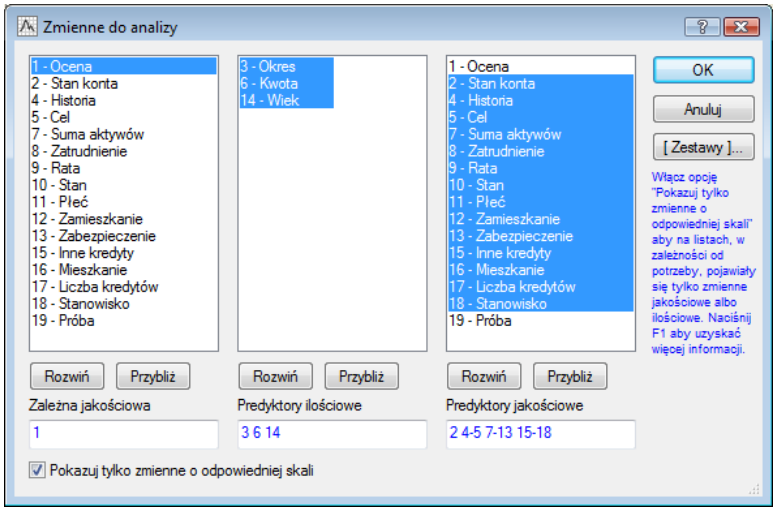
Obszar ***Miary powiązania*** umożliwia wybór trzech miar siły predykcyjnej ***V Cramera***, ***IV*** (*Information Value*) oraz wskaźnika ***Gini***. (szczegóły obliczeniowe patrz: Scorecard Formula Guide).

Obszar ***Wartość nietypowa*** pozwala na uwzględnienie w obliczeniach brakujących danych lub specjalne potraktowanie wartości wskazanej przez użytkownika jako nietypowa. Wybór opcji ***Brak danych*** powoduje, że brakujące dane są traktowane w analizie jako osobna kategoria i tym samym mają wpływ na obliczanie mocy predykcyjnej danej zmiennej. Opcja ***Wskaż wartość*** umożliwia wskazanie w polu ***Wartość*** wartości, która będzie traktowana przez program jako osobna kategoria i przez to specjalnie uwzględniona w obliczeniach.

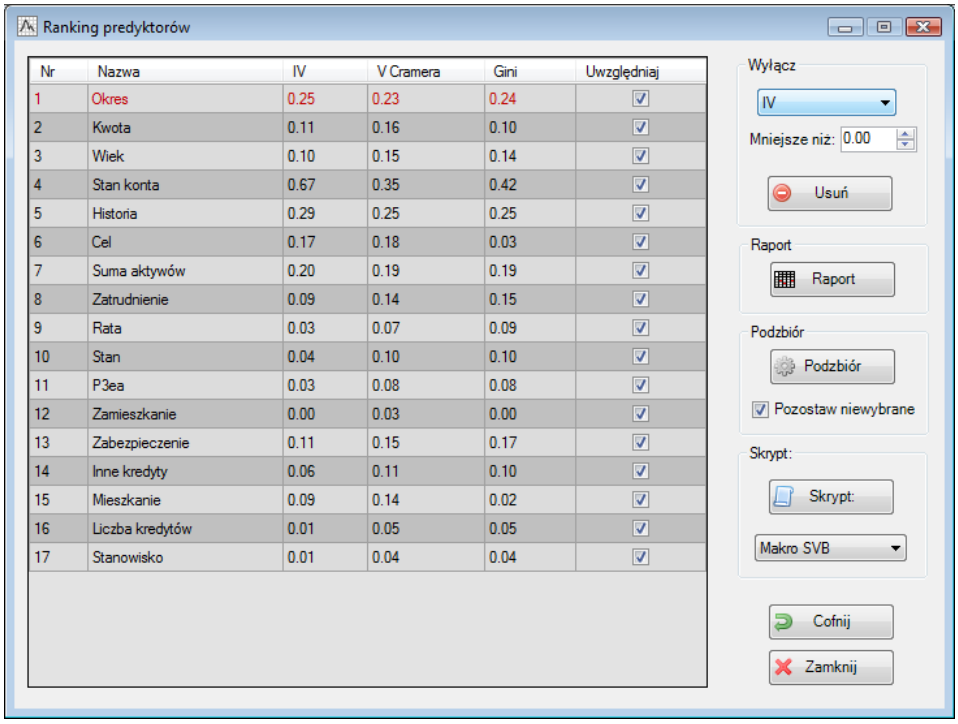
Pole ***Liczba przedziałów*** znajdująca się w obszarze ***Dyskretyzacja pred. ilościowych*** pozwala określić liczbę równolicznych przedziałów, na jakie zostanie podzielona każda zmienna ilościowa przed

obliczeniem miar powiązania. Jeżeli w obszarze ***Wartość nietypowa*** wskazano ***Wskaż wartość*** lub ***Brak danych***, to będą one stanowiły dodatkową kategorię.

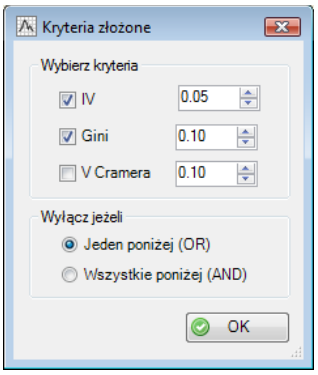
Aby wykonać analizę, klikamy przycisk ***Zmienne***. Zaznaczmy pole ***Pokazuj tylko zmienne o odpowiedniej skali***, wybieramy je zgodnie z poniższym rysunkiem, a następnie zatwierdzamy wybór klikając ***OK***.



Aby obliczyć siłę predykcyjną dla wybranych zmiennych, upewniamy się, że zaznaczone zostały opcje ***V Cramera***, ***IV*** oraz ***Gini***. Pozostałe opcje pozostawiamy bez zmian, a następnie zatwierdzamy wykonanie analizy klikając ***OK***. W rezultacie otrzymujemy okno ***Ranking predyktorów*** z obliczonymi dla wszystkich zmiennych miarami powiązania ze zmienną *Ocena*.

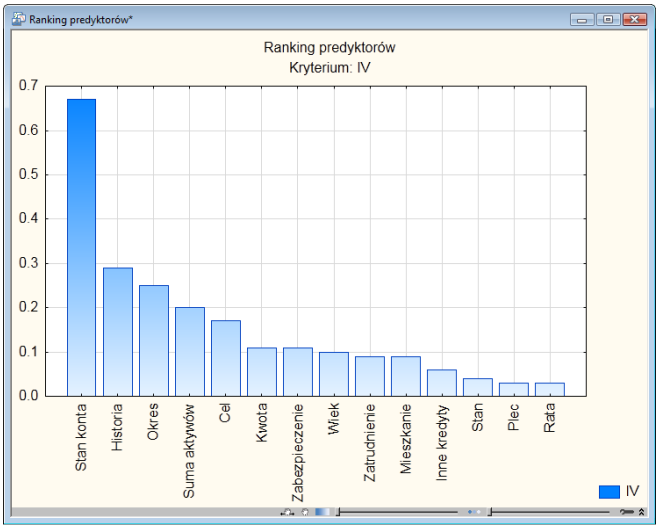


W obszarze ***Wyłącz*** użytkownik ma możliwość włączenia/wyłączenia zmiennych z dalszej analizy. Wybór opcji ***IV***, ***V Cramera*** lub ***Gini*** pozwala na wybór zmiennych na podstawie jednego, wybranego kryterium określonego w polu ***Mniejsze niż***. Wybór opcji ***Złożone*** spowoduje wyświetlenie okna ***Kryteria złożone*** umożliwiając wykonanie selekcji zmiennych na podstawie więcej niż jednego kryterium.



Po wskazaniu kryterium oraz wartości granicznych możemy usunąć zmienne za pomocą przycisku ***Usuń***. Jego naciśnięcie spowoduje odznaczenie zmiennych w tabeli niespełniających kryteriów wyboru.

Przycisk ***Raport*** wyświetla skoroszyt zawierający tabelę oraz wykresy prezentujące siłę predykcyjną wybranych zmiennych.



Możemy zauważyć, że zmienna *Stan konta* ma bardzo dużą siłę predykcyjną, siedem kolejnych zmiennych ma średnią siłę predykcyjną (IV pomiędzy 0,1 a 0,3).

Przycisk ***Podzbiór*** pozwala na wygenerowanie nowego zbioru danych zawierającego jedynie zmienne spełniające kryteria wyboru określone w tabeli (zmienne z zaznaczoną opcją ***Uwzględnij***).

Pozostaw niewybrane powoduje pozostawienie w nowym zbiorze zmiennych, które nie zostały wybrane do rankingu predyktorów (mogą to być na przykład identyfikatory przypadków, identyfikator próby lub inne zmienne niewskazane do analizy).

Przycisk ***Skrypt*** umożliwia wygenerowanie makra (opcja *Makro SVB*) dokonującego selekcji zmiennych lub raportu z numerami wybranych zmiennych.

Na podstawie uzyskanych wyników możemy wskazać graniczną wartość zależności, które traktować będziemy za istotne. Przyjmijmy, że nieistotne są zmienne, dla których wartość *IV* jest mniejsza od 0,02, w obszarze ***Wyłącz*** wybieramy wskaźnik ***IV***, określamy graniczną wartość, a następnie klikamy przycisk ***Usuń***. Spowoduje to anulowanie zaznaczenia pól wyboru znajdujących się w tabeli w kolumnie ***Uwzględniaj***.

W obszarze ***Podzbiór*** upewniamy się, że zaznaczono opcję ***Pozostaw niewybrane*** (chcemy aby w nowym zbiorze danych pozostała zmienna *Próba* niewybrana do analizy), klikamy przycisk ***Podzbiór***, który utworzy arkusz zawierający tylko wybrane zmienne. Alternatywnym rozwiązaniem jest wybranie w obszarze ***Skrypt*** opcji ***Makro***, a następnie kliknięcie przycisku ***Skrypt***, co spowoduje

utworzenie makra wyboru zmiennych eliminującego ze zbioru danych nieistotne predyktory. Uruchomienie wygenerowanego makra na wejściowym zbiorze danych utworzy arkusz wejściowy do dalszych analiz

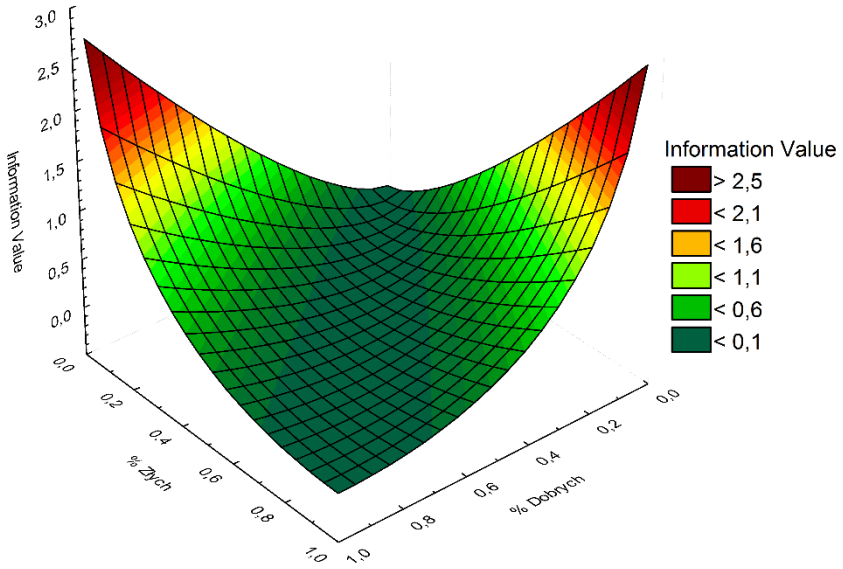
Uwaga. *Information Value jest wskaźnikiem ogólnej mocy predykcyjnej danej zmiennej. Oblicza się go korzystając z następującego wzoru:*

$$IV = \sum_{i=1}^{n} (\text{Distr Good}_i - \text{Distr Bad}_i) * \ln\left(\frac{\text{Distr Good}_i}{\text{Distr Bad}_i}\right)$$

Gdzie n oznacza liczbę atrybutów (poziomów) zmiennej, a Distr Good i Distr Bad odnoszą się do procentowego udziału złych i dobrych kredytów i-tej klasy w stosunku do całości. Korzystając ze wskaźnika IV zmienne można podzielić względem ich mocy predykcyjnej na:

- *Brak zdolności predykcyjnej – mniej niż 0.02*
- *Słaba – 0.02 -0.1*
- *Średnia – 0.1-0.3*
- *Mocna 0.3 i więcej*

W bardzo prostym przypadku, gdy mamy tylko jedną klasę, zależność można przedstawić na wykresie. Widać, że jeżeli liczby złych i dobrych kredytów są sobie równe (przekątna) to IV jest równe 0, i rośnie wraz ze wzrostem dysproporcji w tych dwóch klasach – im różnica jest większa tym większa wartość IV.



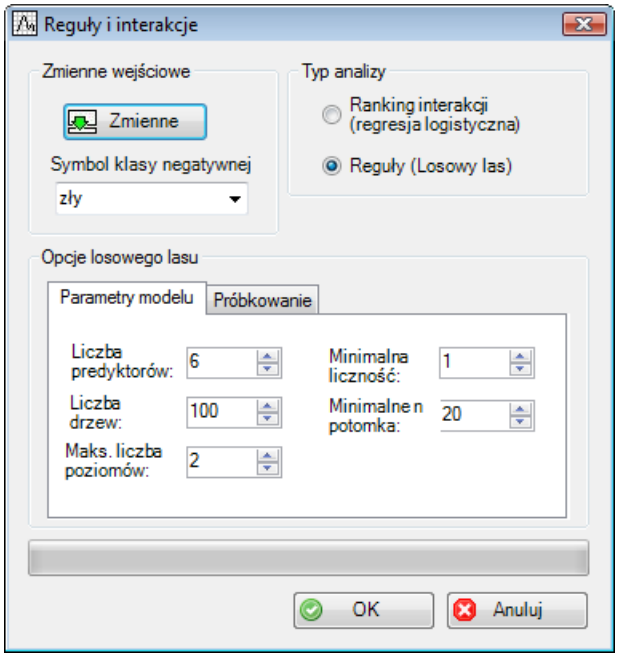
Uwaga*: Niekiedy zdarza się, że ogólna moc predykcyjna danej zmiennej jest na zbyt wysokim poziomie – powyżej 0,5 – aby traktować ją jako normalny predykator. Taka zmienna zdominowałaby model i wpływała niekorzystnie na jego właściwości. Występowanie takiej zmiennej może także świadczyć o braku homogeniczności zbioru danych. Można wtedy wykorzystać tą zmienną do wstępnej segmentacji zbioru danych, a następnie budować osobne modele dla każdego z segmentów. Tak przygotowane modele zwykle będą znacznie lepiej działały niż pojedynczy model budowany dla całego zbioru danych (oczywiście budowa osobnych modeli będzie możliwa jedynie w przypadku zapewnienia odpowiedniej podaży danych, zwłaszcza przypadków zaliczanych do rzadszej grupy).*

3.2. Reguły i interakcje

Moduł ***Reguły i interakcje*** umożliwia wyszukanie zestawu reguł pozwalających na identyfikację podgrup o wysokim prawdopodobieństwie przynależności do jednej z modelowanych klas. Proces identyfikacji reguł odbywa się za pomocą metody *Losowy Las (Random Forest)* -więcej informacji Random Forests Introductory Overview and Technical Notes. Aby uruchomić moduł wymagana jest licencja na moduł *Losowy Las*. Jakość utworzonych reguł możemy ocenić za pomocą przyrostu (*lift*) dla obydwóch klas oraz liczności i odsetka negatywnych elementów w klasie. Wybrane reguły możemy przedstawić w postaci drzewa decyzyjnego oraz zapisać do raportu. Na podstawie wybranych reguł mamy możliwość przygotowania dwustanowych zmiennych pochodnych. Dodatkowo moduł pozwala na utworzenie rankingu interakcji pomiędzy parami zmiennych w oparciu o model regresji logistycznej. Moduł dla każdej możliwej pary predyktorów buduje model logistyczny zawierający parę zmiennych oraz model zawierający tę samą parę zmiennych i ich interakcję. Użytkownik ma możliwość oceny siły interakcji za pomocą testu LR (więcej informacji Logistic regression formula guide).

Przykład 3.– Identyfikacja reguł za pomocą Losowego Lasu

Przykład identyfikacji interakcji zaprezentujemy opierając się na pliku *CreditScoring.sta* wykorzystywanym w poprzednim przykładzie. Z menu ***Zestaw skoringowy*** wybieramy polecenie ***Przygotowanie danych*** a następnie ***Reguły i interakcje*** otwierając okno o tej samej nazwie. Wybieramy zmienne w sposób analogiczny do wcześniejszego przykładu a następnie wskazujemy ***Symbol klasy negatywnej*** (w naszym przypadku jest to klasa „*zły*”).



W obszarze ***Typ analizy*** wybieramy opcję ***Reguły (Losowy las)***. Za pomocą parametrów znajdujących się w obszarze ***Opcje losowego lasu*** sterujemy maksymalną głębokością drzewa, czy też parametrami określającymi warunki podziału kolejnych węzłów drzewa.

Na karcie ***Parametry modelu*** dostępne są następujące opcje:

Liczba predyktorów określa, jaka liczba predyktorów będzie losowana podczas budowy każdego z drzew składowych – będzie to podzbiór zmiennych wskazanych przy wyborze zmiennych.

Liczba drzew – określa liczbę drzew składowych, jaka będzie budowana podczas tworzenia losowego lasu.

Maks. liczba poziomów jest parametrem zatrzymania budowy drzewa. Dopuszczalne są drzewa zawierające 2 lub 3 poziomy (nie licząc korzenia drzewa).

Minimalna liczność określa minimalną liczność węzła macierzystego, jaka może podlegać dalszym podziałom.

Minimalne n potomka określa minimalną liczność węzła potomka, jaki może zostać utworzony w wyniku podziału.

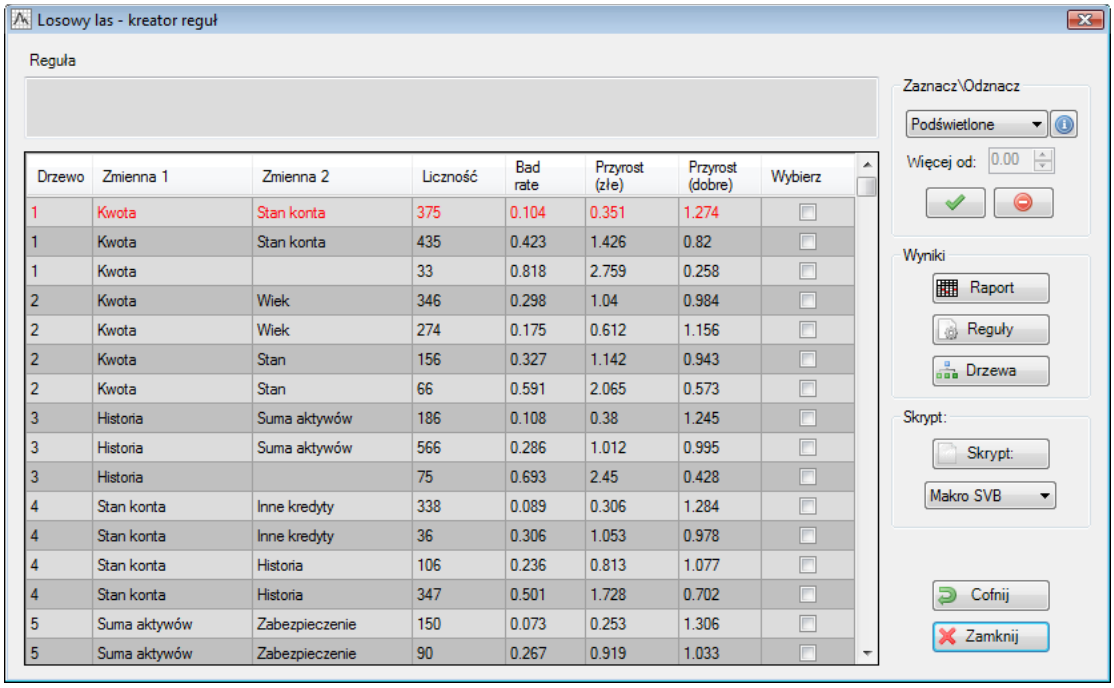
Dodatkowo na karcie ***Próbkowanie*** znaleźć można poniższe opcje:

Proporcja podprób – określa, jaki podzbiór przypadków będzie uwzględniony w próbach bootstrapowych przygotowywanych do budowy kolejnych drzew.

Generator liczb losowych – wartość określa jądro generatora liczb losowych wykorzystywane podczas losowania zmiennych oraz przypadków do budowy kolejnych drzew.

Więcej informacji na temat opcji analizy znajduje się w pomocy elektronicznej do modułu *Losowy Las* programu *STATISTICA*.

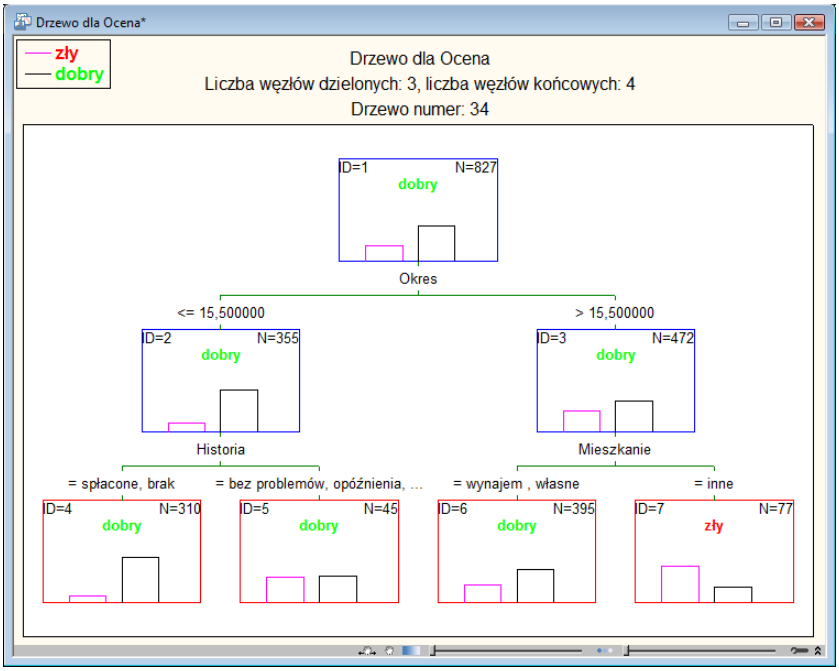
Akceptujemy domyślne ustawienia programu zatwierdzając wykonanie analizy za pomocą przycisku ***OK***. Po wykonaniu analizy wyświetlone zostanie okno ***Losowy las – kreator reguł***, w którym wyświetlone zostały wygenerowane reguły.



Drzewo	Zmienna 1	Zmienna 2	Liczność	Bad rate	Przyrost (złe)	Przyrost (dobre)	Wybierz
1	Kwota	Stan konta	375	0.104	0.351	1.274	
1	Kwota	Stan konta	435	0.423	1.426	0.82	
1	Kwota		33	0.818	2.759	0.258	
2	Kwota	Wiek	346	0.298	1.04	0.984	
2	Kwota	Wiek	274	0.175	0.612	1.156	
2	Kwota	Stan	156	0.327	1.142	0.943	
2	Kwota	Stan	66	0.591	2.065	0.573	
3	Historia	Suma aktywów	186	0.108	0.38	1.245	
3	Historia	Suma aktywów	566	0.286	1.012	0.995	
3	Historia		75	0.693	2.45	0.428	
4	Stan konta	Inne kredyty	338	0.089	0.306	1.284	
4	Stan konta	Inne kredyty	36	0.306	1.053	0.978	
4	Stan konta	Historia	106	0.236	0.813	1.077	
4	Stan konta	Historia	347	0.501	1.728	0.702	
5	Suma aktywów	Zabezpieczenie	150	0.073	0.253	1.306	
5	Suma aktywów	Zabezpieczenie	90	0.267	0.919	1.033	

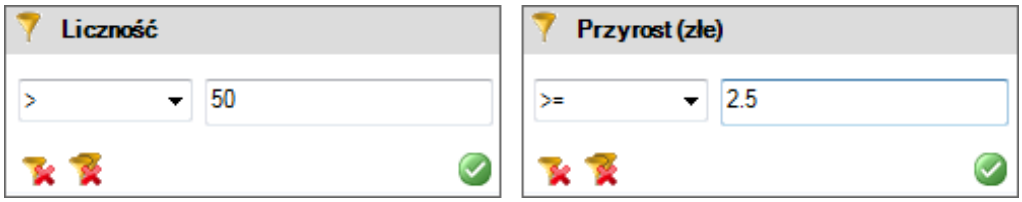
Każdą regułę możemy ocenić za pomocą wskaźników ***Bad rate*** oraz przyrostu zarówno dla klasy pozytywnej ***Przyrost (dobre)*** jak i negatywnej ***Przyrost (złe)***. Szczegóły obliczeniowe patrz Scorecard Formula Guide.

Po zaznaczeniu w tabeli wybranej zmiennej, w polu ***Reguła*** pojawia się jej pełna treść. Klikając na polu w kolumnie ***Drzewo*** możemy wyświetlić diagram drzewa, na podstawie którego określono bieżącą regułę.

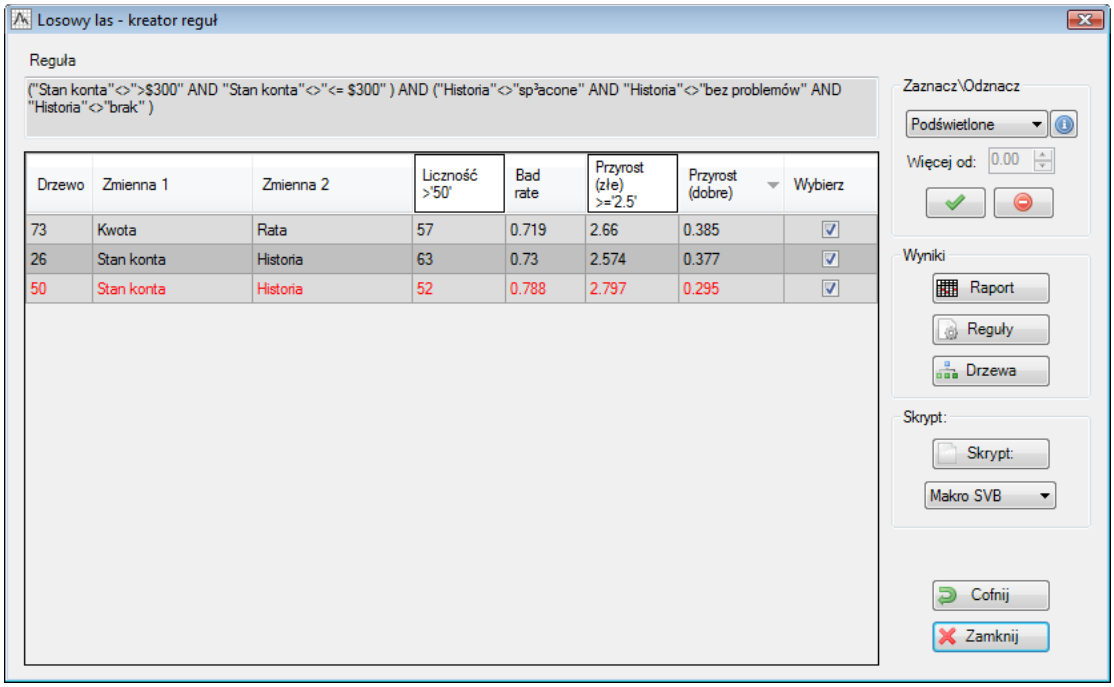


Za pomocą opcji zawartych w obszarze ***Zaznacz\odznacz*** możemy filtrować interesujące nas reguły zaznaczając te, które spełniają nasze kryteria odnośnie jakości.

Mamy również możliwość filtrowania utworzonych reguł za pomocą bardziej zaawansowanych kryteriów. Kliknięcie prawym przyciskiem myszy na wybranym nagłówku kolumny w tabeli pozwala na zdefiniowanie szczegółowych filtrów wyboru. Załóżmy, że chcielibyśmy wyświetlić reguły o liczności większej niż 50 oraz przyroście klasy złej nie mniejszym niż 2,5. Klikamy zatem prawym przyciskiem myszy na odpowiednich nagłówkach kolumn tabeli i definiujemy filtry podobnie jak przedstawiono poniżej:



Po zdefiniowaniu filtrów w oknie ***Losowy las – kreator reguł*** wyświetlone zostaną jedynie reguły spełniające określone kryteria. Możemy je zaznaczyć klikając pole wyboru w kolumnie ***Wybierz***.

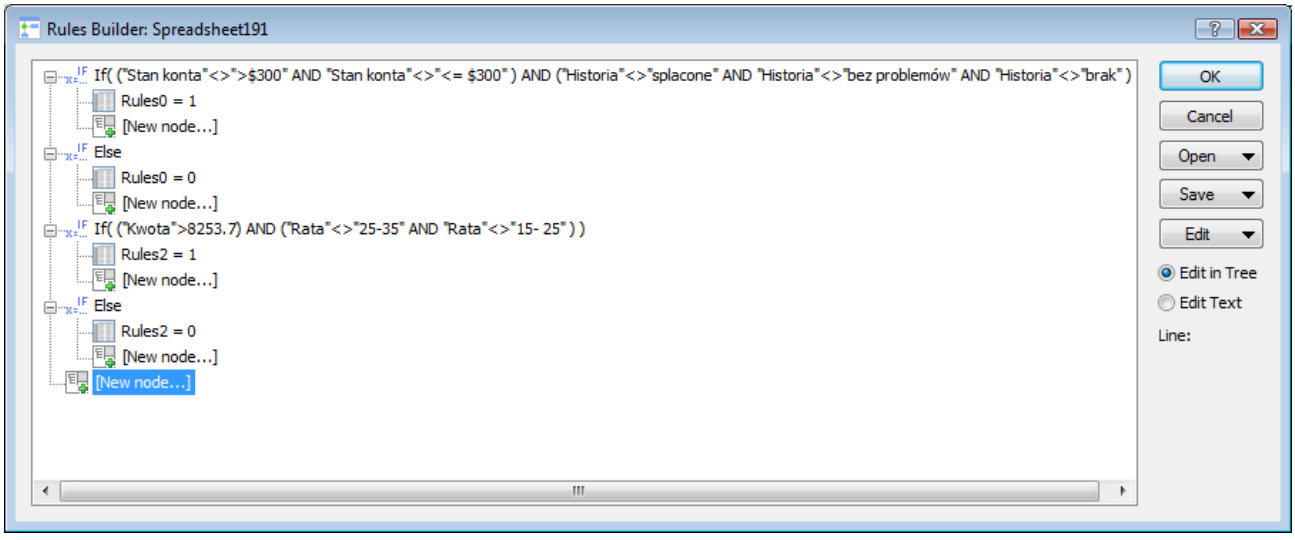


Po wybraniu interesujących reguł w obszarze ***Wyniki*** możemy wygenerować tabelę z dokładnym tekstem reguły oraz wskaźnikami jej jakości.

Klikając przycisk ***Reguły*** uzyskujemy arkusz danych zawierający pierwotny zestaw cech oraz zmienne zero-jedynkowe reprezentujące każdą z wygenerowanych przez moduł reguł.

Przycisk ***Drzewa*** utworzy diagramy wszystkich drzew, do których odnosiły się wybrane reguły.

Za pomocą opcji ***Skrypt*** możemy wygenerować skrypt z wybranymi regułami w postaci makra Visual Basic bądź raportu *STATISTICA*. Dodatkowo, jeżeli użytkownik posiada odpowiednią licencję może skorzystać z opcji ***Reguły*** co spowoduje wyświetlenie utworzonych reguł w specjalnym oknie, gdzie możliwa będzie ich dodatkowa edycja, uruchomienie oraz zapis w bazie metadanych. Więcej informacji zobacz *STATISTICA Konstruktor Reguł.*



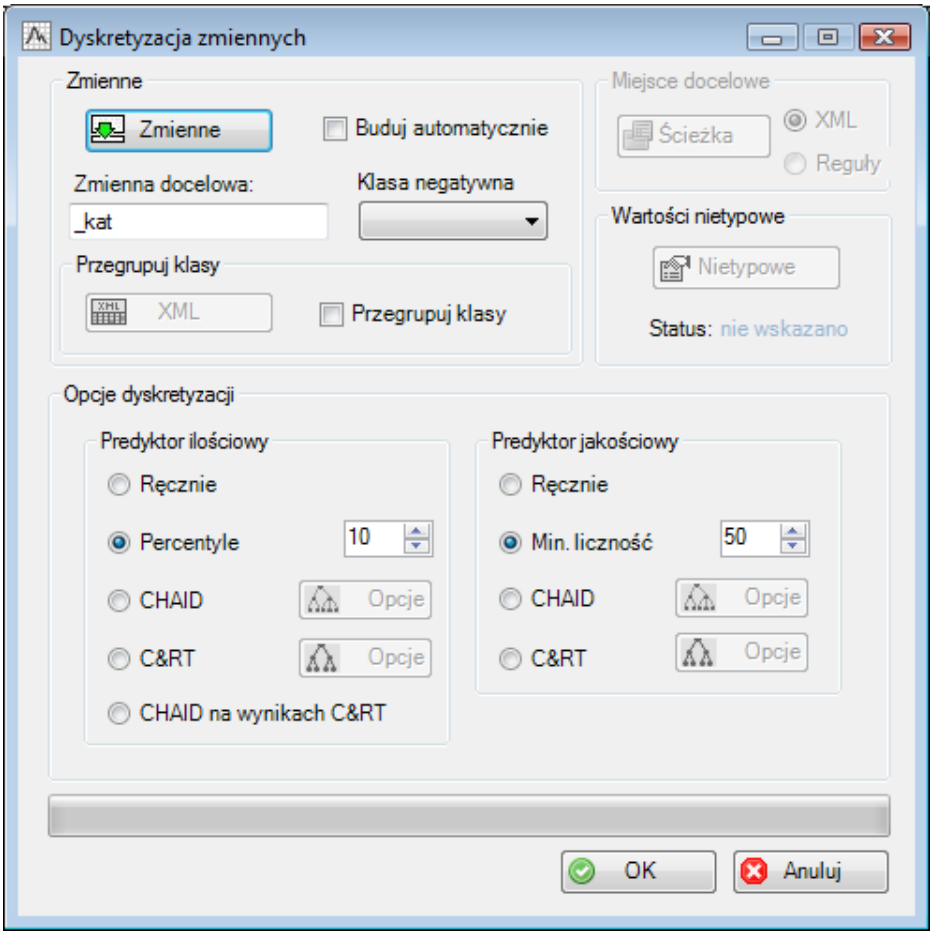
Wybrane reguły mogą zostać wdrożone do bazy systemu *STATISTICA Enterprise* i używane przez jego użytkowników podobnie jak inne obiekty *STATISTICA Enterprise*.

3.3. Dyskretyzacja zmiennych

Ponieważ tablica skoringowa jest narzędziem stosowanym dla danych dyskretnych (każda z analizowanych cech jest podzielona na przedziały), kolejnym etapem procesu przygotowania tablicy jest dyskretyzacja zmiennych ilościowych oraz ewentualna rekategoryzacja zmiennych jakościowych. Służy do tego moduł ***Dyskretyzacja zmiennych***, który tworzy przedziały odwzorowane następnie na karcie skoringowej. W programie zaimplementowano miary umożliwiające ocenę dobroci podziału oraz mocy predykcyjnej przekształcanej zmiennej: *WoE* (*weight of evidence*), za pomocą której mierzymy moc predykcyjną danego poziomu zmiennej oraz *IV* (*Information Value*), za pomocą której badamy moc całej cechy. Oczywiście proces dyskretyzacji może odbywać się automatycznie, jednak bardziej zalecane jest przekształcanie interakcyjne, bazujące na analizie stosownych raportów informujących o wartościach *WoE* oraz *IV*, a także przebiegu wykresu *WoE* (szczegóły patrz przykład 4. i 5.). Na podstawie tych informacji badacz musi ocenić, czy proponowany sposób dyskretyzacji jest prawidłowy i optymalny. Po znalezieniu optymalnego sposobu dyskretyzacji, ustalony przepis zapisywany jest w postaci skryptu *XML* interpretowanego przez moduł ***Budowa tablicy skoringowej***.

Przykład 4.– Dyskretyzacja zmiennej ilościowej

Przykład opiera się na pliku danych *CreditScoring.sta* wykorzystywanym w poprzednich przykładach. Po jego otwarciu z menu ***Zestaw skoringowy*** wybieramy polecenie ***Przygotowanie danych*** a następnie ***Dyskretyzacja zmiennych***, aby wyświetlić okno ***Dyskretyzacja zmiennych***.



Użytkownik może sparametryzować wykonanie analizy za pomocą następujących opcji:

Opcja ***Buduj automatycznie*** pozwala wykonać automatyczną dyskretyzację wielu zmiennych równocześnie na podstawie wybranych ***Opcji dyskretyzacji***.

Zmienna docelowa określa przyrostek, jaki będzie dodawany do oryginalnej nazwy zmiennej po jej dyskretyzacji.

Klasa negatywna umożliwia wybór symbolu klasy niepożądanej spośród dwóch klas wybranej zmiennej zależnej.

Opcje zawarte w obszarze ***Miejsce docelowe*** są aktywne jedynie w przypadku, gdy wybrano opcję ***Buduj automatycznie***. Przycisk ***Ścieżka*** pozwala określić miejsce docelowe zapisu wyników analizy, opcje ***XML*** oraz ***Reguły*** pozwalają na wybór formatu zapisu definicji kategorii utworzonych dla wybranych zmiennych.

Za pomocą opcji ***Przegrupuj klasy*** użytkownik ma możliwość wczytania wcześniej przygotowanej definicji dyskretyzacji zapisanej w formacie XML. Dzięki tej opcji może on powrócić do stanu sprzed zapisu i skorygować odpowiednio sposób podziału zmiennej.

Wartości nietypowe są aktywne, jeżeli do analizy wybrano przynajmniej jeden predyktor ilościowy. Mogą być przydatne w sytuacji, gdy w naszym zbiorze występują braki danych. Przed analizą, braki danych powinny zostać zastąpione wybraną przez użytkownika wartością, którą następnie wskazujemy jako wartość nietypowa. Wartość ta będzie traktowana jako osobna kategoria w analizie.

Opcje dyskretyzacji – ***Predyktor ilościowy*** pozwalają na wybór odpowiedniej strategii dyskretyzacji zmiennych ilościowych.

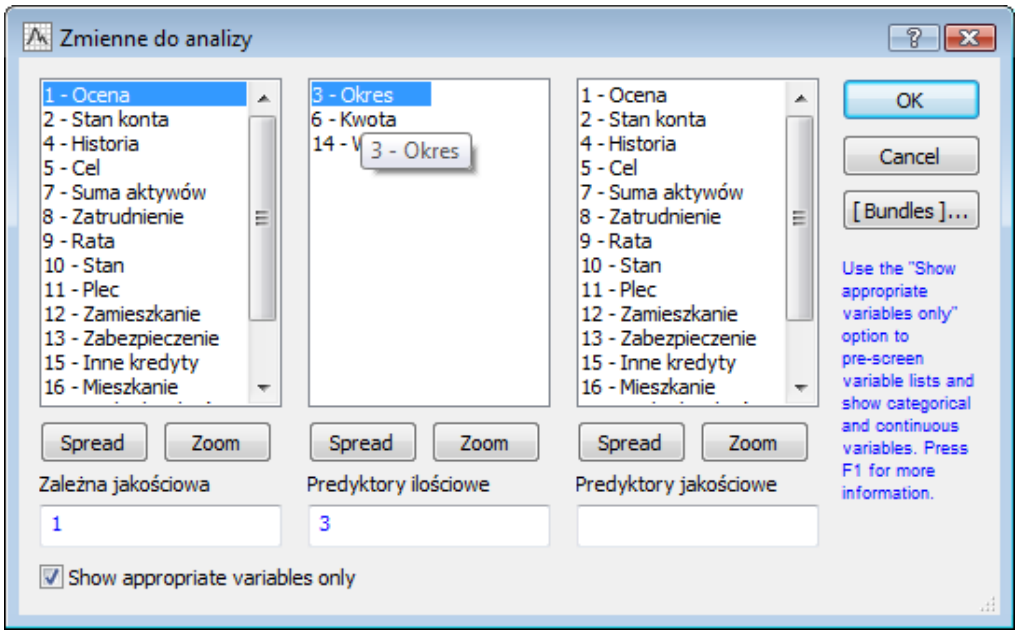
> Opcja ***Ręcznie*** nie tworzy żadnych wstępnych kategorii, użytkownik musi zdefiniować ręcznie własną kategoryzację.
>
> ***Percentyle*** dzieli zmienną ilościową na wskazaną liczbę równolicznych przedziałów. Użytkownik może zmieniać liczbę kategorii od 2 do 100.
>
> ***CHAID*** tworzy model drzew CHAID na podstawie zmiennej zależnej oraz wybranego predyktora. Granice podziałów otrzymane w wyniku budowy modelu utworzą wyjściowe granice klas. Więcej informacji na temat algorytmu CHAID oraz jego ***Opcji*** można znaleźć w pomocy elektronicznej programu *STATISTICA*.
>
> ***C&RT*** tworzy model drzew CART na podstawie zmiennej zależnej oraz wybranego predyktora. Granice podziałów otrzymane w wyniku budowy modelu utworzą wyjściowe granice klas. Więcej informacji na temat algorytmu CART oraz jego ***Opcji*** można znaleźć w pomocy elektronicznej programu *STATISTICA*.
>
> ***CHAID na wynikach CART*** tworzy w pierwszej kolejności model drzew CART na podstawie zmiennej zależnej oraz wybranego predyktora. Granice podziałów otrzymane w wyniku budowy modelu utworzą wyjściowe granice klas. Wyjściowe granice są punktem startowym do budowy modelu CHAID tworzącego finalne granice klas.

Opcje dyskretyzacji – ***Predyktor jakościowy*** pozwalają na wybór odpowiedniej strategii dyskretyzacji zmiennych jakościowych.

> Opcja ***Ręcznie*** nie tworzy żadnych wstępnych kategorii, użytkownik musi zdefiniować ręcznie własną kategoryzację.
>
> ***Min. Liczność*** agreguje do wspólnego atrybutu kategorie, których liczność jest mniejsza niż wskazana w opcji.
>
> ***CHAID*** tworzy model drzew CHAID na podstawie zmiennej zależnej oraz wybranego predyktora. Węzły końcowe otrzymane w wyniku budowy modelu utworzą wyjściowe kategorie. Więcej informacji na temat algorytmu CHAID oraz jego ***Opcji*** można znaleźć w pomocy elektronicznej programu *STATISTICA*.

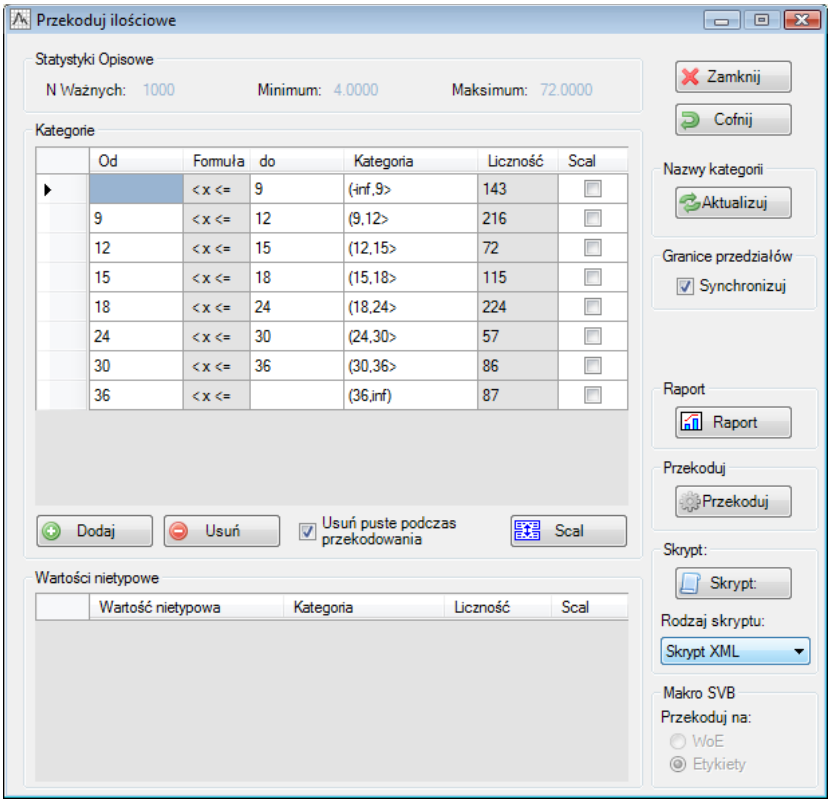
C&RT tworzy model drzew CART na podstawie zmiennej zależnej oraz wybranego predyktora. Węzły końcowe otrzymane w wyniku budowy modelu utworzą wyjściowe kategorie. Więcej informacji na temat algorytmu CART oraz jego ***Opcji*** można znaleźć w pomocy elektronicznej programu *STATISTICA*.

W pierwszym kroku analizy wybieramy zmienne, które będą używane podczas przekodowania. W tym celu klikamy przycisk ***Zmienne*** i w wyświetlonym oknie wskazujemy zmienną *Zależną jakościową* informującą o statusie kredytu (w naszym przypadku jest to zmienna *Ocena)* oraz zmienną, którą chcemy przekształcać (lista *Predyktory ilościowe*), w naszym przypadku będzie to zmienna *Okres*.



W kolejnym kroku na liście ***Klasa negatywna*** wskazujemy *zły* jako symbol klasy negatywnej. Wybieramy opcję podziału na domyślną liczbę percentyli. Zatwierdzamy analizę klikając przycisk ***OK***. Wyświetlone zostanie okno ***Przekoduj ilościowe***, w którym, w tabeli ***Kategorie***, określone zostały przedziały zaproponowane przez wybraną opcję analizy.

Uwaga. *Budowa atrybutów nigdy nie powinna odbywać się w sposób całkowicie zautomatyzowany. Podział na percentyle lub klasy otrzymane algorytmem CHAID lub CART należy traktować tylko jako wstępny podział, który powinien zostać dokładnie przeanalizowany i zmodyfikowany tak, aby uwzględniał posiadaną wiedzę biznesową. Brak staranności w budowie atrybutów będzie skutkował słabym modelem o małej zdolności do generalizacji.*



W oknie ***Przekoduj ilościowe*** możemy dowolnie zmieniać granice przedziałów edytując pola ***Od*** oraz ***do*** zawarte w tabeli. Nazwy przedziałów możemy edytować w polach kolumny ***Kategoria***.

Przyciski ***Dodaj*** oraz ***Usuń*** umożliwiają odpowiednio dodawanie nowych kategorii i usuwanie już istniejących.

Opcja ***Usuń puste podczas przekodowania*** usuwa kategorie bez żadnego przypadku po kliknięciu przycisku ***Przekoduj***.

Przycisk ***Scal*** pozwala łączyć ze sobą kategorie z zaznaczoną opcją ***Scal*** (mogą to być zarówno regularne kategorie zmiennej jak i kategorie utworzone na wartościach nietypowych).

Przycisk ***Aktualizuj*** zmienia nazwy kategorii tak, by odzwierciedlały aktualne granice przedziałów.

Opcja ***Synchronizuj*** powoduje, że zmiany granic przedziałów wprowadzone w jednym atrybucie przenoszą się do sąsiedniego atrybutu.

Przycisk ***Raport*** tworzy raport składający się z trzech dokumentów: wykresu WoE opisującego profil ryzyka, szczegółowego raportu podsumowującego powstałe kategorie oraz szczegółowego opisu powstałych atrybutów. Przycisk ten jest aktywny tylko po kliknięciu przycisku ***Przekoduj***. Każda zmiana w definicji atrybutu sprawia, że przycisk ten jest nieaktywny.

Przycisk ***Przekoduj*** powoduje przekodowanie zmiennej zgodnie z bieżącą definicją atrybutów. Kliknięcie przycisku ***Przekoduj*** uaktywnia przyciski ***Raport*** oraz ***Skrypt***.

Przycisk ***Skrypt*** pozwala utworzyć dokument zawierający definicję atrybutów. Wybór opcji ***Skrypt XML***, tworzy plik, który może być użyty w modułach ***Budowa karty skoringowej*** oraz ***SURVIVAL***. Dodatkowo skrypt taki może być użyty w opcji ***Przegrupuj klasy*** pozwalającej przedefiniować definicję atrybutów. Opcja ***Makro SVB*** i ***Węzeł Data Miner*** pozwala na pisanie skryptów tworzących zmienne pochodne oparte na definicjach atrybutów. Opcja ***Plik reguł (SRX)*** oraz ***Reguły*** mogą być wykorzystywane do przygotowywania karty skoringowej za pomocą osobnego modułu *STATISTICA Konstruktor Reguł*. Wybrane reguły mogą zostać wdrożone do bazy systemu *STATISTICA Enterprise* i używane przez jego użytkowników podobnie jak inne obiekty *STATISTICA Enterprise*.

Uwaga. *Weight of Evidence (WoE) jest bardzo pomocną miarą w ocenie atrybutów danej zmiennej, bazującą na logarytmie szansy zajścia określonego zdarzenia.*

$$\text{WoE} = \left[\ln\left(\frac{\text{Distr Good}}{\text{Distr Bad}} \right) \right] \cdot 100$$

Podczas przygotowywania profili ryzyka należy zawsze obliczyć wskaźnik WoE dla każdego z atrybutów (przedziałów zmiennej) i ocenić różnice między nimi. Zbliżone wartości WoE sugerują aby połączyć dwa lub więcej atrybutów w jeden. Podczas łączenia należy kierować się następującymi zasadami

- *w każdej klasie powinno być przynajmniej 5% obserwacji i 5% złych kredytów*
- *wartości WoE powinny się wyraźnie różnić pomiędzy atrybutami*
- *trend obserwowany w wartościach WoE powinien zgadzać się z wiedzą biznesową.*

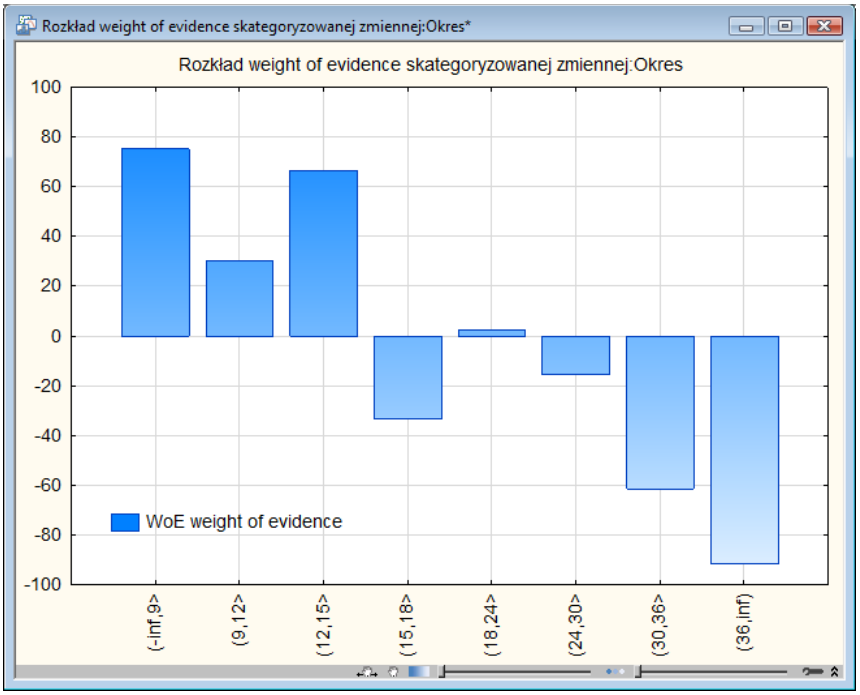
Proces budowy atrybutów dla zmiennej *Okres* rozpoczniemy od sprawdzenia liczności i charakterystyki otrzymanych w wyniku podziału na percentyle wstępnych klas. W celu sprawdzenia liczności poszczególnych klas oraz oceny jakości podziału klikamy przycisk ***Przekoduj***. Po naciśnięciu tego przycisku aktywne stają się przyciski ***Raport*** oraz ***Skrypt***. Za pomocą przycisku ***Raport*** wyświetlamy informację o jakości podziału (miary *WoE* oraz *IV*, a także wykres *WoE*). Informacja ta umożliwia ocenę mocy predykcyjnej zmiennej i poszczególnych jej kategorii oraz podjęcie decyzji o ewentualnej korekcie proponowanego podziału.

Dane: Tabela Woe*

Kategoryzowana zmienna:Okres

Okres_kat	Dobry	Zły	Suma	Portfel	Procent dobrych	Procent złych	Procent ogółem	IV	WoE weight of
(-inf,9>	119	24	143	16,78%	17,00%	8,00%	14,30%	0,07	75,38
(9,12>	164	52	216	24,07%	23,43%	17,33%	21,60%	0,02	30,13
(12,15>	59	13	72	18,06%	8,43%	4,33%	7,20%	0,03	66,53
(15,18>	72	43	115	37,39%	10,29%	14,33%	11,50%	0,01	-33,18
(18,24>	158	66	224	29,46%	22,57%	22,00%	22,40%	0,00	2,56
(24,30>	38	19	57	33,33%	5,43%	6,33%	5,70%	0,00	-15,42
(30,36>	48	38	86	44,19%	6,86%	12,67%	8,60%	0,04	-61,37
(36,inf)	42	45	87	51,72%	6,00%	15,00%	8,70%	0,08	-91,63
Ogół	700	300	1000	30,00%	100,00%	100,00%	100,00%	0,25	

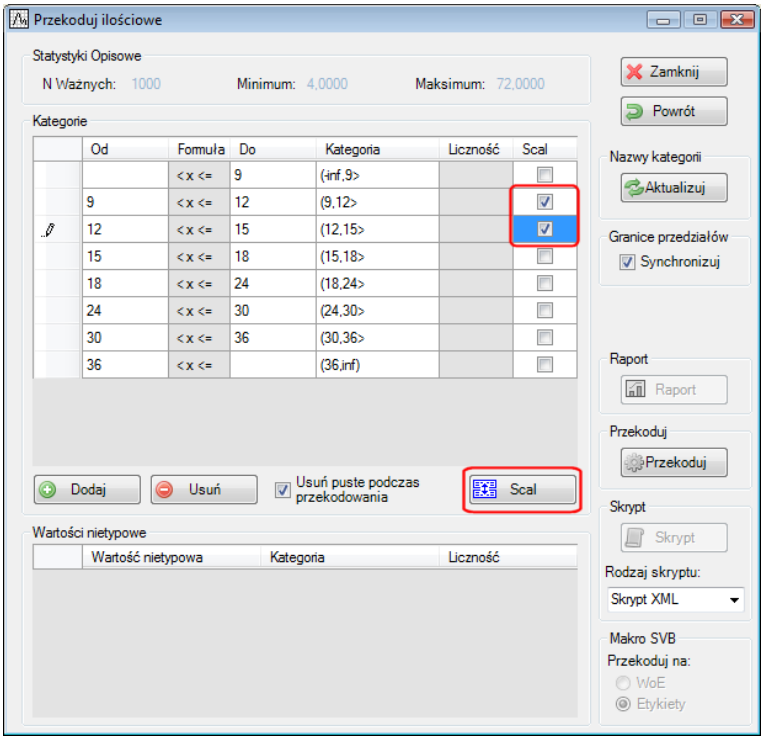
Wykres *WoE* umożliwia łatwą ocenę, czy obserwowany trend wartości *WoE* zgadza się z wiedzą biznesową.



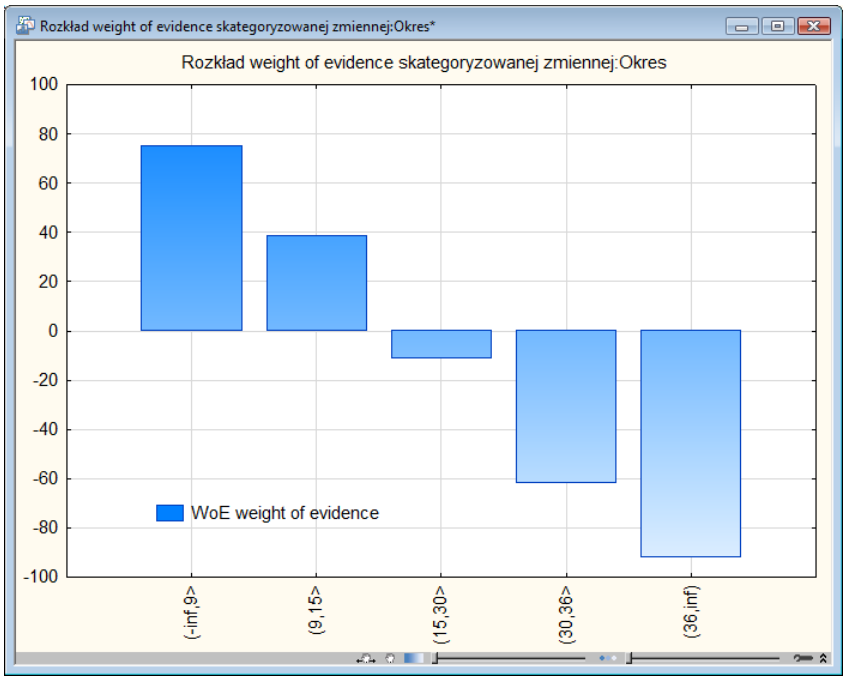
Analizując powyższe wyniki, na podstawie wartości IV możemy stwierdzić, że zmienna *Okres* mieści się w przedziale średniej mocy predykcyjnej. Można zauważyć, że im dłuższy jest okres kredytowania tym większe jest ryzyko generowane przez dłużników. Pierwsze trzy atrybuty mają wartości większe od zera co oznacza niskie ryzyko. Przy prawej krawędzi wykresu dwa ostatnie atrybuty przyjmują ujemne wartości WoE co oznacza dosyć duży poziom ryzyka.

Można także zauważyć lokalne odwrócenia trendu, które mogą być skutkiem nieliniowej zależności między analizowaną zmienną a ryzykiem. Często jednak są one związane z losowymi zaburzeniami danych i jako takie będziemy je traktować w tym przykładzie. Wygładzimy więc trend by je wyeliminować. W pierwszym kroku scalimy kategorię 2 i 3, a w drugim 4, 5 i 6.

W tym celu w wierszach 2 i 3 zaznaczamy opcję ***Scal***, a następnie klikamy przycisk ***Scal***. Podobną operację wykonujemy dla kolejnych trzech kategorii.



Po wykonaniu przekodowania i ponownym wygenerowaniu raportu, wykres *WoE* wygląda następująco:



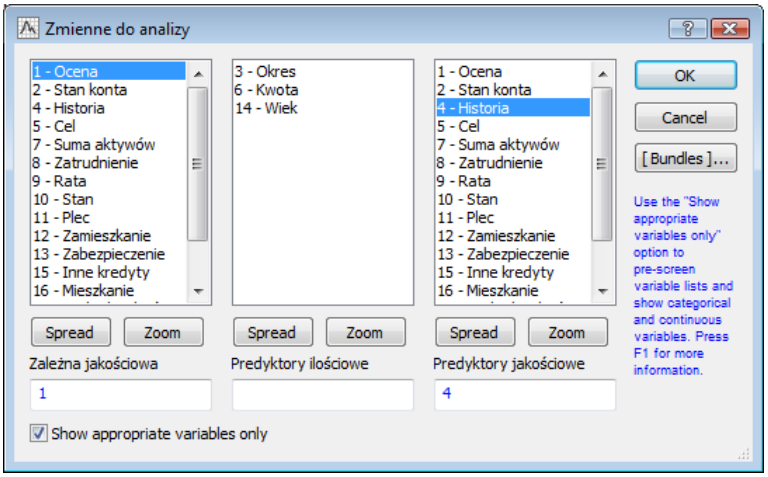
Po scaleniu ze sobą odpowiednich klas udało nam się otrzymać logiczny i monotoniczny trend wartości WoE. Załóżmy, że taka dyskretyzacja dobrze odzwierciedla prawdziwy profil ryzyka. Zapiszemy więc tak przygotowany skrypt dyskretyzacji do pliku XML klikając przycisk ***Skrypt***.

Uwaga. *Użytkownik ma możliwość zmiany wcześniej przygotowanych atrybutów, które zostały zapisane w postaci skryptu XML. W tym celu wystarczy w oknie* ***Dyskretyzacja zmiennych*** *w obszarze* ***Przegrupuj klasy*** *kliknąć przycisk* ***XML*** *i wybrać odpowiedni plik.*

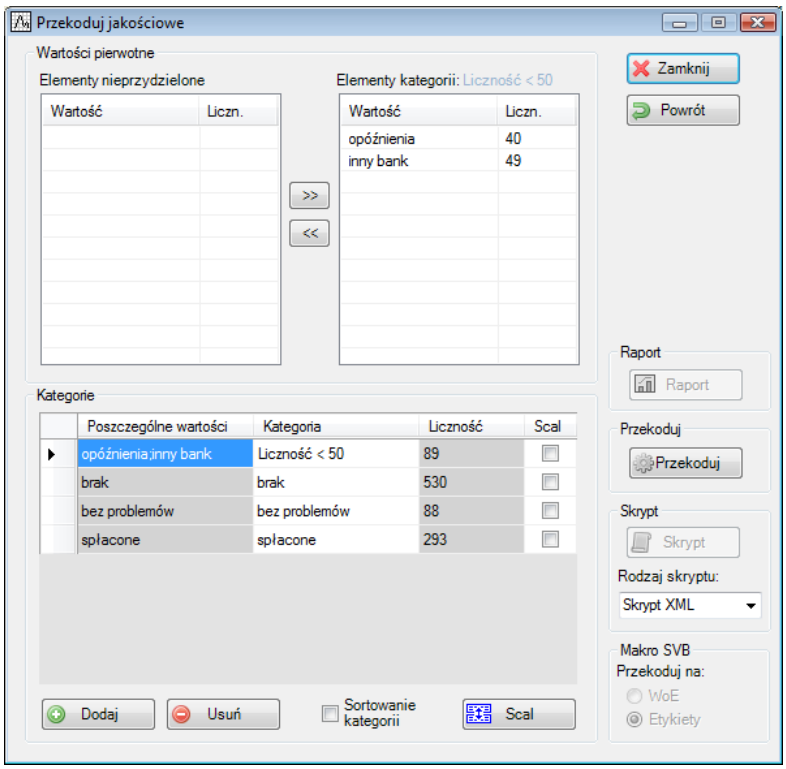
Przykład 5. – Rekategoryzacja zmiennej jakościowej

W przypadku zmiennych jakościowych naszym celem jest zminimalizowanie liczby klas danej zmiennej poprzez łączenie ze sobą klas o małej liczności oraz klas o podobnym wskaźniku *WoE*.

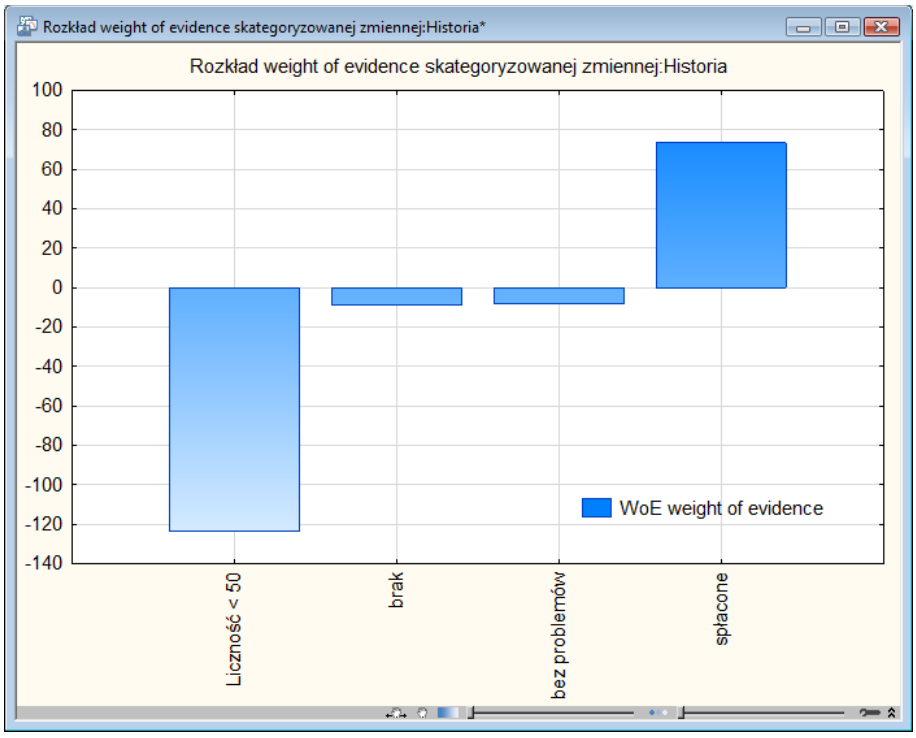
Tym razem do analizy wybierzemy zmienną *Historia*. Ponieważ wybrana zmienna jest zmienną jakościową, w oknie ***Dyskretyzacja zmiennych*** klikamy ***Zmienne***, a następnie wybieramy zmienne zgodnie z poniższym zrzutem.



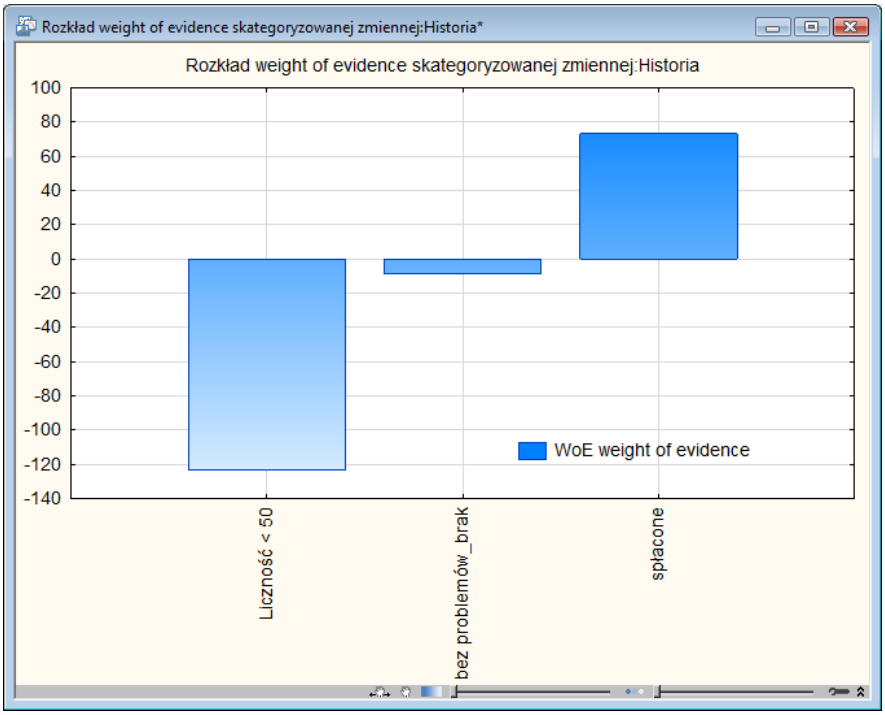
Po wybraniu zmiennych w obszarze ***Predyktor jakościowy*** wybieramy opcję ***Min. liczność***, pozostawiając wartość tego parametru na poziomie 50. Po kliknięciu przycisku ***OK*** wyświetlone zostanie okno ***Przekoduj jakościowe***.



Warto zauważyć, że liczność klas *Opóźnienia* oraz *Inny bank* jest mniejsza od wartości granicznej wynoszącej 50, dlatego obie klasy zostały ze sobą scalone tworząc kategorię *Liczność < 50*. Zobaczmy jak wygląda wykres *WoE* dla powstałych klas. W tym celu klikamy przycisk ***Przekoduj***, a następnie, za pomocą przycisku ***Raport***, uzyskujemy skoroszyt z raportem.



Na wykresie widać wyraźnie, że wartość wskaźnika *WoE* dla klas *brak* oraz *bez problemów* jest praktycznie taka sama. Dla poprawy jakości modelu należy scalić obie te klasy w sposób analogiczny, jak w przypadku zmiennej ilościowej.



Dane: Tabela Woe*

Kategoryzowana zmienna:Historia

Historia_kat	Dobry	Zły	Suma	Portfel	Procent złych	Procent dobrych	Procent ogółem	IV	WoE weight of evidence
Licznosc < 50	36	53	89	59,55%	17,67%	5,14%	8,90%	0,15	-123,41
bez problemów_brak	421	197	618	31,88%	65,67%	60,14%	61,80%	0,00	-8,79
spłacone	243	50	293	17,06%	16,67%	34,71%	29,30%	0,13	73,37
Ogół	700	300	1000	30,00%	100,00%	100,00%	100,00%	0,29	

Analizując tabelę *WoE* możemy zauważyć, że siła predykcyjna zmiennej, mierzona na podstawie wartości wskaźnika *IV*, jest dosyć duża (na granicy średniej i mocnej) – wynosi 0,29.

Podobnie jak w przypadku zmiennej ilościowej, po wykonanej rekategoryzacji, zapisujemy przepis przekodowania w pliku *XML*. Wykorzystamy ten plik podczas budowy modelu.

Przykład 6. – Rekategoryzacja zmiennej zawierającej braki danych lub obserwacje odstające

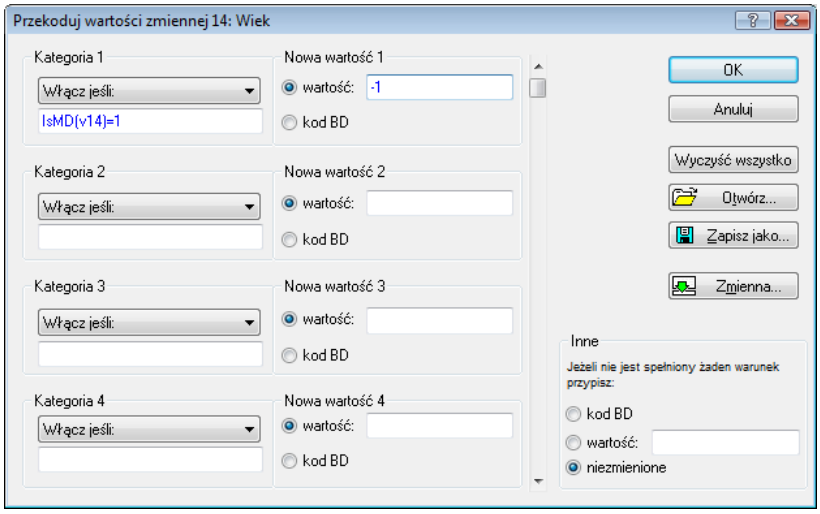
Czasami zdarza się, że zmienne zawierają braki danych lub obserwacje odstające. Jest kilka możliwości postępowania w takich przypadkach:

- usunięcie przypadków z dalszej analizy
- zastąpienie braków np. średnią czy medianą
- wykorzystanie statystycznych metod jak metoda k-najbliższych sąsiadów.

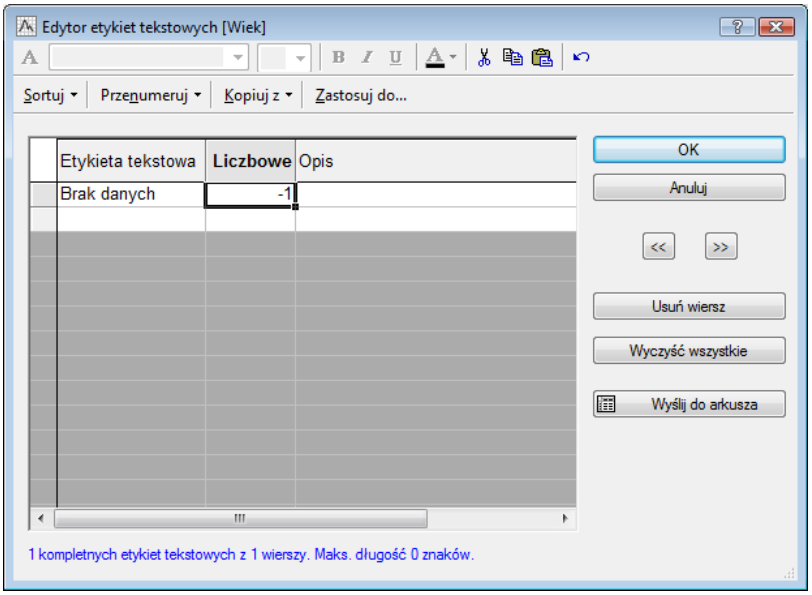
Biorąc pod uwagę specyfikę budowy modeli skoringowych możemy postąpić jeszcze inaczej. Zastąpimy braki danych pewną stałą wartością spoza zakresu zmienności analizowanej cechy umożliwiając wykorzystanie informacji zawartej właśnie w brakach danych, traktując brak odpowiedzi jak każdy inny poziom zmiennej.

W przykładzie wykorzystamy plik *BrakiDanych.sta*, który w zmiennej *Wiek* zawiera braki danych.

W celu uwzględnienia braków danych w dalszej analizie klikamy na zmienną *Wiek*, a następnie wybieramy ***Dane*** | ***Przekoduj***. W pierwszej kategorii wpisujemy formułę (*IsMD(vXX)=1*) , gdzie XX to numer zmiennej, a w polu ***Nowa wartość 1*** wpisujemy na przykład -1. Wybór zatwierdzamy przyciskiem ***OK***.



W przekodowanej zmiennej możemy dodatkowo utworzyć odpowiednią etykietę tekstową. Przywołujemy okno specyfikacji zmiennej klikając dwukrotnie na jej nagłówek, a następnie klikamy przycisk ***Etykiety tekstowe***. W kolumnie ***Etykieta tekstowa*** wpisujemy na przykład *Brak danych*, a w kolumnie ***Liczbowe*** wartość, na którą przekodowaliśmy braki danych w poprzednim kroku – w tym przypadku będzie to -1.

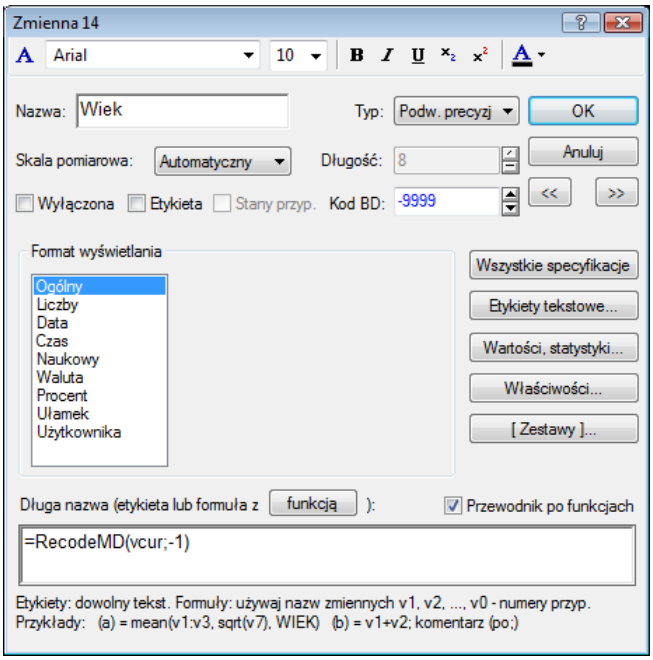


Po kliknięciu ***OK*** zostanie utworzona odpowiednia etykieta tekstowa. Inny sposób zamiany braków danych to wykorzystanie długiej nazwy zmiennej. W tym celu klikamy dwukrotnie na nagłówku zmiennej *Wiek* i w polu *Długa nazwa* wpisujemy *=RecodeMD(vcur;-1).*

Uwaga. *Formuły arkusza wpisywane po znaku równa się „=" to bardzo efektywny sposób operowania na danych zawartych w arkuszach STATISTICA. W tym przypadku posługując się funkcją RecodeMD bardzo szybko możemy przekodować braki danych na zadaną wartość. Funkcja przyjmuje dwa argumenty*

- *Numer zmiennej – określa, w której zmiennej ma zostać wykonane przekodowanie, vcur – zwraca numer bieżącej zmiennej.*

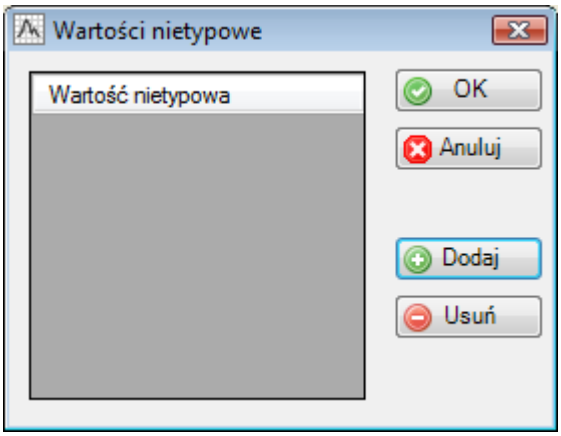
- *Kod – określa na jaką wartość ma zostać przekodowany brak danych.*



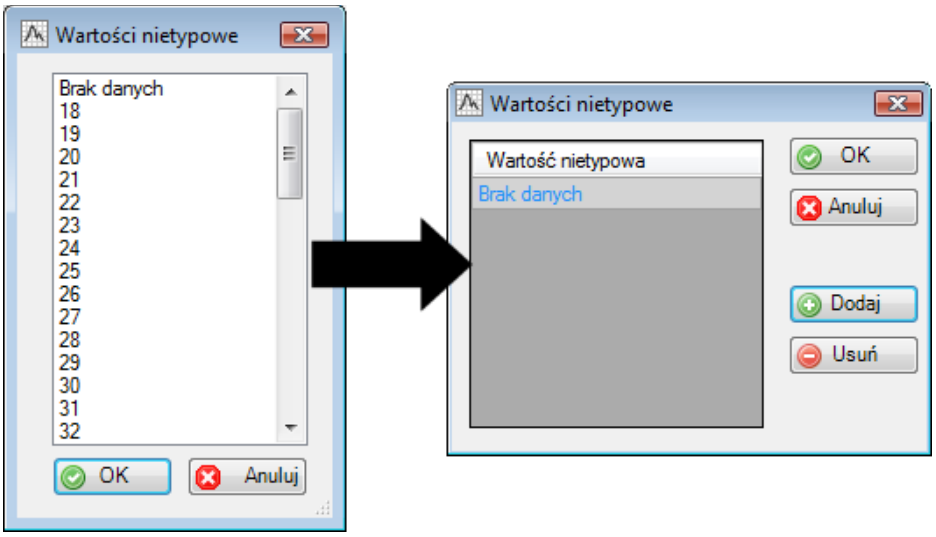
Następnie musimy podobnie jak poprzednio utworzyć dla wartości -1 etykietę tekstową.

W tym momencie zmienna jest już przygotowana do dyskretyzacji. Uruchamiamy więc moduł wybierając z menu ***Przygotowanie danych*** | ***Dyskretyzacja zmiennych*** i jako predyktor ilościowy wybieramy zmienną *Wiek*, oraz wskazujemy kod złego kredytu.

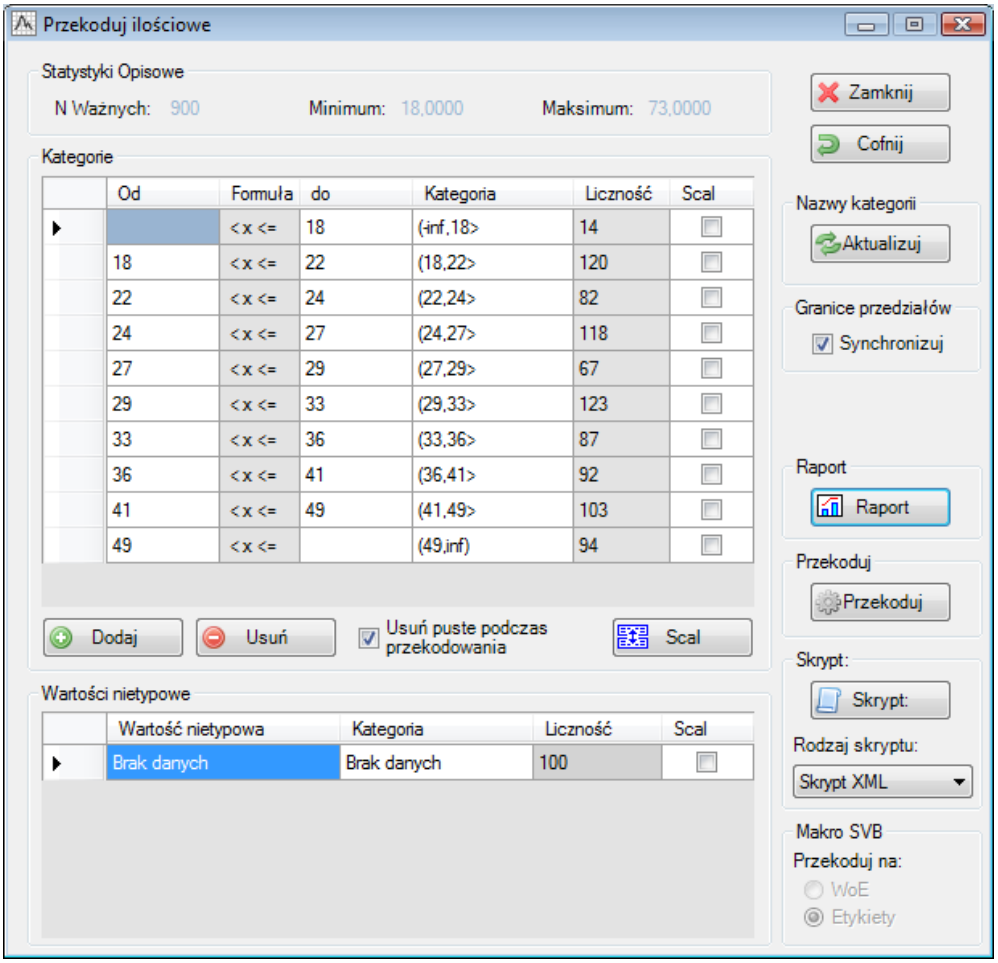
Kolejnym krokiem będzie uwzględnienie w analizie informacji o wartościach nietypowych. Klikamy przycisk ***Nietypowe***, a następnie przycisk ***Dodaj***.



Po kliknięciu przycisku zostanie wywołane okno zawierające listę wszystkich występujących w danej zmiennej wartości. Z listy wybieramy utworzoną wcześniej kategorię ***Brak danych*** oraz zatwierdzamy wybór klikając ***OK***.



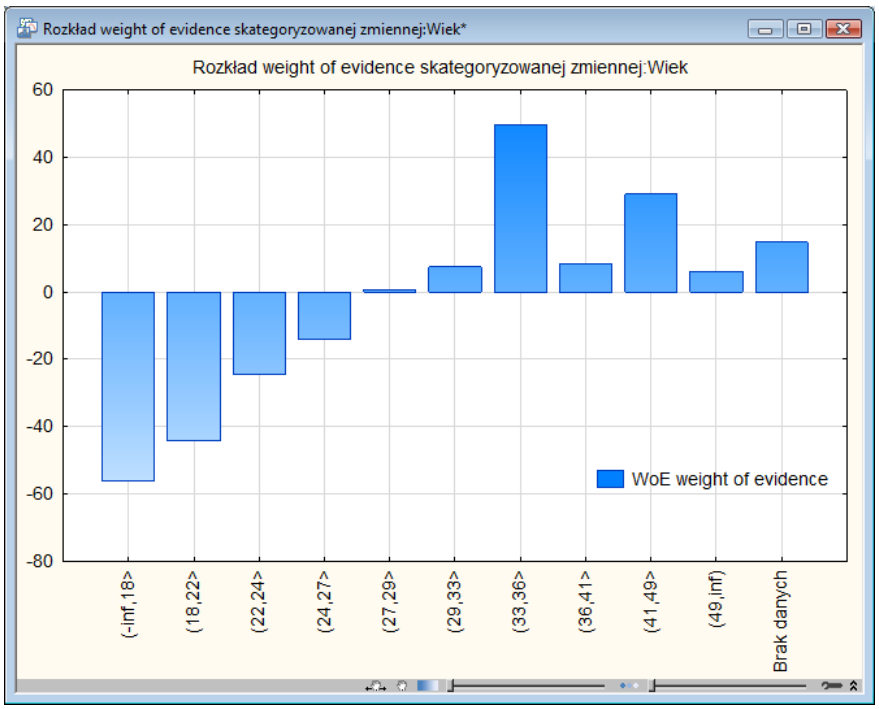
Podobnie jak w poprzednim przykładzie wybieramy opcję podziału na wskazaną liczbę percentyli. Zatwierdzamy analizę klikając przycisk ***OK***. Wyświetlone zostanie okno ***Przekoduj ilościowe***, w którym, w tabeli ***Kategorie***, określone zostały przedziały zaproponowane przez wybraną opcję analizy. Ponieważ w zmiennej występowały wartości nietypowe, utworzona została dodatkowa klasa widoczna w dolnej części okna w obszarze ***Wartości nietypowe***.



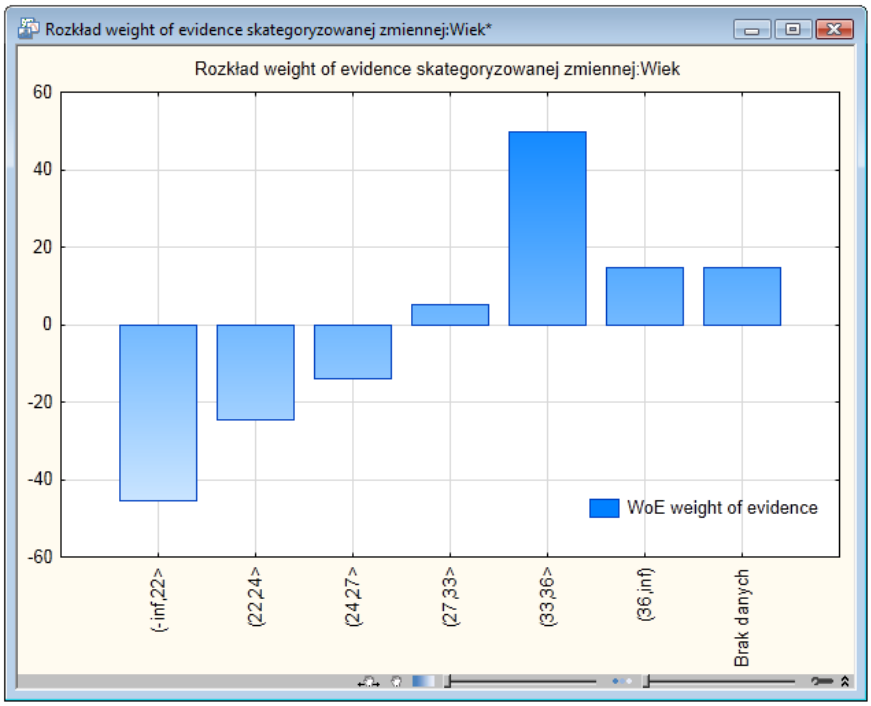
Podobnie jak poprzednio kategorie możemy dowolnie łączyć, zmieniać granice i nazwy przedziałów edytując odpowiednie pola, a także korzystać z przycisków ***Dodaj*** i ***Usuń***.

W celu sprawdzenia liczności poszczególnych klas oraz oceny jakości podziału klikamy przycisk ***Przekoduj***. Po naciśnięciu tego przycisku aktywne stają się przyciski ***Raport*** oraz ***Skrypt***. Za pomocą

przycisku ***Raport*** wyświetlamy informację o jakości podziału (miary *WoE* oraz *IV*, a także wykres *WoE*). Informacja ta umożliwia ocenę mocy predykcyjnej zmiennej i poszczególnych jej kategorii oraz podjęcie decyzji o ewentualnej korekcie proponowanego podziału.

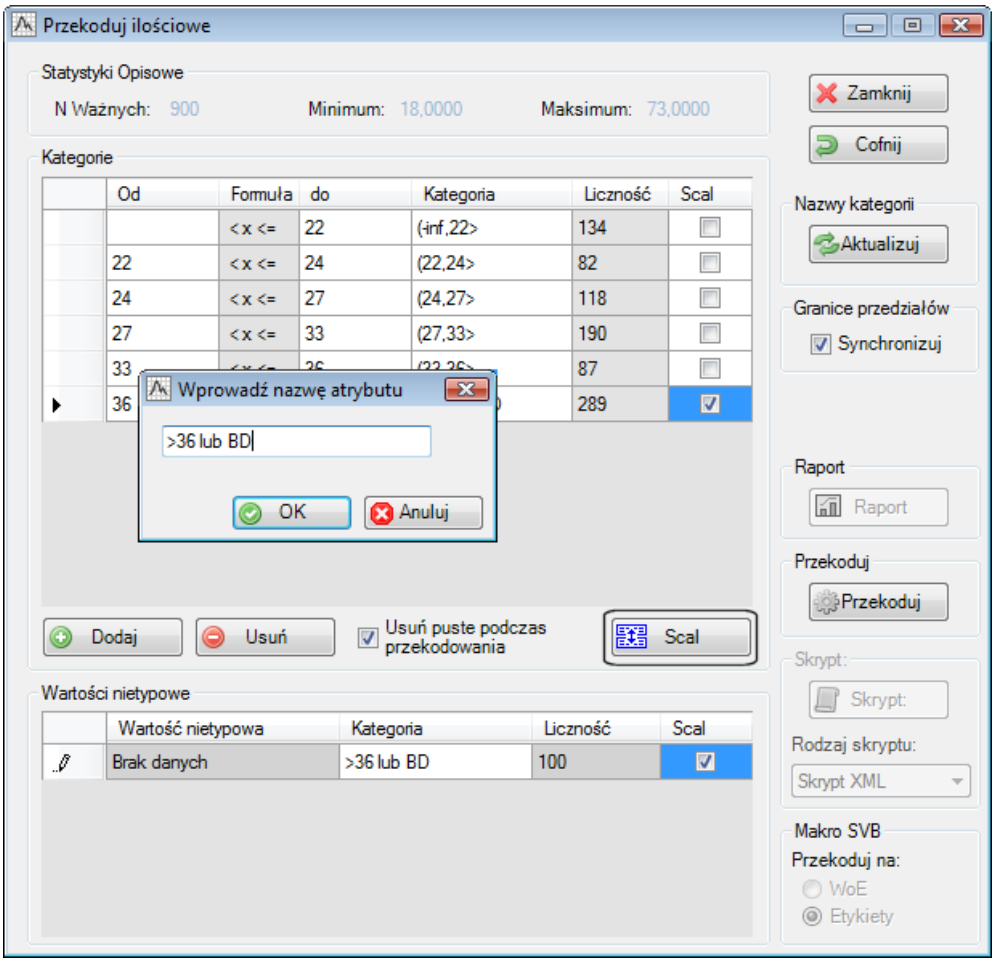


Jak widać na wykresie pojawiła się dodatkowa klasa umieszczona przy prawej krawędzi wykresu – *Brak danych*. Analizując profil ryzyka na wykresie widzimy, że trzy ostatnie klasy (oprócz braków danych) mają podobne wartości WoE. Podobnie klasa 5 i 6. Dodatkowo klasa (-inf,18> ma bardzo małą liczność, dlatego też zostanie połączona z klasą sąsiednią. Po scaleniu wykres WoE wygląda następująco:



Łatwo zauważyć, że dwie ostatnie klasy *(36, inf)* i *Brak danych* mają prawie identyczny poziom WoE. Aby je scalić należy w oknie ***Przekoduj ilościowe*** w kolumnie ***Scal*** obu klasom zaznaczyć opcję, a

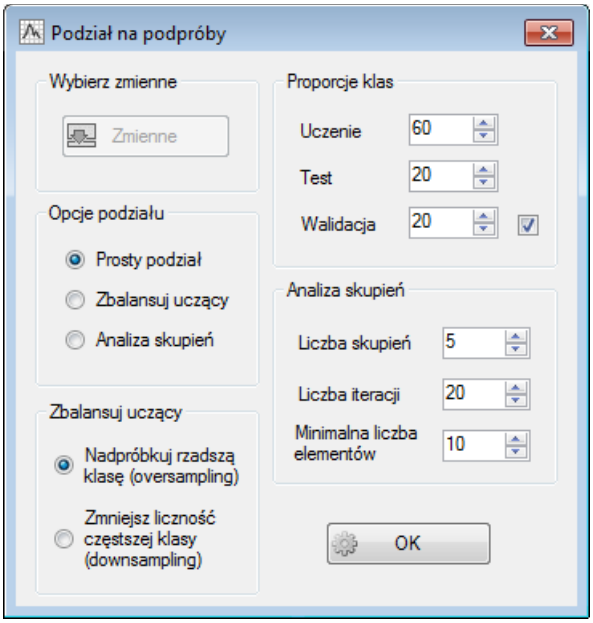
następnie kliknąć ***Przekoduj***. W wyświetlonym oknie ***Wprowadź nazwę atrybutu*** wprowadzamy wspólną nazwę nowej klasy.



Po scaleniu profil ryzyka wygląda logicznie i tak skategoryzowaną zmienną można wykorzystać do budowania modelu.

3.4. Podział na podpróby

Moduł Podział na podpróby pozwala na przygotowanie prób uczącej, testowej oraz (opcjonalnie) walidacyjnej na potrzeby budowy modeli skoringowych.



Moduł oferuje trzy główne opcje podziału zbioru na podpróby:

Prosty podział umożliwia na podzielenie zbioru danych w sposób losowy na klasy Uczącą, Testową oraz (opcjonalnie) Walidacyjną, zgodnie z proporcjami określonymi w grupie *Proporcje klas*.

Zbalansuj uczący pozwala na przygotowanie próby uczącej zawierającej równe proporcje klas zmiennej zależnej. Opcja ta jest przydatna w sytuacji, gdy problem, jaki chcemy analizować jest zadaniem klasyfikacyjnym, z dwoma klasami zmiennej zależnej. W zależności od wyboru opcji w grupie *Zbalansuj uczący*, program dokona nadpróbkowania (*oversampling*) rzadszej klasy lub zmniejszy liczebność częstszej klasy (*downsampling*).

Analiza skupień pozwala dokonać podziału na podpróby na podstawie losowania przypadków ze skupień utworzonych na podstawie analizy k-średnich dla dostępnych predyktorów. Metoda jest przydatna zwłaszcza w sytuacji mniejszych zbiorów danych, pozwala uniknąć nieproporcjonalnego rozłożenia się danej klasy przypadków w jednej z grup.

W wyniku działania modułu analityk otrzymuje nowy zbiór danych, który oprócz pierwotnego zestawu zmiennych zawiera dodatkowo kolumnę informującą do jakiej klasy (uczenie, test, walidacja) trafił konkretny przypadek.

4. Modelowanie

4.1. Budowa tablicy skoringowej

W module ***Budowa tablicy skoringowej*** zmienne wybrane do budowy modelu są przekształcane za pomocą skryptów XML, bądź plików reguł utworzonych w module ***Dyskretyzacja zmiennych***, a następnie na podstawie przekształconych danych budowany jest model regresji logistycznej. Podczas budowy modelu logistycznego mamy do wyboru różne strategie doboru cech. Domyślnie uwzględniane są wszystkie efekty. Istnieje również możliwość wyboru bardziej zaawansowanych strategii doboru cech, począwszy od *wprowadzania postępującego* i *eliminacji wstecznej*, aż po metody typu *„step-wise”*, czyli *metodę krokową postępującą* i *krokową wsteczną*. Kolejną opcją jest budowa modelu z wykorzystaniem strategii *bootstrap*. Po zbudowaniu modelu możemy ocenić jakość jego dopasowania (między innymi za pomocą miar *AIC - Akaike Information Criterion* oraz *BIC – Bayesian Information Criterion*), zbadać poziom korelacji i kowariancji cech oraz wyświetlić wartości ocen parametrów regresji. W celu utworzenia tablicy skoringowej z modelu logistycznego, należy w kolejnym kroku podać parametry skali, na podstawie których konstruowana jest tablica. Zbudowaną tablicę można następnie zapisać w dowolnej postaci (zaimplementowane formaty to XML, *STATISTICA* Visual Basic, Excel oraz plik reguł, jednak możliwe jest przygotowanie mechanizmu zapisującego tablicę skoringową do formatu wskazanego przez użytkownika). Jeżeli to możliwe należy podzielić zbiór danych na zbiór uczący i testowy. Na zbiorze uczącym przeprowadzić budowę tablicy skoringowej, a zbiór testowy wykorzystać do walidacji otrzymanych wyników.

Regresja logistyczna jest metodą, która wymaga by wszystkie zmienne niezależne (wejściowe) były cechami ilościowymi. W praktyce w grupie zmiennych niezależnych występują bardzo często cechy jakościowe (ponieważ wszystkie zmienne poddaliśmy dyskretyzacji, to *de facto* dysponujemy jedynie takimi zmiennymi). Analityk chcący uwzględnić tego typu cechy w modelu logistycznym zmuszony jest zamienić ich reprezentację na formę akceptowalną przez metodę regresji.

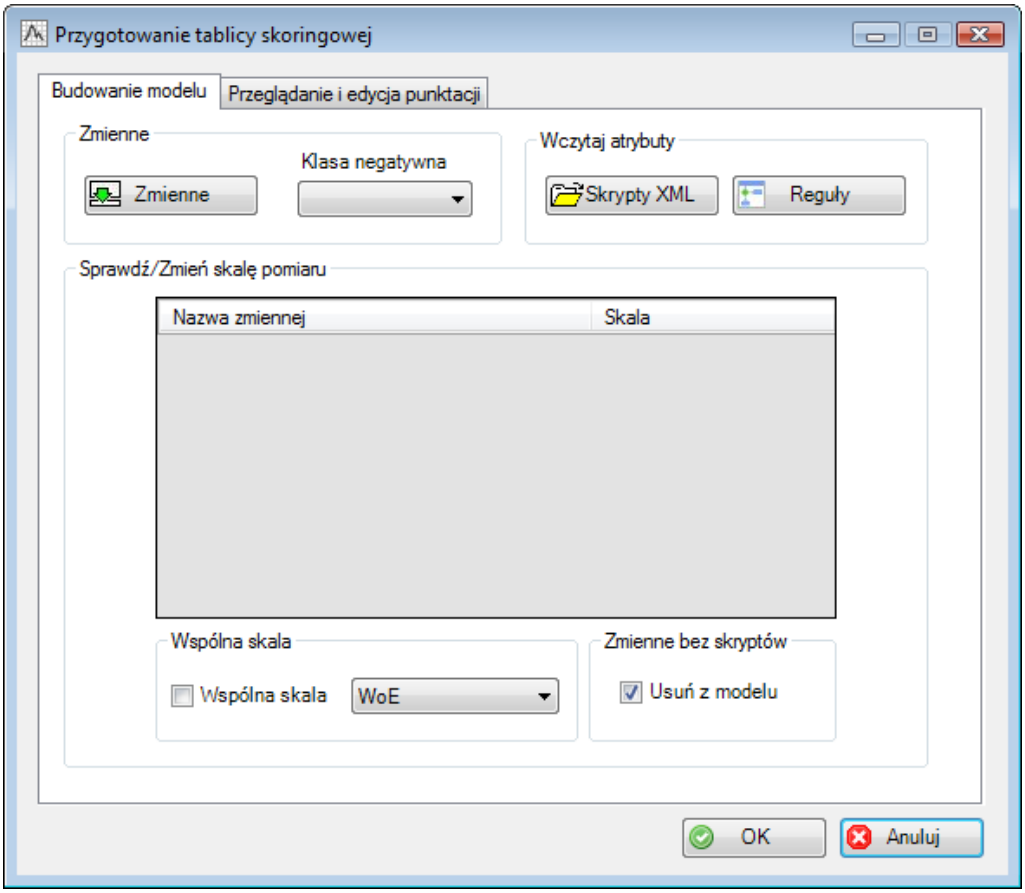
W programie wykorzystywane są dwa sposoby kodowania:

- Kodowanie typu ANOVA nazywane kodowaniem z sigma ograniczeniami (ang. terminy *effect coding* oraz *sigma restricted model*),
- Kodowanie typu WOE (*Weight of Evidence*).

W sytuacji, gdy chcemy utworzyć kartę skoringową na podstawie cech, których wartości zostały pogrupowane na klasy homogeniczne względem badanego zjawiska (co zakłada etap budowy atrybutów) zalecanym sposobem kodowania zarówno zmiennych ilościowych, jak i jakościowych są wartości *WOE* dla poszczególnych atrybutów. Kodowanie typu *WOE* nie tylko rozwiązuje problem różnych wartości skali poszczególnych zmiennych, ale również uwzględnia siłę i kierunek wpływu każdego atrybutu w tej samej skali. Jeśli poprawnie przeprowadzono proces budowy atrybutów, alokacja punktów w zbudowanym modelu odzwierciedlała będzie różnice w poszczególnych cechach uchwycone podczas wstępnej analizy.

Przykład 7. – Budowa tablicy skoringowej

Tablicę skoringową zbudujemy dla zbioru *CreditScoring.sta*, używanego w przykładach kategoryzacji. W celu zbudowania tablicy skoringowej, z menu ***Zestaw Skoringowy*** wybieramy polecenie ***Modelowanie*** | ***Budowa tablicy skoringowej***. Po wybraniu tej opcji otwarte zostanie okno ***Przygotowanie tablicy skoringowej***.



Użytkownik może parametryzować analizę za pomocą następujących opcji:

Klasa negatywna pozwala wskazać jedną z klas dychotomicznej zmiennej zależnej jako negatywną.

Wczytaj skrypty – opcje umożliwiają wczytanie definicji atrybutów zmiennych przygotowanych w module ***Dyskretyzacja zmiennych***. Za pomocą przycisku ***Skrypty XML*** wczytujemy pliki w formacie XML. Jeżeli mamy dostęp do licencji Konstruktora Reguł *STATISTICA* możemy wczytać definicje dyskretyzacji z pliku reguł.

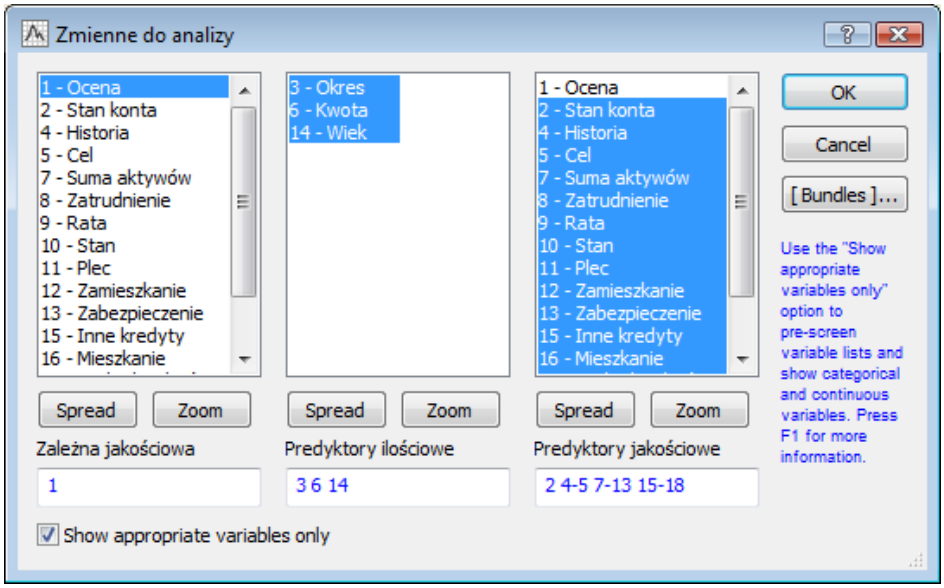
W tabeli ***Sprawdź/zmień skalę pomiaru*** wyświetlane są nazwy wybranych zmiennych oraz w kolumnie ***Skala*** opcje pozwalające określić sposób kodowania predyktorów. Zmienne, dla których nie wczytano skryptów dyskretyzacji mają dostępną jedynie opcję ***Dane surowe***. Po wczytaniu skryptów pojawiają się dodatkowe opcje: ***WoE*** umożliwiająca przekodowanie zmiennych do wartości *WoE* przypisanych do atrybutów oraz ***Jakościowa***, która powoduje przekodowanie zmiennej na sztuczne zmienne zero-jedynkowe (zgodnie ze schematem kodowania z sigma-ograniczeniami).

Obszar ***Wspólna skala*** pozwala określić wspólny sposób przekodowania predyktorów.

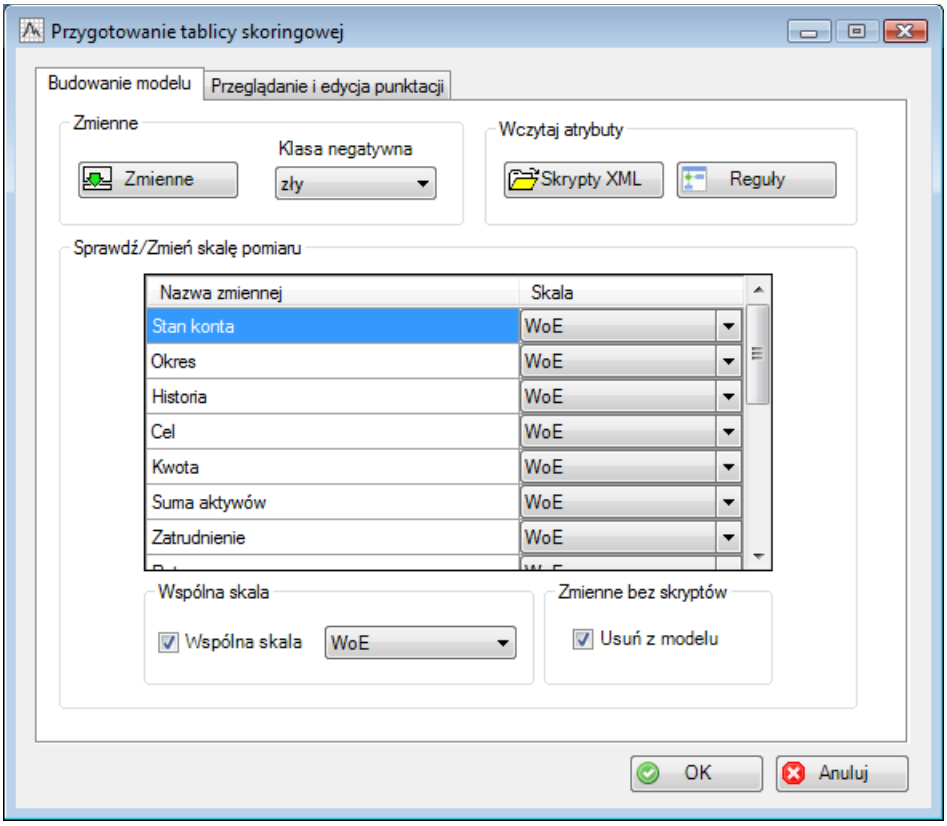
Opcja ***Usuń z modelu*** powoduje usunięcie z dalszego etapu budowania modelu predyktorów, dla których nie wczytano skryptu. Budowa modelu na danych surowych jest możliwa, jednak uwzględnienie „surowej” zmiennej w modelu uniemożliwi przeskalowanie modelu do karty skoringowej.

Na karcie ***Przeglądanie i edycja punktacji*** mamy możliwość wczytania gotowej karty skoringowej zapisanej wcześniej w formacie XML, a następnie przejścia do okna, w którym możemy przejrzeć postać modelu lub dokonać ewentualnych eksperckich korekt.

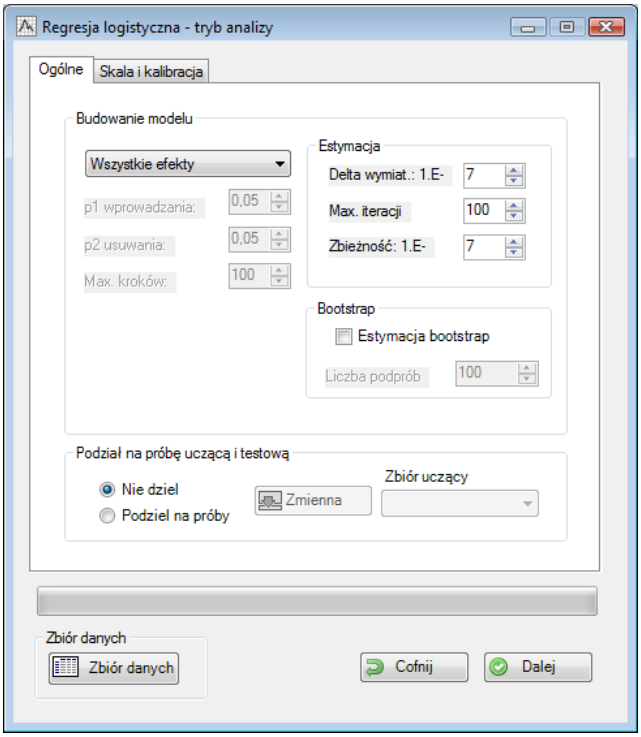
Proces budowy tablicy skoringowej rozpoczynamy od wybrania zmiennych, na podstawie których zbudowany zostanie model logitowy (zmienna *Ocena* będzie zmienną informującą o klasie kredytu, pozostałe zmienne wybieramy jako predyktory).



Następnie należy wczytać skrypty kategoryzacji/dyskretyzacji zapisane w formacie *XML*, przygotowane wcześniej w module ***Dyskretyzacja zmiennych*** (przykładowe skrypty znajdują się także w katalogu z plikami przykładowymi). Po wczytaniu skryptów zaznaczamy opcję ***Wspólna skala***, a następnie wybieramy opcję WoE co spowoduje, że przed zbudowaniem modelu zmienne zostaną przekodowane zgodnie z tym schematem. Aby zapewnić, że model będzie budowany jedynie dla tych cech, dla których wczytano skrypt *XML*, upewniamy się, że włączona jest opcja ***Usuń z modelu***. Przed wykonaniem tego ćwiczenia należy zatem przygotować skrypty kategoryzacji dla wszystkich zmiennych, które chcemy uwzględnić w modelu.



Zatwierdzamy analizę klikając przycisk ***OK***. Przywołane zostanie okno ***Regresja logistyczna – tryb analizy***.



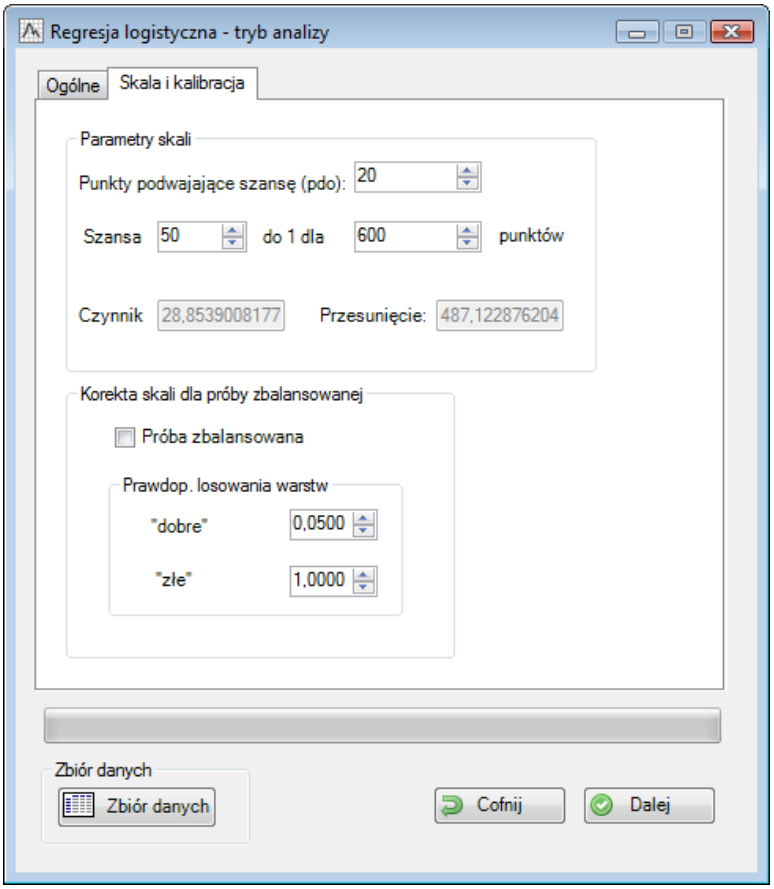
W oknie tym, na karcie ***Ogólne*** dostępnych jest szereg opcji związanych z estymacją modelu oraz wyborem scenariusza doboru zmiennych do modelu. Szczegółowy ich opis dostępny jest w pomocy dla modułu GLZ programu *STATISTICA*. Pozostałe opcje opisano poniżej.

Uwaga *– w STATISTICA od wersji 12 dostępny jest Interakcyjny Konstruktor Modeli (dostępny w menu Statystyka | Modele zaawansowane), który umożliwia użytkownikowi przeprowadzenie procesu doboru zmiennych do modelu w sposób interakcyjny – tu użytkownik decyduje o fakcie i kolejności wprowadzenia bądź odrzucenia zmiennych z modelu. Po każdym kroku obliczane są odpowiednie krańcowe statystyki dla pozostałych zmiennych „kandydujących". Więcej informacji zobacz:* [STATISTICA Interakcyjny Konstruktor Modeli](#).

Opcja ***Estymacja bootstrap*** umożliwia zbudowanie modelu regresji metodą wielokrotnego próbkowania. W zależności od wyboru opcji na liście rozwijalnej ***Budowanie modelu*** program wykona odmienne obliczenia. W przypadku wybrania opcji ***Wszystkie efekty*** jako sposobu budowy modelu, dla podanej ***Liczby podprób*** generowane są modele regresji, a następnie parametry modeli są uśredniane. Strategia ta daje często dobre efekty w sytuacji niewielkiej liczności zbioru danych. Jeżeli natomiast wybrano jedną z pozostałych strategii modelowania, ***Estymacja bootstrap*** pozwala ocenić wiarygodność procesu krokowego doboru zmiennych do modelu. W wyniku tej analizy badacz ma możliwość prześledzenia jak często poszczególne zmienne trafiały do modeli budowanych krokowo dla kolejnych prób *bootstrapowych*.

W obszarze ***Podział na próbę uczącą i testową*** mamy możliwość wskazania zmiennej, której wartości będą identyfikatorami próby uczącej i testowej. Aby mieć możliwość wskazania zmiennej z informacją o próbie włączamy opcję ***Podziel na próby***.

Na karcie ***Skala i kalibracja*** mamy możliwość dodatkowej parametryzacji zbudowanego modelu.



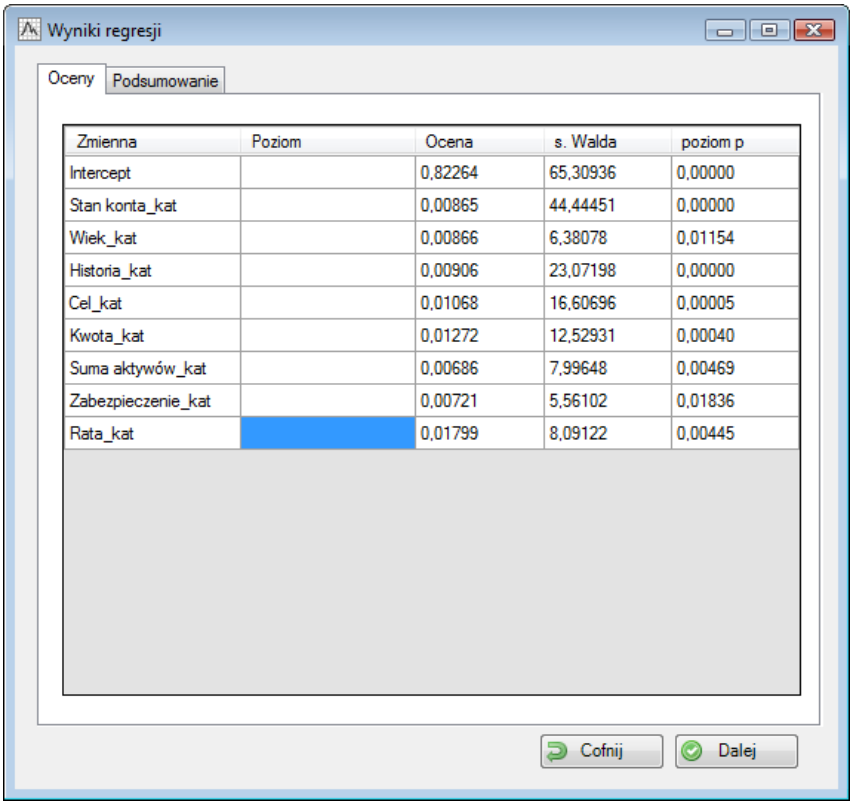
Obszar ***Parametry skali*** określa parametry skalujące model regresji logistycznej do postaci karty skoringowej. Pole ***Punkty podwajające szansę (pdo)*** określa jak szybko wraz ze wzrostem skoringu powinna rosnąć szansa bycia *dobrym*. W przypadku domyślnej wartości (20), szansa bycia dobrym klientem będzie się podwajała co 20 punktów. Kolejne opcje informują o poziomie odniesienia dla

interpretacji punktacji. ***Szansa*** 50 ***do 1 dla*** 600 ***punktów*** informuje, że klient, który uzyskał 600 punktów ma szansę 50 do 1 na bycie „dobrym klientem". Parametry ***Czynnik*** oraz ***Przesunięcie*** są automatycznie obliczane na podstawie powyższych parametrów. Szczegóły obliczeniowe patrz Scorecard Formula Guide.

Częstą praktyką analityków jest budowa modeli skoringowych dla zbioru zbalansowanego, w którym ilość złych i dobrych przypadków jest na zbliżonym poziomie. Budowa tego typu modelu daje jednak nieprawidłowe oszacowania poziomu ryzyka wynikające z nierzeczywistych proporcji „dobrych" i „złych" w próbie. Jeżeli model jest budowany na zbiorze zbalansowanym, w celu korekty poziomu ryzyka należy w obszarze ***Korekta skali dla próby zbalansowanej*** włączyć opcję ***Próba zbalansowana***. W opcji ***„dobre"*** podajemy jaką frakcję „*dobrych*" z próby oryginalnej wylosowaliśmy do próby uczącej. Analogicznie w opcji ***„złe"*** podajemy, jaką frakcję „*złych*" z próby oryginalnej wylosowaliśmy do próby uczącej.

Aby wykonać kolejne kroki analizy, w obszarze ***Budowanie modelu*** wybieramy opcję ***Krokowa wsteczna***. W obszarze ***Podział na próbę uczącą i testową*** wybieramy opcję ***Podziel na próby*** a następnie po kliknięciu przycisku ***Zmienna*** wskazujemy zmienną ***Próba***, jako zmienną identyfikującą próby. Na liście ***Zbiór uczący*** wskazujemy klasę ***Uczenie***. Przypadki należące do tej klasy posłużą jako zbiór uczący do analizy.

Po zbudowaniu modelu wyświetlone zostanie okno ***Wyniki regresji***, składające się z dwóch kart.



Na karcie ***Oceny*** można dokonać modyfikacji parametrów modelu, edytując wartości uzyskane w kolumnie ***Ocena***. Zauważmy, że wszystkie wartości ocen parametrów regresji są dodatnie.

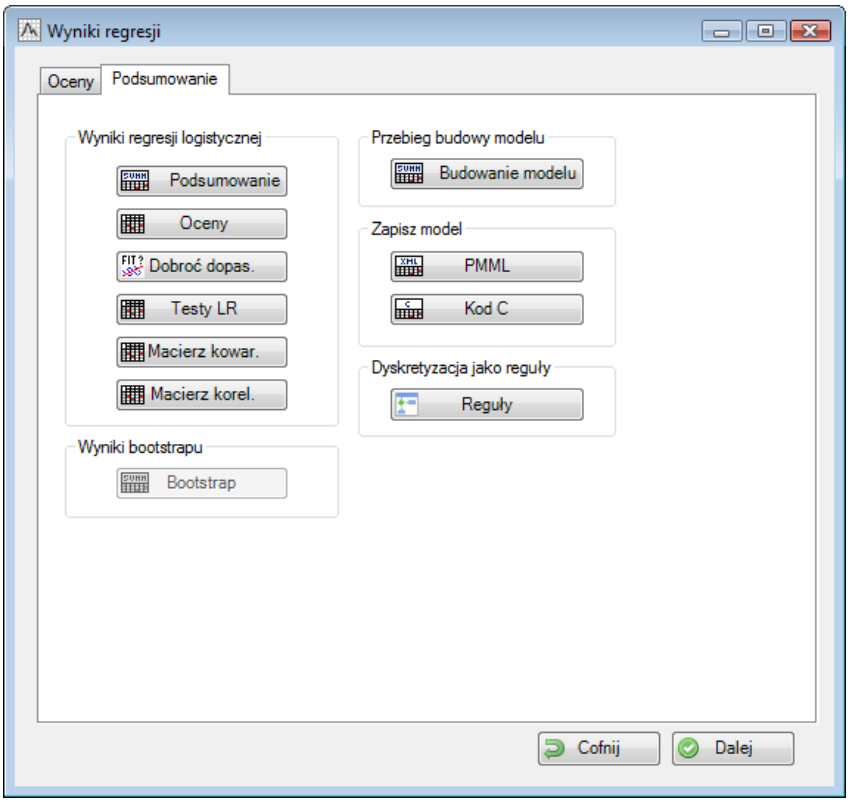
W sytuacji, gdy wybraliśmy kodowanie WoE, a niektóre oceny parametrów (nie licząc wyrazu wolnego) są ujemne, świadczy to o błędnej specyfikacji modelu wynikającej z nadmiernej korelacji pomiędzy predyktorami. Zaleca się w wtedy przegląd zmiennych wejściowych i eliminację nadmiarowych predyktorów.

Na karcie ***Podsumowanie*** znajduje się szereg wskaźników związanych z regresją logistyczną:

- Oceny parametrów

- Dobroć dopasowania
  - AIC
  - BIC
  - Miary pseudo R^2
- Testy ilorazu wiarygodności
- Macierze wariancji i kowariancji parametrów regresji oraz macierz korelacji

Użytkownik ma także możliwość zapisania modelu w formie pliku PMML lub kodu C.



W obszarze ***Wyniki bootstrapu*** mamy możliwość wygenerowania raportu *bootstrap*. Przycisk ten jest aktywny, jeżeli w oknie ***Tryb analizy*** zaznaczono opcję ***bootstrap***.

Jeżeli przed zbudowaniem modelu w obszarze ***Budowanie modelu*** została wybrana opcja ***Wszystkie efekty*** kliknięcie przycisku ***Bootstrap*** wyświetli arkusz zawierający wartości parametrów wszystkich zbudowanych modeli, ponadto wyświetlony zostanie arkusz z uśrednionymi ocenami parametrów regresji oraz bootstrapowymi przedziałami ufności. Uzupełnieniem wyników są wykresy prezentujące rozkłady wartości parametrów regresji.

Jeżeli przed zbudowaniem modelu w obszarze ***Budowanie modelu*** została wybrana opcja ***Krokowa postępująca***, ***Krokowa wsteczna***, ***Wprowadzanie postępujące***, ***Eliminacja wsteczna***, po naciśnięciu przycisku ***Bootstrap*** wyświetlony zostanie arkusz i wykres z informacjami na temat liczby i odsetka modeli bootstrapowych, w których znalazły się poszczególne predyktory. Dzięki temu możemy zwalidować proces doboru zmiennych do modelu i skorygować jego wyniki.

Przycisk ***Budowanie modelu*** wyświetla raport z przebiegu doboru zmiennych w regresji krokowej. W wynikach zostały opisane wszystkie kroki regresji i proces przejścia od początkowego zestawu predyktorów do finalnego modelu.

Dane: Ocena - Wyniki budowy modelu (Arkusz181)*

Ocena - Wyniki budowy modelu (Arkusz181)
Rozkład: DWUMIANOWY , F. wiążąca: LOGIT
Modelowane prawdopodobieństwo Ocena = dobry

Krok	Efekt	Stopnie Swobody	Walda Stat.	Walda p	S.pkt Stat.	S.pkt p	War. Stan
Krok 7	Cel_kat	1	16,60714	0,000046			W modelu
	Historia_kat	1	23,07217	0,000002			W modelu
	Wiek_kat	1	6,38084	0,011536			W modelu
	Kwota_kat	1	12,52941	0,000401			W modelu
	Zabezpieczenie_kat	1	5,56106	0,018364			W modelu
	Suma aktywów_kat	1	7,99658	0,004687			W modelu
	Stan konta_kat	1	44,44502	0,000000			W modelu
	Rata_kat	1	8,09129	0,004448			W modelu
	Okres_kat	1			2,882026	0,089573	Poza
	Zatrudnienie_kat	1			2,123377	0,145066	Poza
	Stan_kat	1			2,184470	0,139409	Poza
	Inne kredyty_kat	1			1,753993	0,185376	Poza
	Mieszkanie_kat	1			0,880456	0,348077	Poza
	Płeć_kat	1			0,842362	0,358722	Poza

W celu przygotowania tablicy skoringowej klikamy przycisk ***Dalej***, wyświetlone zostanie okno ***Zestaw skoringowy***. W oknie tym znajduje się tabela ze wszystkimi poziomami (atrybutami) każdej zmiennej, która weszła do modelu oraz z przypisanymi wartościami skoringu. W tabeli znajduje się również informacja o wartości wskaźnika *WoE*, ocenie parametru regresji oraz, jeżeli to możliwe, statystyce Walda oraz wartości *p*. Dodatkowo dla każdej cechy wyliczana jest ***Wartość neutralna*** (ważony poziom skoringu) przydatna, gdy wartość badanej cechy jest nieznana lub spoza zakresu.

Zestaw skoringowy

Zmienna	Zakres	WoE	Ocena	s. Walda	poziom p	Skoring	Skoring zaokr.
Kwota	(6630,4;inf)	-53,318	0,01272	12,52931	0,00040	44,288	44
Kwota	Wartość ne...	-	-			64,421	64
Suma akty...	140-700	76,214	0,00686	7,99648	0,00469	78,943	79
Suma akty...	700-1400	76,214	0,00686	7,99648	0,00469	78,943	79
Suma akty...	>1400	76,214	0,00686	7,99648	0,00469	78,943	79
Suma akty...	brak	-25,245	0,00686	7,99648	0,00469	58,860	59
Suma akty...	<140	-25,245	0,00686	7,99648	0,00469	58,860	59
Suma akty...	Wartość ne...	-	-			64,765	65
Rata	> 35	25,131	0,01799	8,09122	0,00445	76,903	77
Rata	25-35	15,547	0,01799	8,09122	0,00445	71,927	72
Rata	15- 25	6,454	0,01799	8,09122	0,00445	67,207	67
Rata	< 15	-15,730	0,01799	8,09122	0,00445	55,692	56
Rata	Wartość ne...	-	-			63,500	64
Zabezpiecz...	brak	46,103	0,00721	5,56102	0,01836	73,449	73
Zabezpiecz...	nieruchomo...	-58,608	0,00721	5,56102	0,01836	51,665	52
Zabezpiecz...	samochód	-3,188	0,00721	5,56102	0,01836	63,194	63
Zabezpiecz...	polisa	-3,188	0,00721	5,56102	0,01836	63,194	63
Zabezpiecz...	Wartość ne...	-	-			64,960	65
Wiek	(-inf;24>	-41,386	0,00866	6,38078	0,01154	53,516	54
Wiek	(24;31>	-7,676	0,00866	6,38078	0,01154	61,939	62
Wiek	(31;50>	35,667	0,00866	6,38078	0,01154	72,770	73
Wiek	(50;inf)	-0,990	0,00866	6,38078	0,01154	63,610	64
Wiek	Wartość ne...	-	-			63,674	64

Zamknij
Cofnij
Wartości skoringu można edytować
Raport
Raport
Zapisz do Excela
Zapisz
Reguły
Reguły
Skrypt:
Skrypt:
Skrypt XML

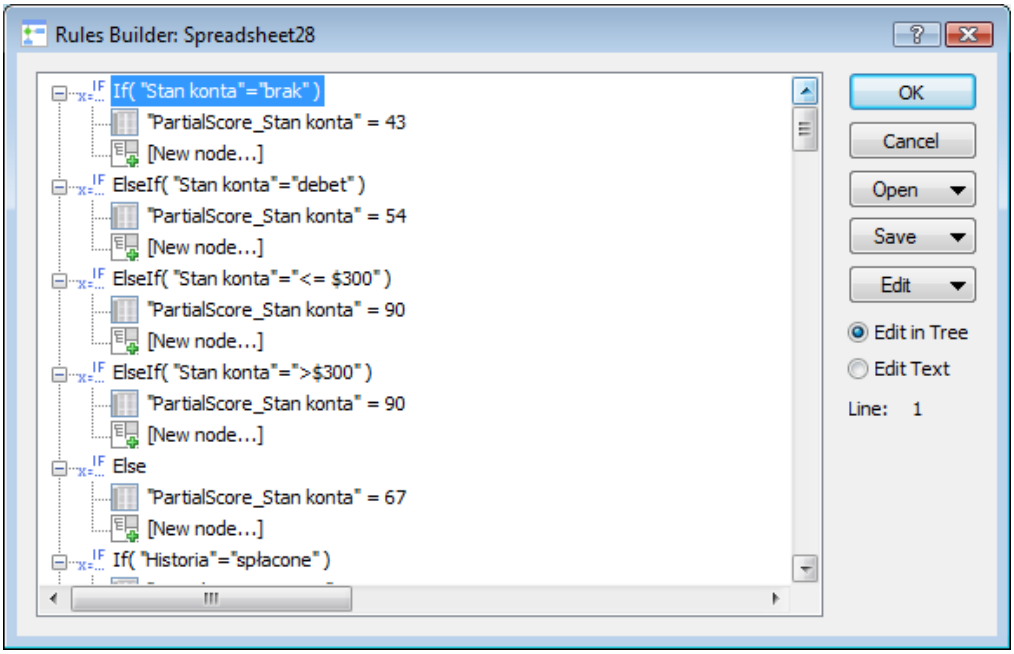
Klikając na odpowiedniej komórce w kolumnie ***Skoring*** możemy w sposób ekspercki zmienić wartości uzyskanej punktacji.

Przygotowaną tablicę skoringową zapisujemy do arkusza *STATISTICA* klikając przycisk ***Raport***.

Uwaga. *Po zbudowaniu tablicy skoringowej można wrócić do początkowej fazy analizy w celu dokonania zmian na przykład w dyskretyzacji parametrów czy sposobie doboru zmiennych do modelu. Następnie tak przygotowane tablice skoringowe można ze sobą porównywać w module Ocena i kalibracja.*

Za pomocą przycisku ***Zapisz*** z obszaru ***Zapisz do Excela*** zapisujemy stan wyświetlonej karty skoringowej do pliku MS Excel.

Przycisk ***Reguły*** spowoduje wyświetlenie utworzonego modelu w specjalnym oknie reguł (jeżeli użytkownik posiada odpowiednią licencję), gdzie możliwa będzie ich dodatkowa edycja, uruchomienie oraz zapis w bazie metadanych. Więcej informacji zobacz *STATISTICA Konstruktor Reguł*.



Wybrane reguły mogą zostać wdrożone do bazy systemu *STATISTICA Enterprise* i używane przez jego użytkowników podobnie jak inne obiekty *STATISTICA Enterprise*.

Przycisk ***Skrypt*** pozwala zapisać zbudowaną kartę skoringową do pliku *XML*, dzięki czemu możemy jej użyć w kolejnych modułach ***Zestawu skoringowego***. Kartę skoringową można także zapisać w postaci makra *STATISTICA* Visual Basic, (możemy uruchomić je w programie *STATISTICA* w celu obliczenia wartości skoringu dla nowych danych), a także w postaci kodu węzła programu *STATISTICA Data Miner* (możemy umieścić i uruchamiać go w graficznej przestrzeni roboczej).

4.2. Analiza wniosków odrzuconych

Budując model skoringowy jedynie na podstawie wniosków zaakceptowanych, już na wstępie zgadzamy się na pewne obciążenie całej analizy. Istnieje kilka rozwiązań tego problemu, pierwszym, najbardziej ryzykownym jest akceptacja przez pewien okres czasu wszystkich, bądź losowo wybranej części wniosków, które normalnie byśmy odrzucili, tzw. *Buying data*. Innym mniej ryzykownym sposobem jest przydzielenie wniosków odrzuconych do dwóch klas *„zły”* i *„dobry”* i dołączenie ich do procesu budowy karty skoringowej tak jak wniosków zaakceptowanych.

Moduł ***Wnioski odrzucone*** umożliwia zbudowanie tablicy skoringowej uwzględniającej całość populacji – także tych klientów, którzy zostali odrzuceni. Do wyboru są dwie metody uzupełniania brakującej informacji „*dobry/zły*” w zmiennej zależnej.

- paczkowanie (*parceling*),

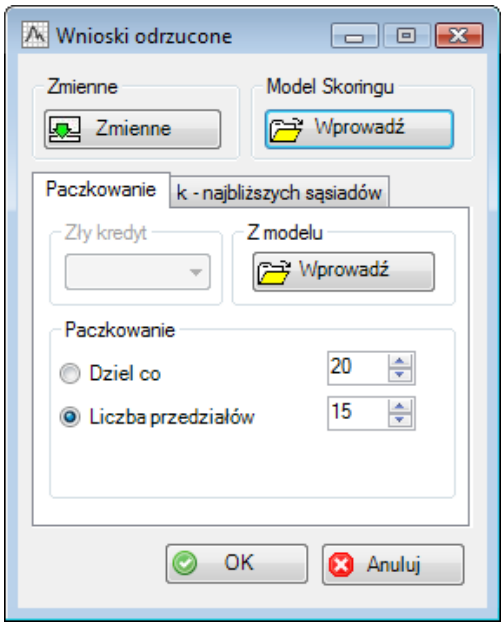
- k - najbliższych sąsiadów.

Metoda k – najbliższych sąsiadów to jeden z algorytmów używanych w statystyce do prognozowania wartości pewnej zmiennej losowej. Może również być używany do klasyfikacji. Dla każdego odrzuconego wniosku wyznaczany jest zbiór k najbliższych (najbardziej podobnych) wniosków zaakceptowanych. Wniosek odrzucony otrzymuje tą z ocen „*dobry/zły*”, która w tym zbiorze występuje częściej.

Metoda paczkowania (*parceling*) – polega na podzieleniu pewnego zakresu skoringu na przedziały, następnie oszacowaniu frakcji *dobrych/złych* kredytów w każdej klasie dla wniosków, które zostały zaakceptowane i odtworzeniu tych frakcji dla wniosków odrzuconych.

Przykład 8. – paczkowanie (*parceling*)

Moduł ***Wnioski odrzucone*** uruchamiamy poleceniem ***Modelowanie*** | ***Wnioski odrzucone*** znajdującym się w menu ***Zestaw skoringowy***. Po wyborze tej opcji, wyświetlone zostanie okno ***Analiza wniosków odrzuconych***. Przykładową analizę wykonamy dla pliku *WnioskiOdrzucone.sta* znajdującego się w plikach dołączonych do dokumentacji.



Wykonanie analizy może być parametryzowane za pomocą następujących opcji:

Przycisk ***Wprowadź*** znajdujący się w obszarze ***Model Skoringu*** umożliwia wczytanie karty skoringowej zapisanej w formacie XML. Po jej wczytaniu automatycznie uruchomiona zostanie procedura obliczenia skoringu dla wszystkich przypadków (zarówno zaakceptowanych, jak i odrzuconych) znajdujących się w zbiorze danych.

Na karcie ***Paczkowanie*** znajdują się następujące opcje:

Zły kredyt umożliwia wybranie jednej z klas zmiennej zależnej jako symbolu klasy niepożądanej.

Przycisk ***Wprowadź*** znajdujący się w obszarze ***Z modelu*** umożliwia wczytanie przygotowanego wcześniej w tym module schematu uzupełniania wniosków odrzuconych.

Opcja ***Dziel co*** podzieli analizowany zbiór danych na przedziały, z których każdy będzie obejmował liczbę punktów zawartą w odpowiadającym jej polu.

Liczba przedziałów określa, na jaką liczbę przedziałów ma zostać podzielony zbiór danych.

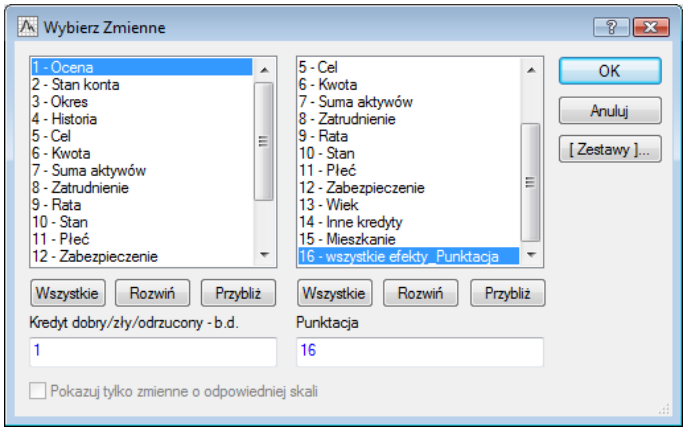
Na karcie ***k-najbliższych sąsiadów*** mamy do dyspozycji następujące opcje:

Liczba sąsiadów określa, na podstawie jakiej liczby przypadków o najbliższym skoringu będzie określana klasa *dobry/zły*.

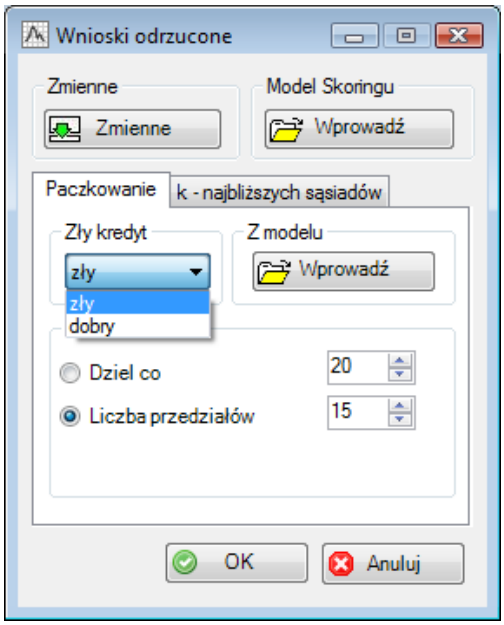
Liczba wzorców określa liczbę przypadków zaakceptowanych, jakie będą brane pod uwagę w określaniu sąsiadów.

Ponieważ w arkuszu danych nie ma kolumny z wyliczonym skoringiem, należy wczytać zbudowany wcześniej model korzystając z przycisku ***Wczytaj*** a następnie wskazując plik *WszystkieEfekty.xml*. Po wczytaniu pliku z modelem zostanie on uruchomiony na bieżącym zbiorze danych a efekt jego działania zostanie zapisany w postaci skoringu w dodatkowej zmiennej.

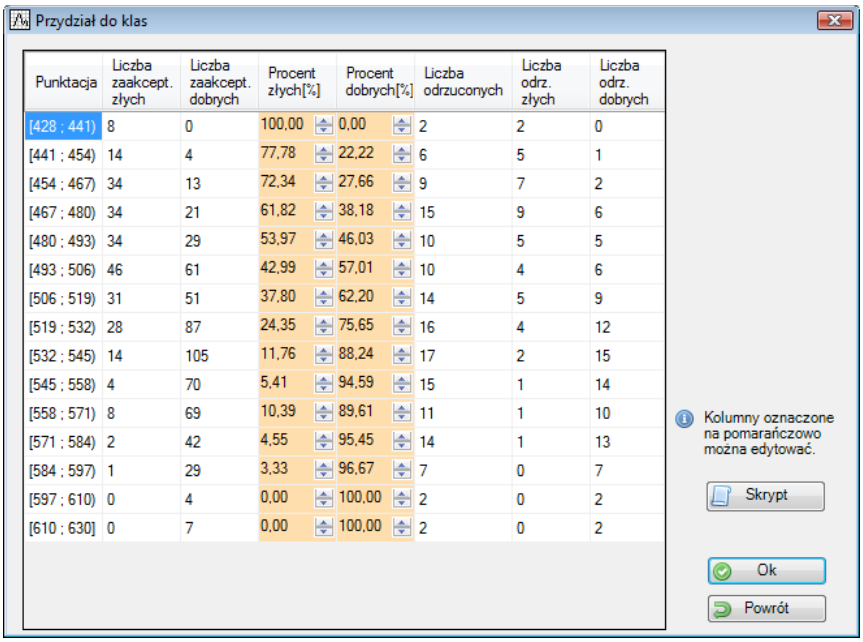
Kolejnym krokiem jest wybranie zmiennych do analizy, należy wskazać zmienną zawierającą „*dobry/zły*" kredyt, oraz obliczony skoring.



Następnie należy wskazać symbol złego kredytu.



Następnie wybieramy opcję ***Liczba przedziałów*** i pozostawiamy jej wartość na domyślnym poziomie. Po kliknięciu ***OK*** pojawi się okno umożliwiające ręczną korektę frakcji *dobrych/złych* kredytów w każdym przedziale (pomarańczowe kolumny), a także zapisania ich do pliku (przycisk ***Skrypt***) w celu późniejszego wykorzystania. Zapisany model paczkowania można wczytać korzystając z przycisku ***Wprowadź*** na zakładce ***Paczkowanie*** w oknie ***Wnioski odrzucone***.

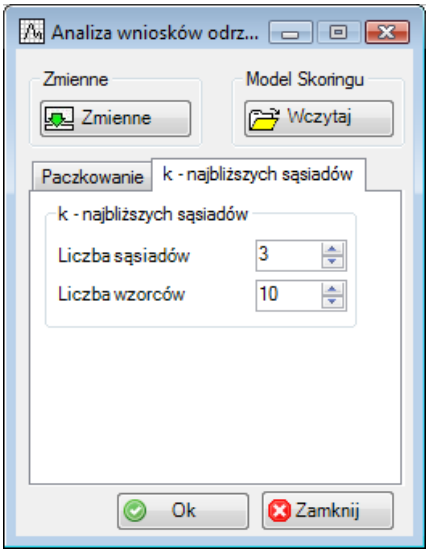


Po zakończeniu modyfikacji i kliknięciu przycisku ***OK*** utworzony zostaje raport podsumowujący oceny *„dobry/zły”* w poszczególnych klasach skoringu i w podziale na wnioski zaakceptowane i odrzucone. Raport zawiera także uzupełniony arkusz danych, który powinien zostać wykorzystany do powtórnej budowy karty skoringowej.

Przykład 9. – k – najbliższych sąsiadów

Moduł ***Wnioski odrzucone*** uruchamiamy poleceniem ***Wnioski odrzucone*** znajdującym się w menu ***Zestaw skoringowy***. Po wyborze tej opcji, wyświetlone zostanie okno ***Analiza wniosków odrzuconych***.

Jeżeli w arkuszu danych nie ma kolumny z wyliczonym skoringiem, należy wczytać zbudowany wcześniej model korzystając z przycisku ***Wczytaj***. Po przejściu na drugą zakładkę należy dokonać selekcji zmiennych i dopasować parametry metody.



Po kliknięciu przycisku ***Ok*** zostanie utworzony raport zawierający oryginalny arkusz danych, arkusz z uzupełnioną kolumną kredyt *„dobry/zły”*, oraz tabela przedstawiająca frakcje *„dobrych/złych”* kredytów dla odrzuconych i zaakceptowanych wniosków kredytowych. Na uzupełnionym arkuszu należy powtórnie przeprowadzić proces budowy tablicy skoringowej.

Dane: Raport*

Tabela liczności (Arkusz29)
Tabela: Ocena(2) x Status(2)

	Ocena	Status Odrzucony	Status Zaakceptowany	Wiersz Razem
Liczba	zły	34	258	292
% z kolumny		22,67%	30,35%	
Liczba	dobry	116	592	708
% z kolumny		77,33%	69,65%	
Liczba	Ogół	150	850	1000

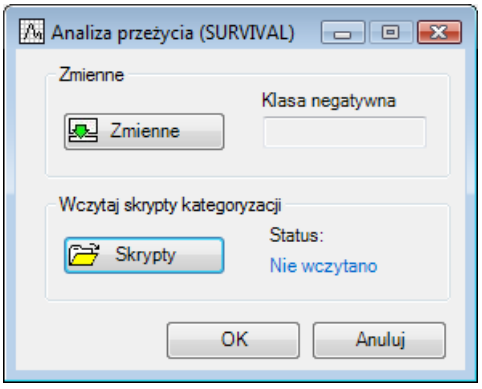
4.3. Analiza przeżycia

Moduł ***Analiza przeżycia (SURVIVAL)*** umożliwia budowę modeli skoringowych za pomocą proporcjonalnego hazardu Coxa, który poza czynnikami wpływającymi na zmienną zależną pozwala uwzględnić również czas zajścia analizowanego zdarzenia. Modele te pozwalają nie tylko określić prawdopodobieństwo zajścia danego zdarzenia, ale również czas w jakim osiągnie ono dany (krytyczny) poziom - np. kiedy dana osoba przestanie spłacać kredyt, bądź kiedy odejdzie do konkurencji.

Przykład 10. Analiza przeżycia

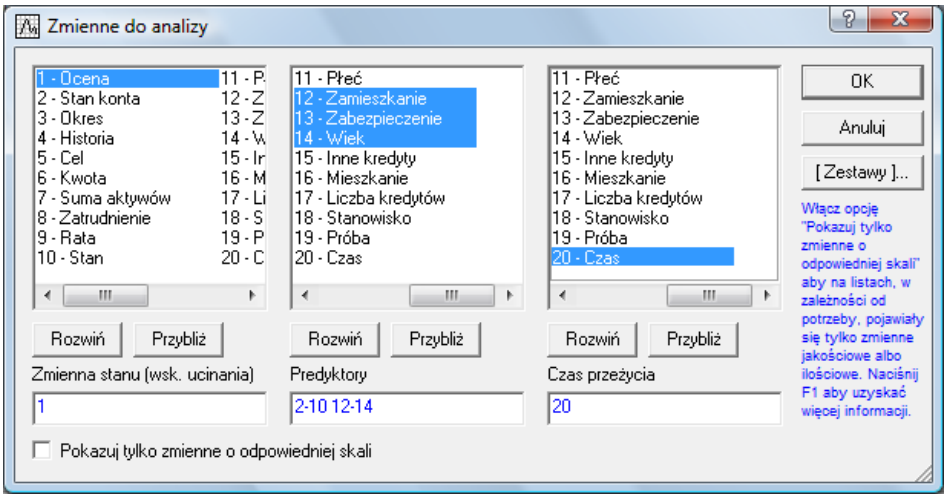
Przykład budowy modelu za pomocą modułu ***Analiza przeżycia (SURVIVAL)*** wykonamy na podstawie zbioru *TimeUczenie.sta*, który, oprócz znanych nam zmiennych, zawiera dodatkowo zmienną *Czas*. Zmienna ta informuje w jakim czasie zakończyła się obserwacja danego przypadku. Jeżeli przypadek **zły** (w analizie przeżycia odpowiada mu tzw. przypadek kompletny, czyli przypadek, w którym zaszło „zdarzenie") zawiera w zmiennej *Czas* wartość 6 oznacza to, że analizowany kredytobiorca w szóstym okresie obserwacji wykazał znaczące odstępstwo od umowy kredytowej (w przypadku analizy odejść klientów oznaczałoby to, że w tym czasie zrezygnował z umowy; w przypadku skoringu medycznego oznaczałoby to okres, w którym nastąpiło „zdarzenie" - np. zgon pacjenta). Jeżeli przypadek **dobry** (przypadek dla którego nie zaobserwowano „zdarzenia" - w analizie przeżycia mówimy, że jest to przypadek ucięty) zawiera w zmiennej *Czas* wartość 12 oznacza to, że w całym okresie obserwacji (przyjmijmy, że będzie to 12 miesięcy) nie wystąpiło znaczące odstępstwo od umowy kredytowej. W sytuacji, gdy przypadek **dobry** w zmiennej *Czas* zawiera wartość mniejszą niż 12, może to oznaczać na przykład wcześniejsze zakończenie spłaty, bądź przeniesienie kredytu do innego banku.

Analizę danych rozpoczniemy od wybrania z ***Zestawu skoringowego*** modułu ***Modelowanie*** | ***Analiza przeżycia*** **(SURVIVAL)**.

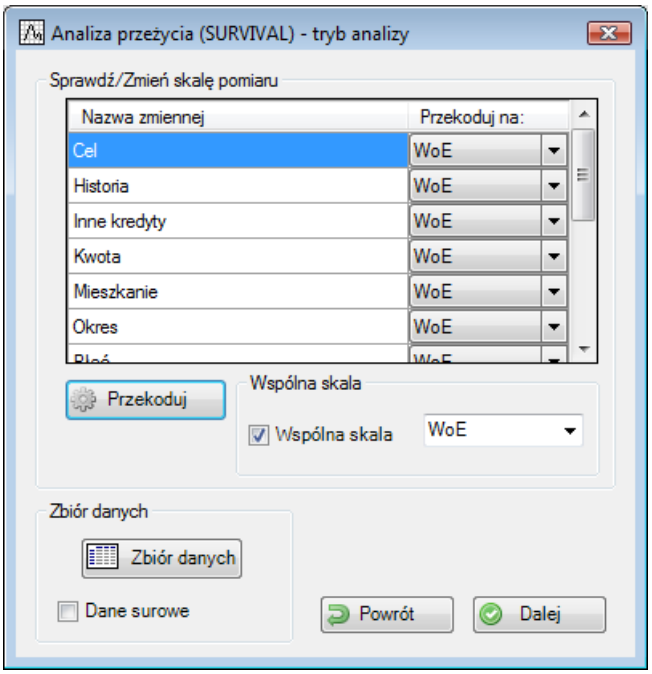


Następnie w oknie o tej samej nazwie wskazujemy zmienne do analizy w sposób analogiczny jak w modelu logistycznym (wybieramy jedynie te predyktory, które okazały się istotne w budowie

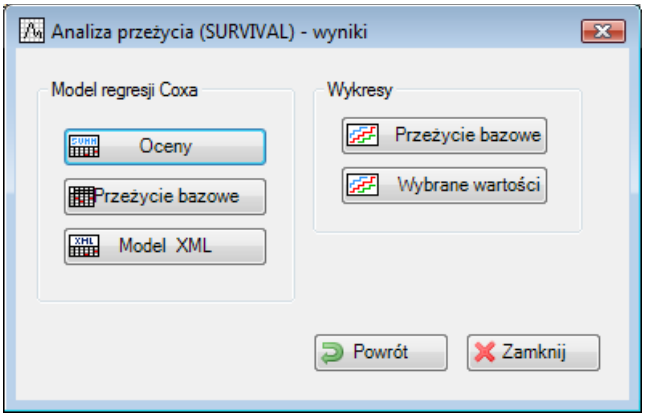
modelu logistycznego). *Zmienną stanu* (wskaźnikiem ucinania w analizie przeżycia) będzie zmienna *Ocena*. Dodatkowo w oknie *Czas przeżycia* wybieramy zmienną *Czas*.



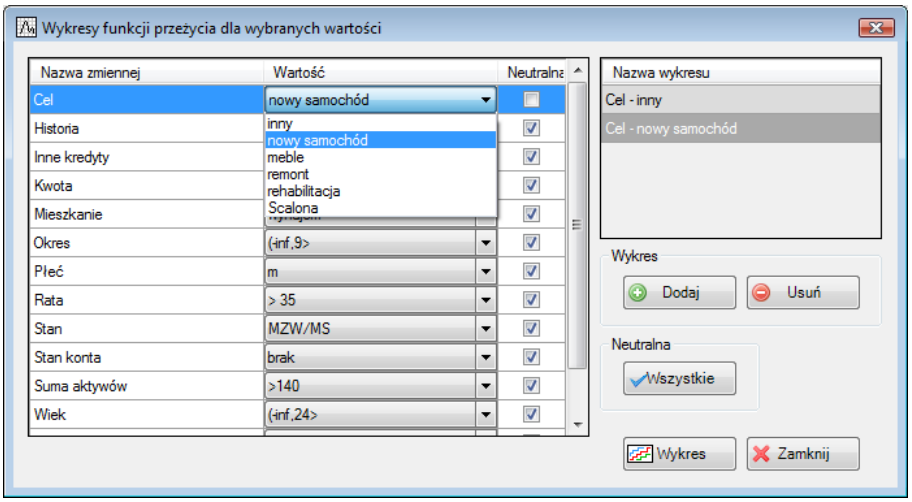
Po wybraniu zmiennych wskazujemy symbol wartości negatywnej - **zły** (odpowiadający obserwacjom kompletnym - w których zaszło „zdarzenie"), a następnie wczytujemy skrypty dyskretyzacji zapisane w plikach *XML*. Zatwierdzamy wykonanie analizy przechodząc tym samym do okna ***Analiza przeżycia (SURVIVAL) – tryb analizy***.



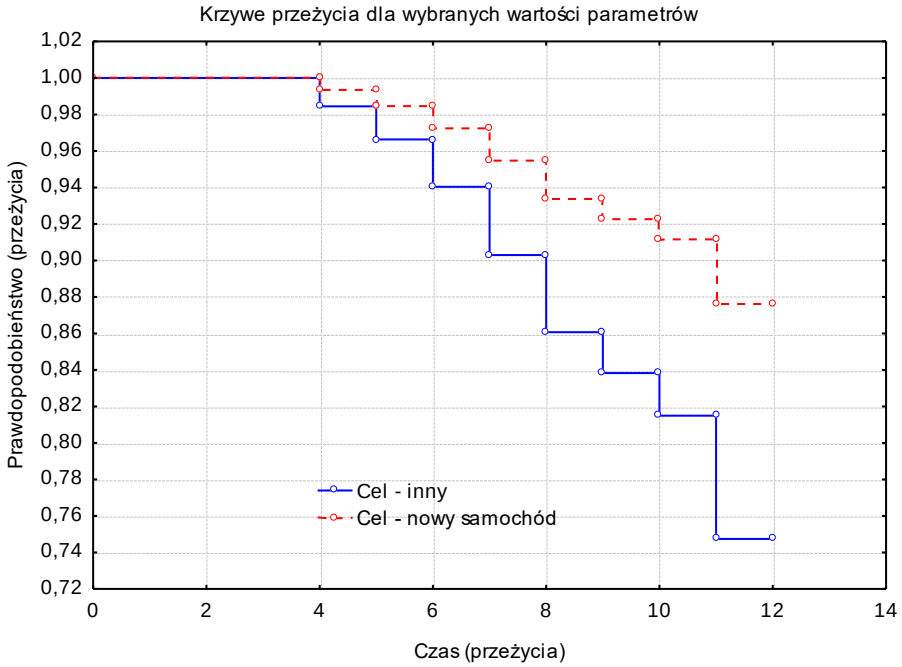
W oknie tym określamy wspólną skalę predyktorów jako: *WoE* i po kliknięciu przycisku ***Przekoduj*** uruchamiamy metodę regresji proporcjonalnego hazardu Coxa bez zmiennych zależnych od czasu, za pomocą przycisku *Dalej*. Po oszacowaniu parametrów modelu wyświetlone zostanie okno ***Analiza przeżycia (SURVIVAL) – wyniki***, w którym będziemy mogli ocenić parametry modelu (przycisk ***Oceny***) i utworzyć za pomocą przycisku ***Przeżycie bazowe*** wykres przeżycia bazowego (przeżycia w sytuacji, gdy wartości wszystkich predyktorów są równe 0).



Dodatkową opcją jest możliwość wykreślenia krzywych przeżycia dla wybranych wartości zmiennych uwzględnionych w modelu. Po kliknięciu przycisku ***Wybrane wartości*** w oknie ***Wykresy funkcji przeżycia dla wybranych wartości***, możemy wybrać dowolną kategorię cech, jakie weszły do modelu, a następnie klikając ***Dodaj***, dodać wybrany układ do zestawu wykresów, jakie zostaną wygenerowane.



W naszym przypadku klikamy najpierw przycisk ***Wszystkie*** w obszarze ***Neutralna***, co spowoduje zaznaczenie opcji we wszystkich wierszach w kolumnie ***Neutralna***. Następnie anulujemy zaznaczenie tej opcji w pierwszym wierszu odpowiadającym zmiennej *Cel*. Na liście rozwijalnej wybieramy *inny* i klikamy ***Dodaj***, a następnie w tabeli ***Nazwa wykresu*** ustawiamy nazwę wykresu, która będzie wyświetlana na wykresie krzywych przeżycia - np. *Cel – inny*. Podobną operację wykonujemy dla kategorii *nowy samochód* a następnie klikamy przycisk ***Wykres*** uzyskując wykres krzywych przeżycia dla wybranych poziomów zmiennych.



Dzięki tej opcji mamy możliwość porównania prawdopodobieństwa przeżycia w różnych okresach czasu (czyli tego, że do określonego momentu nie wystąpiło znaczące odstępstwo od umowy kredytowej), dla wybranych poziomów zmiennych.

Na koniec, za pomocą przycisku ***Zamknij***, wracamy do okna wyników analizy, w obszarze ***Model regresji Coxa*** klikamy przycisk ***Model XML*** i zapisujemy uzyskany model do pliku *XML*. Będziemy mogli go następnie użyć w module ***Obliczanie skoringu***, aby obliczyć wartość prawdopodobieństwa modelowanych zdarzeń w określonym czasie. Wyliczone prawdopodobieństwa będą mogły nam następnie posłużyć jako podstawa do obliczeń jakości modelu oraz obliczenia optymalnego punktu odcięcia.

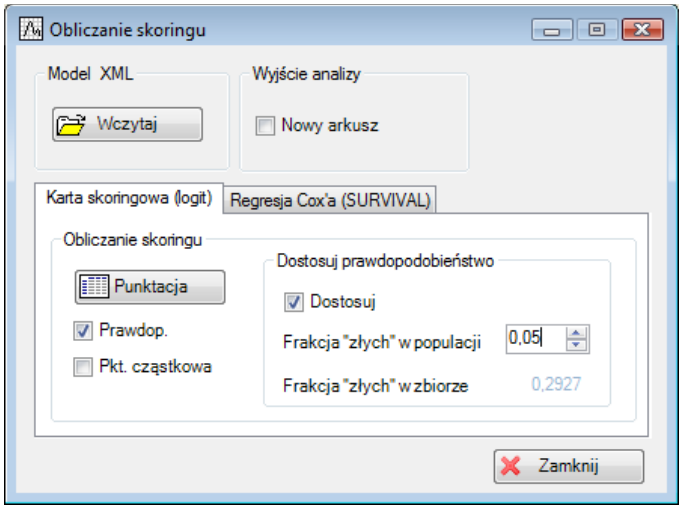
5. Ocena i kalibracja

5.1. Obliczanie skoringu

Moduł ***Obliczanie skoringu*** służy do wdrażania zbudowanych modeli dla nowych danych. Model skoringowy można zbudować korzystając z dwóch modułów: ***Budowa tablicy skoringowej*** i ***Analiza przeżycia (SURVIVAL)***. W zależności od tego, który model został zbudowany, do dyspozycji mamy odpowiednią zakładkę: ***Karta skoringowa (logit)***, jest aktywna, jeśli wczytano kartę skoringową opartą na regresji logistycznej ***Regresja Cox'a (SURVIVAL)***, jeżeli wczytano model SURVIVAL. Po wczytaniu pliku *XML* zawierającego model, moduł automatycznie przełączy się na odpowiednią zakładkę.

Przykład 11. Obliczanie skoringu - logit

Do wyliczenia skoringu wybierzemy model zbudowany wcześniej w module ***Budowa tablicy skoringowej***. Po wybraniu polecenia ***Ocena i kalibracja*** | ***Obliczanie skoringu*** wyświetlone zostanie okno ***Obliczanie skoringu***



Dla modelu budowanego regresją logistyczną moduł pozwala na obliczenie wartości skoringu wraz ze skoringami cząstkowymi dla każdej z cech. Jeśli zaznaczona jest opcja ***Prawdop.*** do raportu dodatkowo dołączone zostaną wartości prawdopodobieństwa generowane przez model, wynikające z modelu logitowego, ewentualnie (opcja ***Dostosuj***) skorygowane o wartość prawdopodobieństwa *a priori* (podawanego przez użytkownika) niespłacenia kredytu w badanej populacji. Wartość tą podajemy w obszarze ***Frakcja „złych" w populacji***.

Klikamy przycisk ***Wczytaj*** w obszarze ***Model XML*** i w wyświetlonym oknie wskazujemy plik *wszystkie efekty.xml* z zapisaną kartą skoringową. Po wczytaniu pliku z modelem, moduł automatycznie wybierze odpowiednią zakładkę: ***Karta skoringowa (logit)***.

Możemy przeprowadzić obliczanie skoringu za pomocą następujących opcji:

Nowy arkusz pozwala na utworzenie kopii wejściowego arkusza danych wraz z nowymi zmiennymi z obliczonym skoringiem i prawdopodobieństwem. Jeśli ta opcja nie jest włączona, zmienne z obliczonym skoringiem zostaną dodane do wejściowego zbioru danych.

Przycisk ***Punktacja*** oblicza skoring na podstawie wczytanego modelu oraz wybranych opcji.

Prawdop. umożliwia obliczenie prawdopodobieństwa bycia „*dobrym*" i „*złym*" klientem. Opcja ta aktywuje opcje zgrupowane w obszarze ***Dostosuj prawdopodobieństwo***.

Opcja ***Pkt. cząstkowa*** pozwala na obliczenie skoringów cząstkowych dla poszczególnych zamiennych zawartych w modelu.

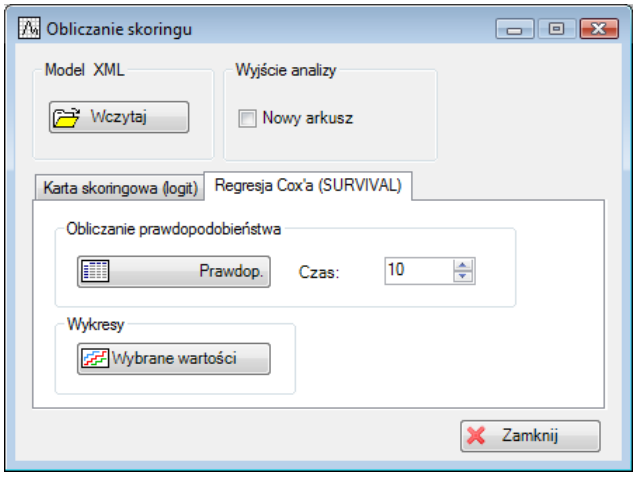
Opcja ***Dostosuj*** pozwala na korektę wartości otrzymanego prawdopodobieństwa wynikającą ze zmiany frakcji „*złych*" klientów w populacji przychodzącej. Szczegóły obliczeniowe, patrz: Scorecard Formula Guide.

> ***Frakcja „złych" w populacji*** umożliwia określenie spodziewanego prawdopodobieństwa bycia „złym" klientem w aktualnej populacji.
>
> ***Frakcja „złych" w zbiorze*** wyświetla prawdopodobieństwo bycia „złym" klientem zaobserwowane w zbiorze uczącym i zapisane w pliku XML z modelem.

Przed obliczeniem skoringu klikamy opcję ***Nowy arkusz***, aby punktacja została obliczona do nowego arkusza danych. Ponieważ chcemy obliczyć dodatkowo wartość prawdopodobieństwa odstępstwa od umowy kredytowej zaznaczamy opcję ***Prawdop.***, a w obszarze ***Dostosuj prawdopodobieństwo*** określamy poziom złych kredytów w populacji na 0,05. Następnie klikamy przycisk ***Punktacja***, *co* spowoduje obliczenie skoringu oraz prawdopodobieństwa dla nowego pliku danych.

Przykład 12. Obliczanie skoringu – regresja Coxa

Do wyliczenia skoringu wybierzemy tym razem model zbudowany za pomocą modelu *SURVIVAL*. Po wybraniu polecenia ***Ocena i kalibracja | Obliczanie skoringu*** wyświetlone zostanie okno ***Obliczanie skoringu***. W oknie tym klikamy przycisk ***Wczytaj*** i wskazujemy plik z zapisanym modelem (*cox.xml*). Po wczytaniu pliku z modelem, moduł automatycznie wybierze odpowiednią zakładkę: ***Regresja Cox'a (SURVIVAL)***.



Przycisk ***Prawdop.***, powoduje obliczenie prawdopodobieństwa przeżycia (czyli na przykład poprawnej spłaty) do zadanego czasu określanego w opcji ***Czas***. W zależności od opcji ***Nowy arkusz*** wynik zostanie dodany do wejściowego zbioru danych lub utworzony zostanie nowy zbiór.

Czas pozwala na określenie, dla jakiego czasu licząc od momentu przyznania kredytu zostanie obliczone prawdopodobieństwo.

Przycisk ***Wybrane wartości*** działa analogicznie do opcji opisanej w module do budowy modeli *SURVIVAL*.

Aby obliczyć prawdopodobieństwo, w opcji ***Czas*** zaznaczamy wartość 10, a następnie klikamy przycisk ***Prawdop.***

5.2. Ocena modeli

Podczas procesu tworzenia karty skoringowej zwykle tworzonych jest kilka modeli wykorzystujących różne metody analizy czy sposoby dyskretyzacji. Na pewnym etapie należy jednak wybrać najlepszy z nich. W *Zestawie Skoringowym* do tego celu służy moduł ***Ocena modeli***. Umożliwia on porównanie zbudowanych modeli skoringowych wykorzystując szereg miar, do których należą:

- Wskaźnik *IV* (*Information Value*),
- *KS* – współczynnik Kołmogorowa-Smirnowa,
- Wskaźnik *Gini*,
- Dywergencja,
- Wskaźnik Hosmera-Lemeshowa,
- *AUC* - pole powierzchni pod krzywą *ROC*,
- Wykres przyrostu *(Lift)*
- Wykres zysku (*Gains*, krzywa *CAP*)

Dodatkowo dla każdego wskaźnika tworzony jest szczegółowy raport, a dla wskaźnika Kołmogorowa-Smirnowa, *Gini*, *ROC*, przyrostu oraz zysku generowane są także odpowiednie wykresy.

Moduł ten dodatkowo daje możliwość:

- generowania raportu punktacji końcowej (*Final score)* wraz z wykresami szans (*Odds)* oraz działań niepożądanych (*Bad rate),*
- generowania raportu cech (*Characteristic Report).*

Poniższy przykład prezentuje sposób użycia modułu do porównania kilku modeli zbudowanych różnymi metodami.

Przykład 13. Ocena i porównanie modeli (logit)

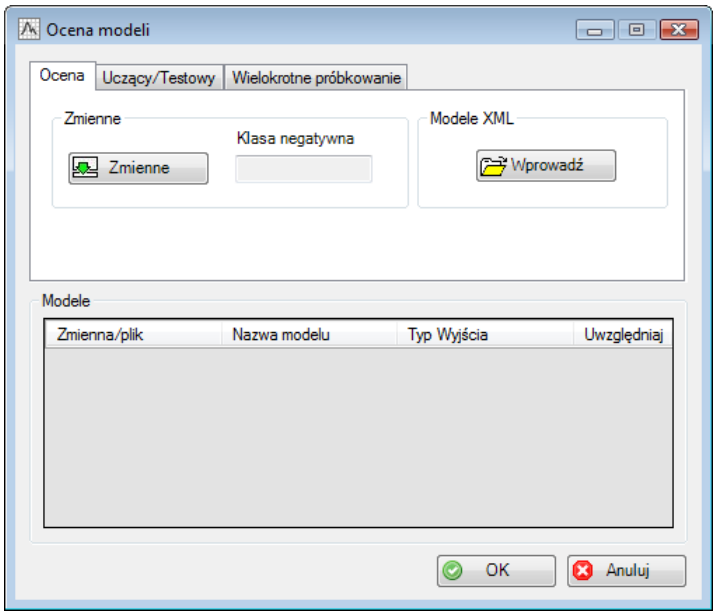
Do oceny modelu konieczne jest, aby w zbiorze danych zawarta była zmienna informująca o klasie kredytu (kredyt *dobry-zły*) oraz wynik skoringu, który może

- znajdować się w arkuszu danych – w takiej sytuacji wskazujemy zmienną (zmienne) zawierającą(e) wartości skoringu w oknie wyboru zmiennych,
- zostać wyliczony na podstawie modelu zapisanego w pliku XML.

Możliwe jest również porównanie wyników skoringu zapisanego w zbiorze danych ze skoringiem wyliczanym na podstawie skryptu *XML*.

Ocenę modeli wykonamy na podstawie zbioru *CreditScoring.sta,* który posłużył nam do budowy modelu we wcześniejszym przykładzie. Oceniać będziemy modele zapisane w postaci skryptów *XML*. Mając na uwadze czytelność wykresów wczytane zostaną tylko dwa modele, jednak nic nie stoi na przeszkodzie, aby wczytać większą liczbę modeli.

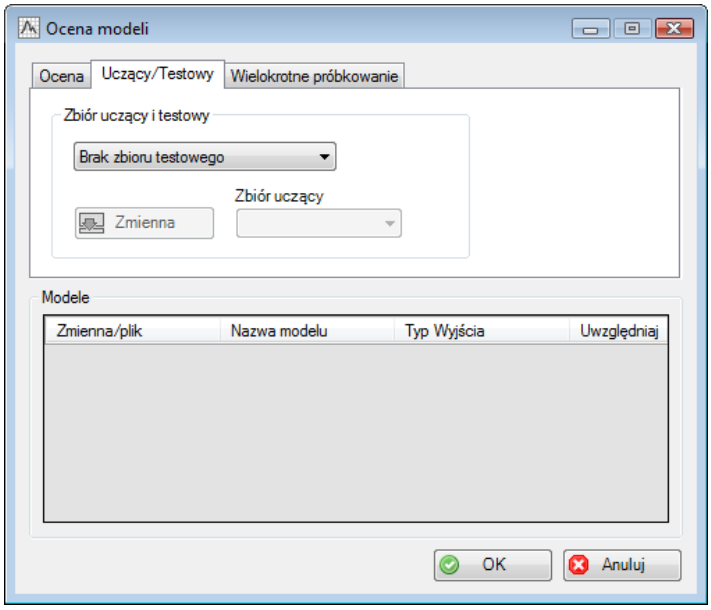
Moduł ***Ocena modeli*** uruchamiamy wybierając z menu ***Zestaw Skoringowy*** polecenie ***Ocena i kalibracja*** | ***Ocena modeli*** wyświetlając okno o tej samej nazwie.



W oknie tym możemy przeprowadzić procedurę oceny modeli za pomocą następujących opcji:

Zmienne – przycisk wyświetla okno dialogowe, w którym możemy wybrać zmienną zależną. Dodatkowo opcjonalnie możliwy jest wybór zmiennych zawierających informację o prawdopodobieństwie zajścia modelowanego zjawiska lub informację o skoringu. Na przykład po zbudowaniu modelu skoringowego obliczyliśmy skoring w module ***Obliczanie skoringu*** i wyniki tych obliczeń chcielibyśmy wykorzystać w bieżącym module, bądź też za pomocą innych metod skoringowych (sieci neuronowych, drzew wzmacnianych) zbudowaliśmy model, jego odpowiedź obliczyliśmy w module ***Data Mining*** | ***Szybkie wdrażanie modeli predykcyjnych***, a teraz chcielibyśmy go ocenić.

Wprowadź – przycisk umożliwia wczytanie modeli kart skoringowych zbudowanych w module ***Budowa tablicy skoringowej***. Uwaga: Za pomocą tej opcji nie można wczytać modeli SURVIVAL. Aby ocenić jakość modeli SURVIVAL oblicz prawdopodobieństwo wynikające z tego modelu za pomocą modułu ***Obliczanie skoringu*** a następnie nowoutworzoną zmienną wybieramy za pomocą przycisku ***Zmienne***.



Na karcie ***Uczący/Testowy*** mamy możliwość wskazania zbioru testowego pozwalającego niezależnie ocenić siłę predykcyjną zbudowanego modelu.

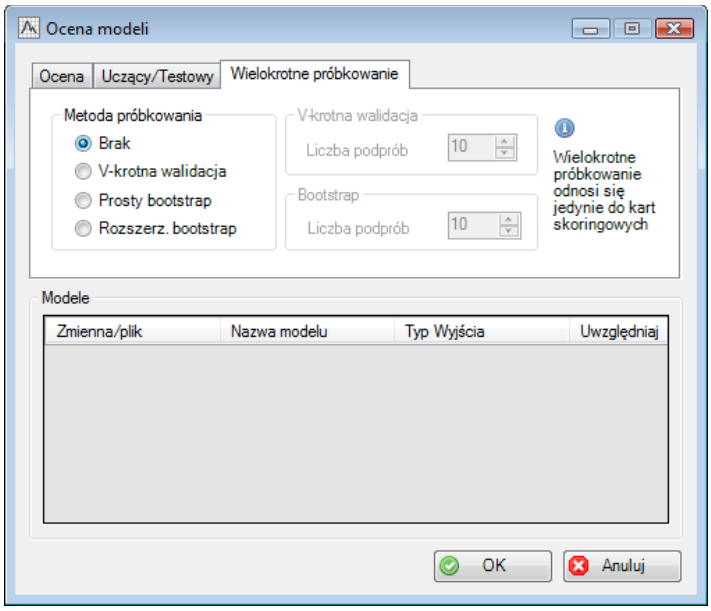
Lista rozwijalna znajdująca się w obszarze ***Zbiór uczący i testowy*** zawiera następujące opcje:

Brak zbioru testowego – opcja pozwala na obliczenie miar dobroci dopasowania na całym zbiorze bez podziału na zbiór uczący i testowy.

Identyfikator zbiorów w zmiennej pozwala na wskazanie zmiennej zawierającej identyfikator próby uczącej i testowej. Po wybraniu zmiennej na liście rozwijalnej ***Zbiór uczący*** wskazujemy symbol próby uczącej. Obliczenia wykonane zostaną niezależnie na obydwóch próbach.

Testowy w osobnym pliku umożliwia wskazanie dodatkowego arkusza z danymi, który będzie traktowany jako zbiór testowy. Obliczenia wykonane zostaną niezależnie na obydwóch plikach.

Na karcie ***Wielokrotne próbkowanie*** mamy możliwość wykonania niezależnej walidacji bez wyodrębniania dodatkowego podzbioru stanowiącego próbę testową.



W obszarze ***Metoda próbkowania*** mamy do wyboru następujące opcje:

Brak nie wykonuje walidacji za pomocą wielokrotnego próbkowania.

V-krotna walidacja po wybraniu tej opcji, zbiór danych zostanie podzielony na v podprób. Liczbę (v) podprób określamy w polu ***Liczba podprób*** w obszarze ***V-krotna walidacja***. Na podstawie zbioru składającego się z v-1 podprób budowany jest model, który następnie walidowany jest na v-tej podpróbie. Operacja ta powtarzana jest v razy tak, aby każdy z podzbiorów mógł pełnić rolę zbioru walidacyjnego. Wyniki uzyskane na podstawie v powtórzeń są następnie łączone i na ich podstawie oblicza się miary dobroci dopasowania modelu.

Prosty bootstrap po wybraniu tej opcji wykonywanych jest *k* prób *bootstrap* (próby te losowane są w schemacie ze zwracaniem i mają liczność próby oryginalnej), których liczba jest określana w polu ***Liczba podprób*** w obszarze ***Bootstrap***. Następnie dla każdej z podprób budowany jest model regresji. Model ten jest następnie oceniany na próbie oryginalnej (próba oryginalna jest zatem próbą testową). Oceny bootstrapowe uzyskuje się po uśrednieniu ocen uzyskanych dla poszczególnych powtórzeń.

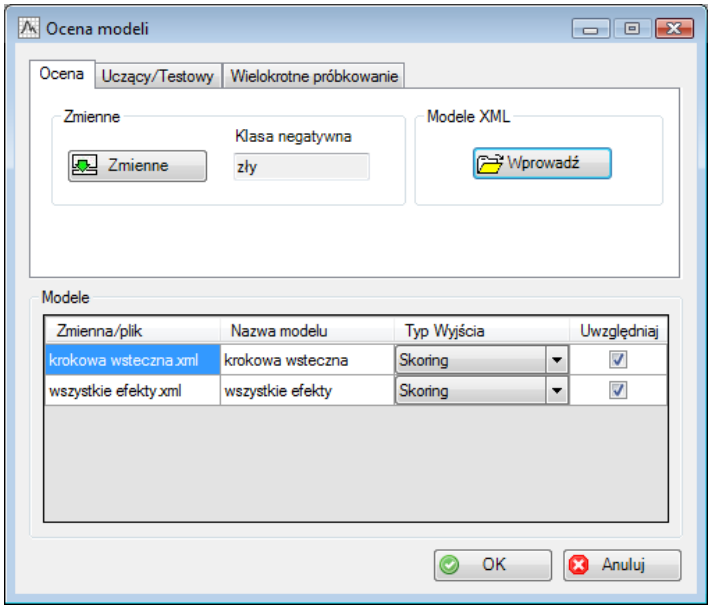
Rozszerz. bootstrap po wybraniu tej opcji, podobnie jak w przypadku opcji wcześniejszej, modele budowane są wielokrotnie na podstawie *k* prób bootstrap, których liczba jest określana w polu ***Liczba podprób*** w obszarze ***Bootstrap***. Po zbudowaniu modelu podobnie jak w prostym bootstrapie model jest stosowany na próbie oryginalnej. Ocena dobroci dopasowania określona na próbie oryginalnej jest następnie odejmowana od analogicznej oceny na próbie bootstrapowej. Wynik tego odejmowania nazywamy optymizmem. Proces ten jest powtarzany *k* razy a następnie otrzymane wartości optymizmu są uśredniane. Uśredniona wartość optymizmu jest ostatecznie odejmowana od oceny dobroci dopasowania oryginalnego modelu uzyskanej na próbie uczącej.

Aby wykonać analizę, w oknie ***Ocena modeli*** wskazujemy zmienną informującą o statusie kredytu, klikamy przycisk ***Zmienne***, a następnie na liście ***Zmienna stanu*** zaznaczamy zmienną ***Ocena***. Na liście ***Skoring/Prawdopodobieństwo (opcjonalnie)*** nie wybieramy żadnej zmiennej.

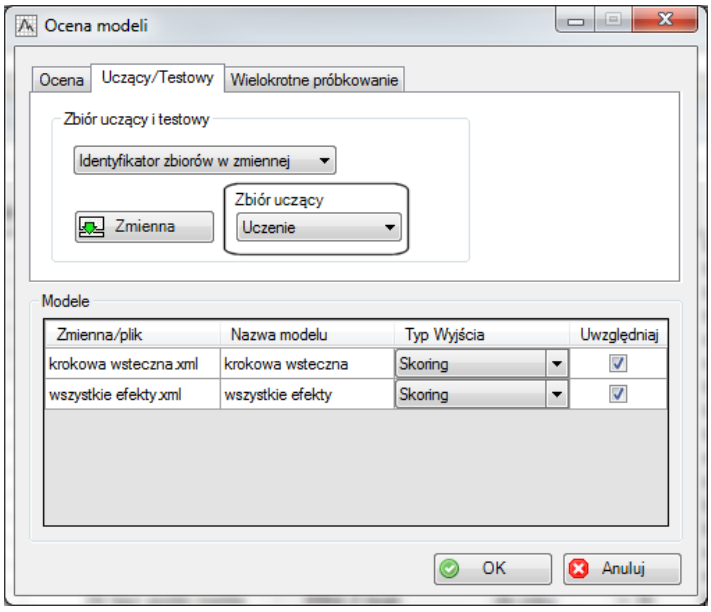
W kolejnym kroku, analogicznie do poprzednich przykładów, określamy symbol złego kredytu klikając dwukrotnie pole ***Klasa negatywna***. Przygotowane modele skoringowe wczytujemy do programu za pomocą przycisku ***Wczytaj***. Po wskazaniu odpowiednich plików *XML* i zatwierdzeniu wyboru, w obszarze ***Modele*** pojawią się informacje o wczytanych modelach. Na liście tej możemy określić nazwę modelu, jaka powinna występować w raporcie oraz określić, czy powinien on być w

nim uwzględniony. Jeśli nasz model oceniamy na podstawie wyniku zapisanego w zbiorze danych (podczas wyboru zmiennych na liście *Skoring/Prawdopodobieństwo (opcjonalnie)* wskazaliśmy przynajmniej jedną zmienną), wtedy na liście ***Typ Wyjścia*** należy określić, czy w zmiennej tej zapisano wartości będące prawdopodobieństwem - opcja ***Prawdop.***, czy też są to liczby naturalne opcja ***Skoring***.

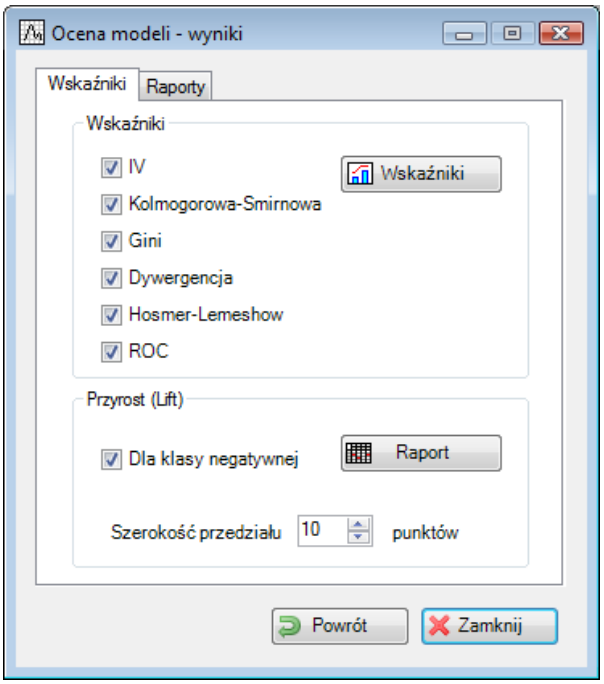
Uwaga. *W module do oceny modeli można porównywać modele (zbudowane na przykład drzewami decyzyjnymi lub dowolną inną metodą), które zamiast skoringu wyliczają prawdopodobieństwo. W tym celu w oknie wyboru zmiennych należy wskazać je na drugiej liście wyboru (musi to być prawdopodobieństwo zajścia zdarzenia POZYTYWNEGO) i w sekcji Modele zmienić parametr w kolumnie Typ wyjścia na Prawdopodobieństwo.*



Modele można także sprawdzać obliczając wskaźniki równocześnie dla zbioru uczącego i testowego. Odpowiednie opcje znajdują się na zakładce ***Uczący/testowy***. Podobnie jak robiliśmy to podczas budowy modelu wskażemy zmienną identyfikującą próbę. Z listy rozwijalnej wybieramy opcję ***Identyfikator zbiorów w zmiennej*** a następnie za pomocą przycisku ***Zmienna*** wskazujemy zmienną ***Próba***.



Po wczytaniu modeli i wskazaniu zmiennej informującej o statusie kredytu oraz próbie zatwierdzamy analizę klikając ***OK*** i przechodzimy do okna ***Ocena modeli – wyniki***.



Grupa ***Wskaźniki*** umożliwia wybór szeregu miar dobroci dopasowania używanych między innymi w ryzyku kredytowym. Szczegóły obliczeniowe można znaleźć w dokumencie Scorecard Formula Guide.

Opcja ***IV*** oblicza wskaźnik *Information Value* analogicznie do obliczeń przedstawionych w opisie modułu tworzącego ranking predyktorów. Obliczenia te są wykonywane bez kategoryzacji wartości skoringu lub prawdopodobieństwa – każda wartość skoringu tworzy osobną kategorię.

Opcja ***Kołmogorowa-Smirnowa*** oblicza statystykę KS wraz z powiązaną z nią wartością prawdopodobieństwa testowego p. Dodatkowo generowane są odpowiednie wykresy powiązane z obliczoną statystyką. Statystyka KS opisuje o ile maksymalnie może różnić się prawdopodobieństwo spłacenia od niespłacenia kredytu wśród wszystkich poziomów skoringu. Przyjmuje wartości od zera (dla danej punktacji kredyt ma takie samo prawdopodobieństwo spłacenia, co niespłacenia) do jeden (w danej klasie są tylko kredyty spłacone lub tylko niespłacone). Wartości KS możemy interpretować w następujący sposób:

- Poniżej 0,2 – model do odrzucenia,
- 0,2 – 0,4 – słaba siła dyskryminacyjna,
- 0,4 – 0,5 – akceptowalna siła dyskryminacyjna,
- 0,5 – 0,6 – duża siła dyskryminacyjna,
- 0,6 – 0,75 – bardzo duża siła dyskryminacyjna,
- Powyżej 0,75 – zbyt dobrze by było prawdziwie

Wykres KS przedstawia skumulowane procenty dobrych i złych przypadków. Maksymalna różnica pomiędzy tymi dwoma dystrybuantami wyznacza wartość KS. Wykres *Wartości KS względem skoringu* pokazuje wartość różnicy pomiędzy dystrybuantami dobrych i złych przypadków dla każdej wartości skoringu. Pozwala to określić zakres wartości skoringu, dla których model ma największą siłę dyskryminacyjną.

Opcja ***Gini*** pozwala na obliczenie wskaźnika Giniego oraz odpowiednich wykresów dla każdego modelu i próby. Dodatkowo tworzony jest zbiorczy wykres przedstawiający

wszystkie modele i próby. Wskaźnik *Gini* mierzy do jakiego stopnia model ma większą siłę predykcyjną od losowego klasyfikatora. Gini przyjmuje wartości z przedziału [0,1], gdzie 0 odpowiada modelowi losowemu a 1 odpowiada modelowi idealnemu. Możemy interpretować tę miarę w następujący sposób:

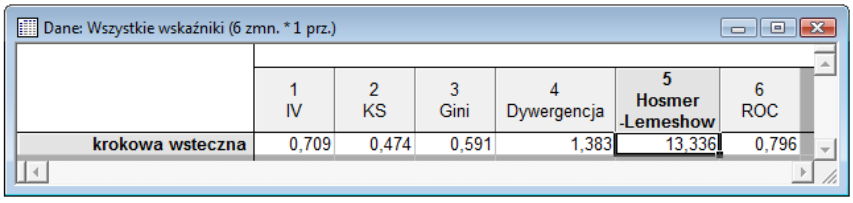
- Poniżej 0,2 – model do odrzucenia
- 0,2 – 0,4 – słaba siła dyskryminacyjna
- 0,4 – 0,6 – akceptowalna siła dyskryminacyjna
- 0,6 – 0,95 – duża siła dyskryminacyjna
- Powyżej 0,95 – zbyt dobrze by było prawdziwie

Opcja ***Dywergencja*** umożliwia obliczenie statystyki dywergencji – miary przedstawiającej różnicę między rozkładami dobrych i złych kredytów(a dokładniej kwadrat różnicy w średniej wartości skoringu przypadków „dobrych" i „złych" wystandaryzowany ich wariancjami). Statystyka ta przyjmuje wartości od 0 do dużych wartości dodatnich.

Statystyka ***Hosmera-Lemeshowa*** jest wskaźnikiem typu Chi kwadrat i mierzy zgodność między odpowiedziami modelu a rzeczywistymi wartościami zmiennej zależnej. Jest podstawą testu dobroci dopasowania.

ROC odnosi się do pola powierzchni pod krzywą ROC (*Receiver Operating Characteristic*). Krzywa ROC jest tworzona przez przedstawienie na jednym wykresie odsetka prawdziwie dodatnich (czułość) oraz odsetka fałszywie dodatnich (1-specyficzność). Pole powierzchni pod krzywą ROC przyjmuje wartości od 0 do 1. Wartość 0,5 oznacza model losowy. Wartość pola powierzchni może być także obliczona na podstawie wskaźnika Gini według poniższego wzoru: $ROC = \frac{Gini + 1}{2}$.

Po kliknięciu przycisku ***Wskaźniki*** generujemy raport zawierający wskaźniki odpowiadające zaznaczonym opcjom – dla każdego wskaźnika obliczony zostanie dodatkowo szczegółowy raport, a dla wskaźnika Kołmogorowa-Smirnowa, *Gini* oraz *ROC* wygenerowane zostaną dodatkowo odpowiednie wykresy.


Dane: Wszystkie wskaźniki (6 zmn. * 1 prz.)

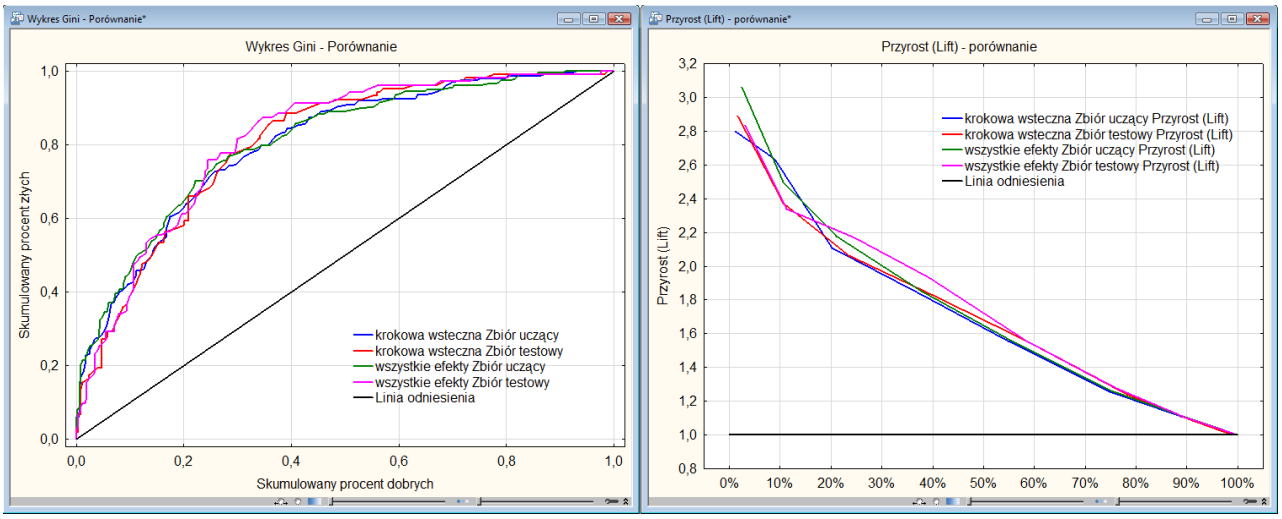
	1 IV	2 KS	3 Gini	4 Dywergencja	5 Hosmer -Lemeshow	6 ROC
krokowa wsteczna	0,709	0,474	0,591	1,383	13,336	0,796

Opcje ***Przyrost (Lift)*** pozwalają na wygenerowanie wykresów Przyrostu (*Lift*) oraz Zysku (*Gains*, *CAP*) dla modeli oraz prób. Więcej informacji na temat tych wykresów zobacz: Lift Chart oraz Gains Chart.

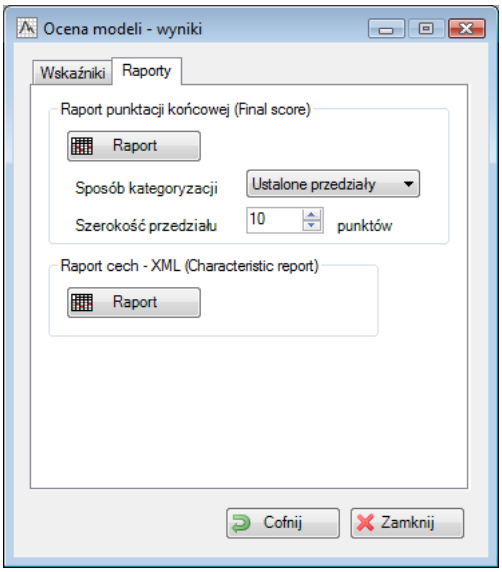
Dla klasy negatywnej – jeśli zaznaczona jest ta opcja, obliczenia zostaną wykonane dla klasy „negatywnej". W przeciwnym przypadku klasą odniesienia będzie klasa pozytywna.

Szerokość przedziału – określa szerokość przedziału, jaka zostanie przyjęta podczas kategoryzacji skoringu przed wykonaniem obliczeń.

W obszarze ***Lift*** po naciśnięciu przycisku ***Raport*** otrzymujemy wykresy Przyrostu (*Lift*) oraz Zysku (*Gains*), a także raport zawierający informację o przyroście oraz przyroście skumulowanym.



Zakładka ***Raporty*** pozwala na wygenerowanie raportu punktacji końcowej (***Final Score)*** oraz raportu cech (***Characteristic report)***.

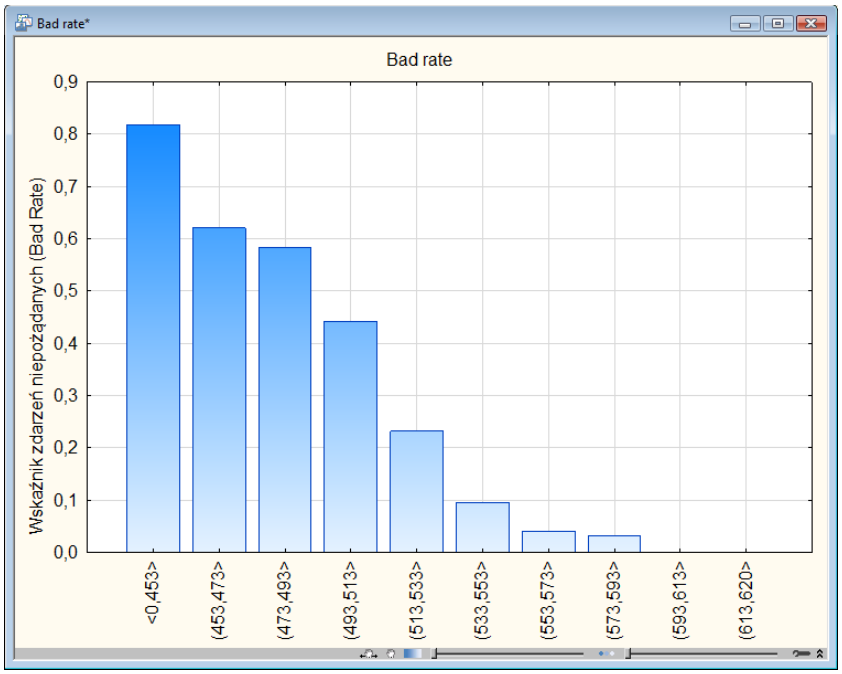


Raport punktacji końcowej (Final score)– pozwala na wygenerowanie raportów przedstawiających szczegółowe statystyki jakości analizowanego portfela dla określonej szerokości przedziałów. Dodatkowo generowane są wykresy *Szans* (*ODDS*) oraz *Zdarzeń niepożądanych* (*Bad Rate*).

> ***Sposób kategoryzacji*** pozwala na przygotowanie wyżej wymienionych raportów i wykresów na podstawie skategoryzowanej punktacji (opcja ***Ustalone przedziały***) lub na podstawie surowych wartości skoringu (opcja ***Poszczególne wartości***).
>
> ***Szerokość przedziału*** określa szerokość przedziału, jaka zostanie przyjęta podczas kategoryzacji skoringu przed wykonaniem obliczeń.

Raport cech – XML (Characteristic report) umożliwia wygenerowanie raportów przedstawiających szczegółowe statystyki jakości analizowanego portfela dla wszystkich kategorii (atrybutów) cech zawartych w modelu. Opcja ta jest aktywna jedynie dla modeli wczytanych do modułu za pomocą plików XML. Raporty te mogą być bardzo przydatne zwłaszcza na etapie monitorowania modeli, aby wykryć ewentualne zmiany w sile predykcyjnej poszczególnych zmiennych.

Poniżej przedstawiono przykładowy wykres *Bad rate* dla jednego z wczytanych modeli.

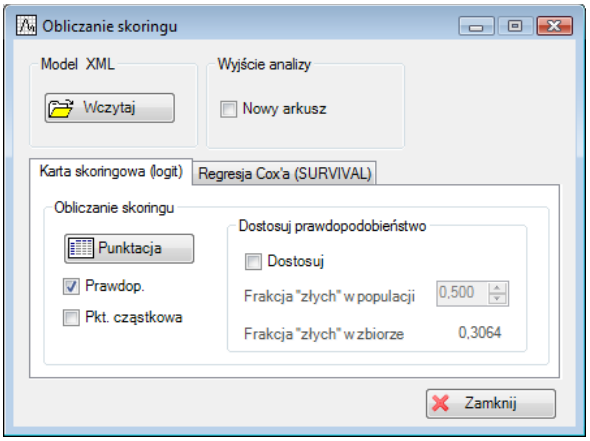


Przykład 14. Ocena modeli na podstawie prawdopodobieństwa

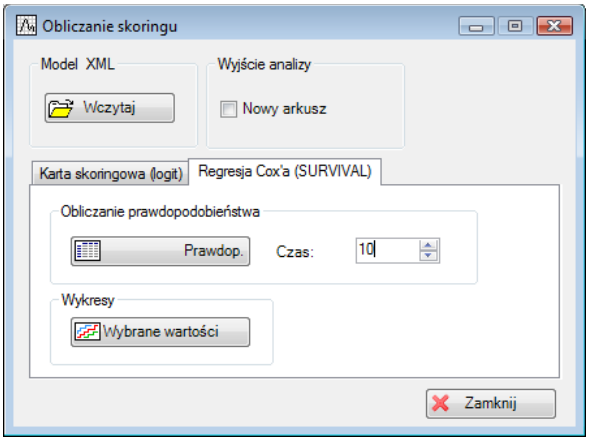
W module do oceny modeli możemy oprócz skoringu (który nie zawsze jest wyliczany) posłużyć się prawdopodobieństwem przynależności do klasy pozytywnej. W tym celu porównamy ze sobą dwa modele jeden zbudowany metodą regresji logistycznej oraz drugi budowany metodą regresji Coxa.

Ocenę modeli wykonamy na podstawie zbioru *TimeTest.sta*, znajdującego się w katalogu ze zbiorami przykładowymi. Pierwszym krokiem będzie wdrożenie modeli dla nowego zbioru danych. W tym celu wybieramy z menu ***Zestaw skoringowy*** kolejno ***Ocena i kalibracja*** | ***Obliczanie skoringu***.

Następnie w oknie ***Obliczanie skoringu*** wczytujemy pierwszy model (na przykład *krokowa wsteczna time.xml* dostępny w katalogu z przykładami) – zbudowany metodą regresji logistycznej i zaznaczamy pole ***Prawdopodobieństwo***, aby oprócz skoringu zostało wyliczone prawdopodobieństwo zarówno dla klasy pozytywnej jak i negatywnej.



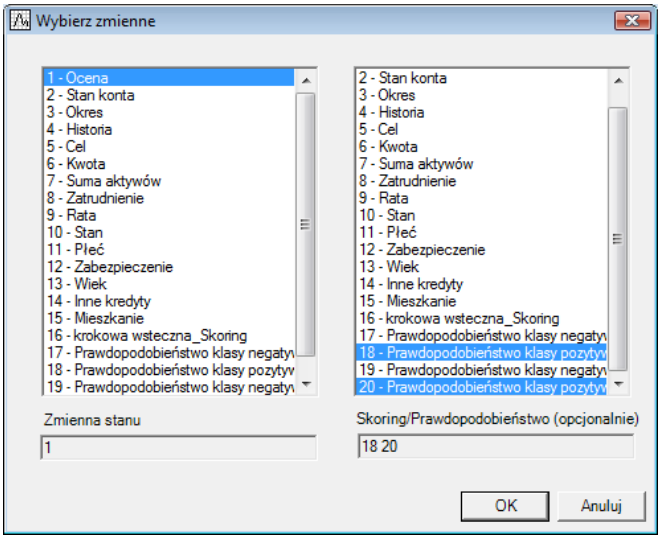
Po kliknięciu ***Punktacja*** odpowiednie kolumny zostaną dodane do zbioru danych. Następnie należy wczytać skrypt zawierający model regresji Coxa (na przykład plik *cox.xml* dostępny w katalogu z przykładami). Po wczytaniu program przełączy się automatycznie na drugą zakładkę, gdzie należy wybrać czas, dla którego ma być obliczane prawdopodobieństwo przynależności do obydwóch klas (na przykład 10).



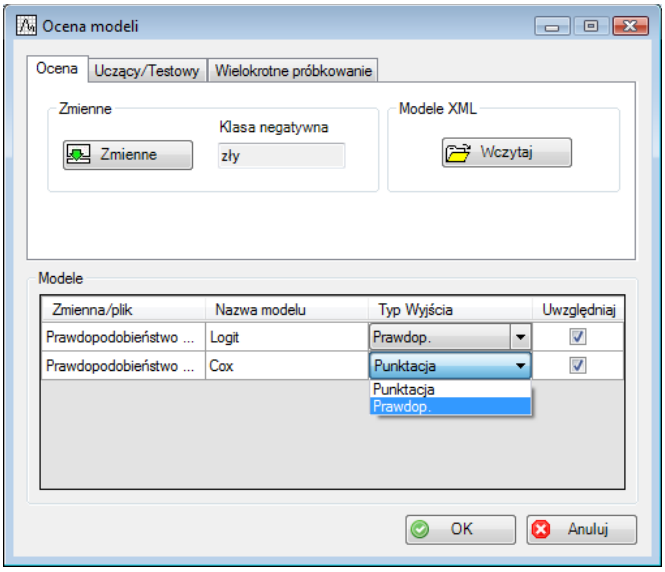
Po kliknięciu przycisku ***Prawdop***. odpowiednie kolumny zostaną dodane do zbioru.

Uwaga. *Użytkownicy programu STATISTICA Data Miner mają również do dyspozycji szereg innych metod skoringowych takich jak sieci neuronowe, drzewa wzmacniane, MARS czy metoda wektorów nośnych. Zbudowane za pomocą tych i innych metod modele można zapisać w postaci kodu PMML a następnie wyliczyć na ich podstawie prawdopodobieństwo modelowanego zdarzenia w module Data Mining | Szybkie wdrażanie modeli predykcyjnych.*

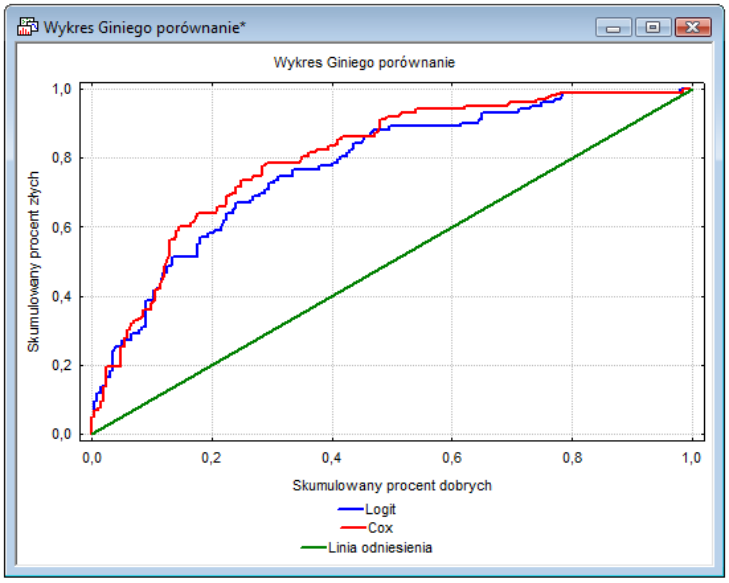
Kolejnym krokiem będzie porównanie tych dwóch modeli – dla każdego modelu mamy bowiem dwie wartości prawdopodobieństwa przynależności do klasy pozytywnej. Uruchamiamy moduł do oceny modeli wybierając z menu ***Zestaw skoringowy*** | ***Ocena i kalibracja*** | ***Ocena modeli***. Następnie klikamy przycisk ***Zmienne*** i wskazujemy zmienne zawierające stan – w naszym przypadku to zmienna ***Ocena*** oraz odpowiednie prawdopodobieństwa klasy pozytywnej.



Należy jeszcze wskazać symbol wartości negatywnej, oraz zmienić ***Typ wyjścia*** dla obu modeli na ***Prawdop***. Aby łatwiej odróżnić modele od siebie zmienimy pierwszemu z nich nazwę na *Logit*, drugiemu zaś na *Cox*.



Po kliknięciu przycisku *OK* otrzymujemy dostęp do identycznego okna jak w poprzednim przykładzie. Również zestaw wskaźników i raportów jest niemal identyczny – niedostępny jest jedynie raport Cech (*Characteristic report).*



5.3. Zarządzanie punktem odcięcia

Po określeniu sposobu przyznawania punktów kredytobiorcom, kolejnym ważnym krokiem jest ustalenie granicznego punktu (bądź punktów) w ocenie, pozwalającego na przydzielenie klientów do różnych klas ryzyka (punkty *cut-off*). Do tego celu służy kolejny moduł ***Zestawu Skoringowego*** – ***Ocena i kalibracja*** | ***Zarządzanie punktem odcięcia***. Za jego pomocą możemy określić punkt odcięcia w sposób automatyczny, na podstawie analizy krzywej *ROC* przy uwzględnieniu podanych przez użytkownika kosztów błędnych klasyfikacji oraz rzeczywistej frakcji złych kredytów w populacji. Dodatkowo użytkownik może ręcznie wprowadzić od jednego do trzech punktów odcięcia oraz ocenić jakość podziału za pomocą zestawu wykresów i raportów.

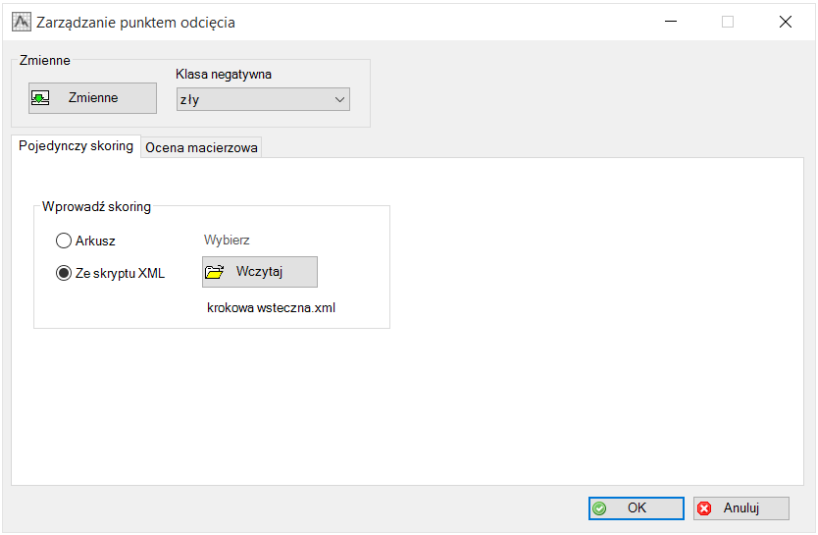
Dla regresji logistycznej mamy możliwość wczytania skoringu w postaci pliku *XML* ze specyfikacją modelu, natomiast dla wszystkich innych metod skoring lub prawdopodobieństwo można wczytać z arkusza danych wskazując odpowiednie zmienne.

Jeżeli chcemy przeprowadzić *analizę macierzową* uwzględniającą dwa sposoby oceny (modele, kategorie ratingowe), to możemy wczytać obydwa z plików XML (dla modeli regresji logistycznej)

lub wskazać w arkuszu odpowiednie zmienne. Jeżeli dysponujemy skryptem dyskretyzacji (utworzonym w module *Dyskretyzacja*) dzielącym punktację na klasy, to możemy również wczytać go w celu wykorzystania w ocenie macierzowej. Wynikiem takiej analizy są klasy ryzyka wyznaczone w oparciu o obydwa modele łącznie.

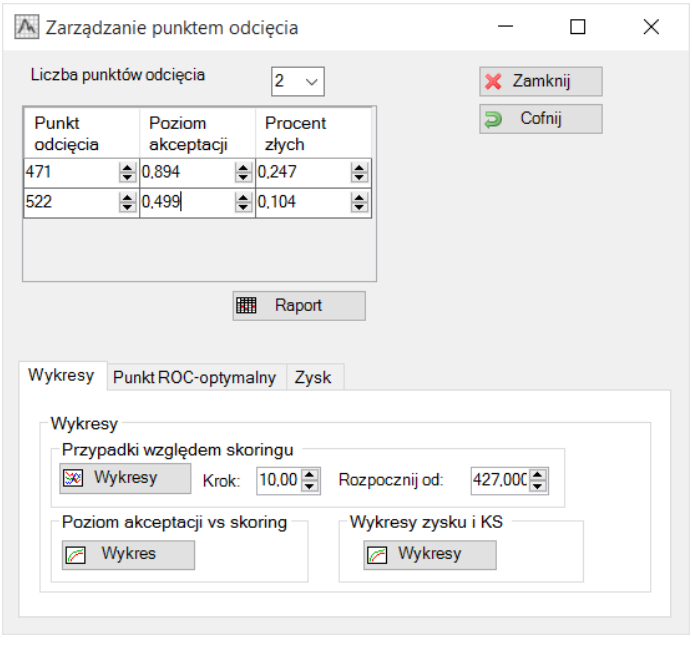
Przykład 15. Wybór punktu odcięcia

Do określenia punktu odcięcia ponownie wykorzystamy model zbudowany w jednym z wcześniejszych przykładów. W oknie ***Zarządzanie punktem odcięcia*** po wybraniu zmiennej ***Ocena***, określeniu symbolu złego kredytu oraz wczytaniu modelu (przycisk *XML*), zatwierdzamy wykonanie analizy przyciskiem ***OK***.



W przypadku wybrania zakładki ***Ocena macierzowa*** możemy w analogiczny sposób wprowadzić drugi model Można też wybrać zmienne na obydwu listach ***Skoring*** i ***Drugi skoring*** dostępnych po kliknięciu przycisku **Zmienne** (jeśli modele nie są wczytywane ze skryptów)**.** W dowolnym momencie możemy zmienić skoring wczytywany z arkusza na wynik wskazanego modelu (przełączając opcje w polu ***Wprowadź skoring***). Przebieg analizy macierzowej punktów odcięcia, wraz z dokładnym opisem powiązanych opcji programu, przedstawiono w przykładzie w dalszej części rozdziału.

Kliknięcie przycisku ***OK*** spowoduje wyświetlenie okna ***Wybór punktu odcięcia – wyniki***.



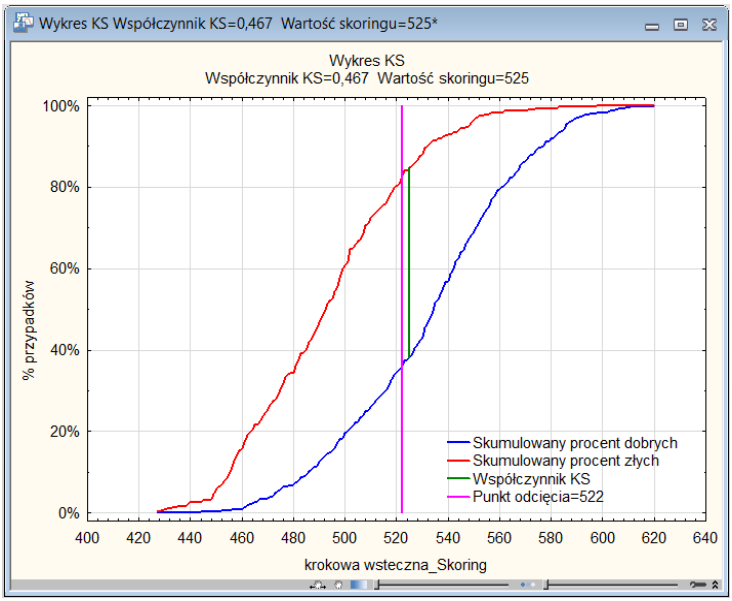
W jego górnej części w obszarze ***Punkt odcięcia*** mamy możliwość ręcznego wprowadzenia do trzech punktów odcięcia (opcja ***Liczba punktów odcięcia***) i określania dla nich poziomu akceptacji oraz procenta złych przypadków (*bad rate*). Każdy punkt odcięcia może być zdefiniowany za pomocą jednej z trzech kolumn tabeli:

> ***Punkt odcięcia*** pozwala wprowadzić określoną wartość skoringu lub prawdopodobieństwa jako proponowany punkt odcięcia. Po jego wprowadzeniu obliczone zostaną pozostałe miary opisane poniżej.
>
> ***Poziom akceptacji*** określa procent klientów, jaki powinien znaleźć się powyżej punktu odcięcia.
>
> ***Procent złych*** określa, jaki jest poziom „*bad rate*" w grupie osób ze skoringiem powyżej punktu odcięcia.

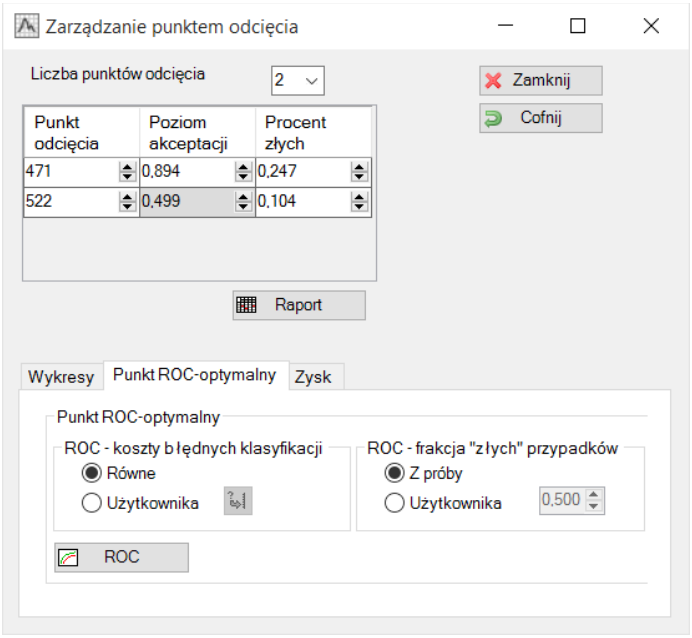
Miary te są ściśle powiązane pomiędzy sobą. Zmiana wartości w dowolnej komórce danego wiersza powoduje zmianę pozostałych wartości w tym wierszu.

Raport – przycisk tworzy arkusze ze szczegółowymi podsumowaniami jakości analizowanego portfela dla wskazanych punktów odcięcia. Dodatkowo tworzony jest wykres przedstawiający trzy parametry opisane powyżej: *Punkt odcięcia*, *Poziom akceptacji*, *Procent złych*.
Karta ***Wykresy*** zawiera opcje umożliwiające przygotowanie raportów dla wybranego punktu odcięcia. Na wszystkich tych wykresach naniesiona zostanie linia reprezentująca zaznaczony punkt odcięcia.

> ***Przypadki względem skoringu***- wykres przedstawia procent lub liczbę dobrych i złych przypadków w każdym przedziale punktacji. Przedziały są definiowane na podstawie opcji ***Krok*** oraz ***Rozpocznij od***.
>
> ***Poziom akceptacji vs skoring*** – wykres przedstawia zmiany w poziomie akceptacji w zależności od poziomu skoringu.
>
> ***Wykresy zysku i KS*** – generuje wykresy *Zysku* i *Kołmogorowa-Smirnowa* w sposób analogiczny do wykresów generowanych w module do oceny modeli.



Na karcie ***Punkt ROC optymalny*** możemy w sposób automatyczny wyznaczyć punkt odcięcia klikając przycisk ***ROC*** i odczytując proponowany punkt z ***Wykresu ROC***. Szczegóły techniczne obliczania punktu ROC optymalnego można znaleźć w dokumencie Scorecard Formula Guide.



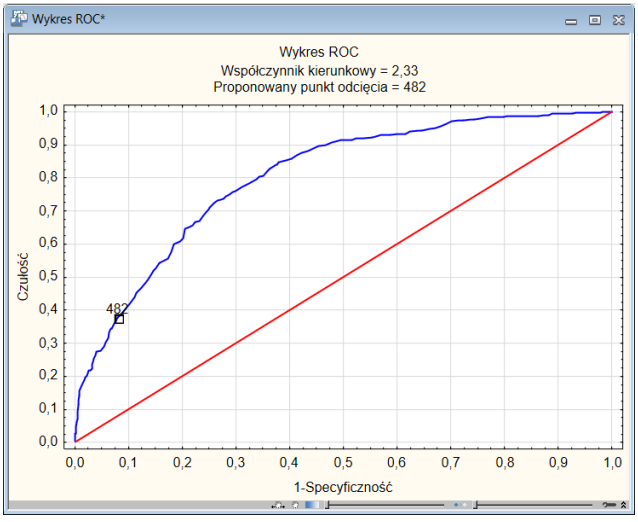
ROC – koszty błędnych klasyfikacji umożliwiają użytkownikowi zdefiniowanie własnych kosztów błędnych klasyfikacji.

> ***Równe*** – przypadki fałszywie dodatnie oraz fałszywie ujemne są traktowane z taką samą wagą.
>
> ***Użytkownika*** – pozwala określić własne koszty dla przypadków fałszywie dodatnich oraz fałszywie ujemnych.

ROC – frakcja "złych" przypadków pozwala określić oczekiwaną frakcję złych przypadków w populacji przychodzącej

> ***Z próby*** – frakcja złych jest obliczana na podstawie wejściowego zbioru danych.
>
> ***Użytkownika*** – jeżeli z pewnych powodów frakcja złych w próbie jest inna niż oczekiwana w populacji przychodzącej, użytkownik ma możliwość wprowadzenia oczekiwanej frakcji złych (*bad rate*).

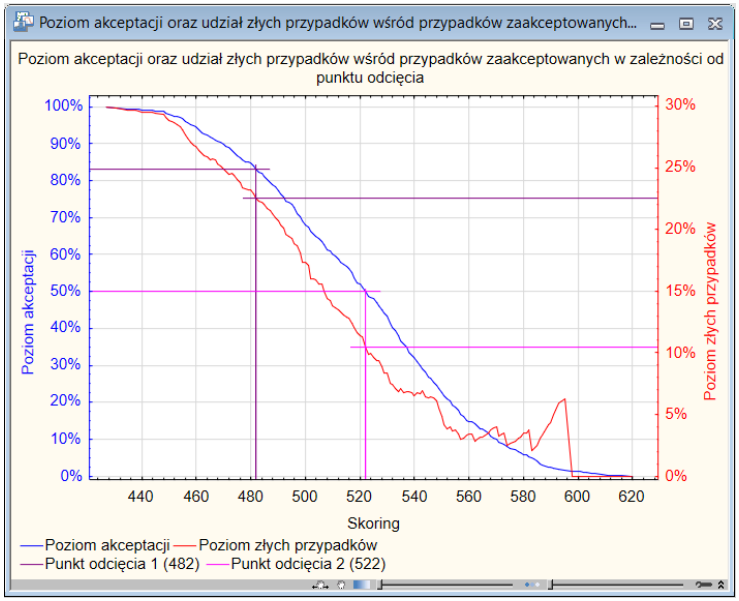
Na podstawie wykresu *ROC* wyznaczamy punkt odcięcia na poziomie 482 punktów. Wartość tą wpiszemy do pierwszego wiersza tabeli w kolumnie *Punkt odcięcia*.



Uwaga. *Parametrami wpływającymi na wartość optymalnego punktu odcięcia są koszty błędnych klasyfikacji oraz oczekiwany poziom złych kredytobiorców w populacji generalnej. Użytkownik ma możliwość zmiany domyślnych wartości, w celu uwzględnienia dodatkowej wiedzy biznesowej.*

Drugi punkt odcięcia wyznaczamy ręcznie przyjmując kryterium poziomu *złych* na poziomie 10%. Po wpisaniu wartości 0,1 w kolumnie ***Procent złych***, w kolumnie ***Punkt odcięcia*** automatycznie pojawiła się wartość 522.

Dla wskazanych punktów odcięcia możemy następnie wygenerować raport (przycisk ***Raport***), pozwalający określić jakość dokonanego podziału. Możemy wykorzystać do tego wykres określający poziom złych kredytów oraz poziom akceptacji dla wskazanych punktów odcięcia lub też ocenić jakość podziału na podstawie tabeli z klasyfikacją zgodną ze wskazanymi punktami.



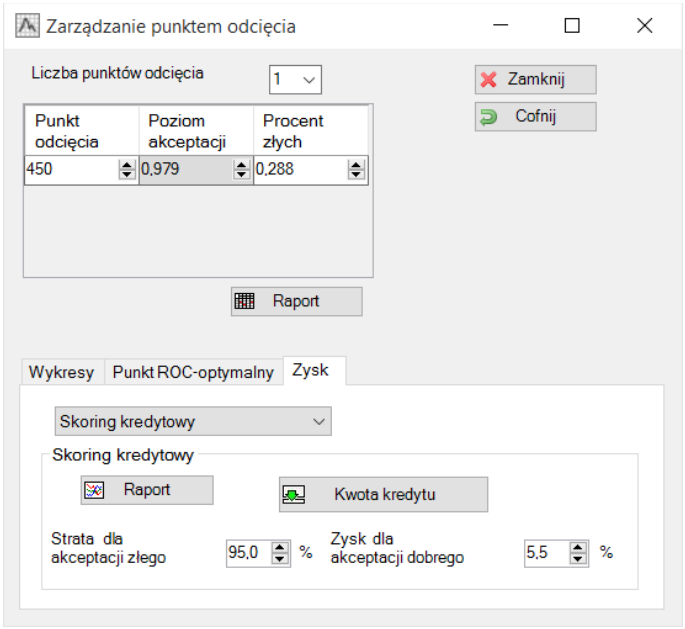
Dane: Punkty odcięcia - podsumowanie* (8 zmn. * 4 prz.)

Skoring	Dobre	Złe	Suma	Portfel	Procent złych	Procent dobrych	Procent ogółem
(-inf;482>	56	113	169	66,86%	37,67%	8,00%	16,90%
(482;522>	197	135	332	40,66%	45,00%	28,14%	33,20%
(522;inf)	447	52	499	10,42%	17,33%	63,86%	49,90%
Ogół	700	300	1000	30,00%	100,00%	100,00%	100,00%

Przykład 16. Optymalny punkt odcięcia z uwzględnieniem częściowej spłaty kredytu

Wyznaczając optymalny poziom odcięcia możemy także uwzględnić możliwość jedynie częściowej spłaty zadłużenia. Wtedy strata banku zmniejsza się o kwotę, którą udało się odzyskać. Analizę tę możemy wykonać na karcie ***Zysk***. Na karcie tej w zależności od wyboru rodzaju skoringu dokonanego za pomocą listy rozwijalnej możemy przeprowadzić analizę zyskowności dla skoringu kredytowego, marketingowego, churn oraz nadużyć. W przypadkach wykraczających poza te najbardziej typowe scenariusze możemy przeprowadzić ogólną analizę zyskowności wybierając opcję *Ogólna macierz zysków/strat*, której wykorzystanie zaprezentowano w następnym przykładzie.

Po wybraniu opcji ***Skoring kredytowy*** użytkownik ma do wyboru następujące parametry:



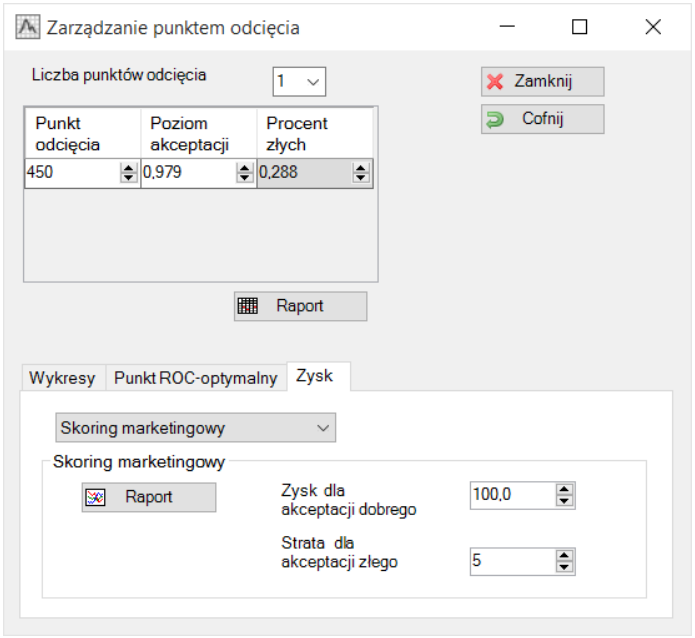
Raport przycisk generuje raport zysku w zależności od punktu odcięcia.

Kwota kredytu umożliwia wskazanie zmiennej zawierającej kwotę kredytu.

Strata dla akceptacji złego określa, jaki procent kwoty „złego" kredytu nie zostaje spłacony i nie może zostać odzyskany.

Zysk dla akceptacji dobrego określa zysk powiązany z „dobrymi" kredytobiorcami wyrażony jako procent udzielonej kwoty kredytu.

Po wybraniu opcji ***Skoring marketingowy*** użytkownik ma do wyboru następujące parametry:

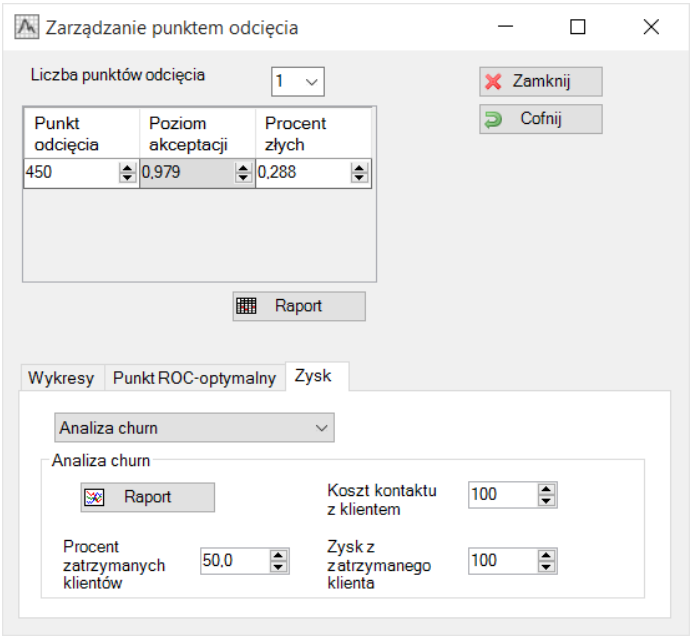


Raport – przycisk generuje raport zysku w zależności od punktu odcięcia.

Zysk dla akceptacji dobrego – określa kwotę, jaką zyskujemy w sytuacji, gdy klient przyjmie naszą ofertę.

Strata dla akceptacji złego – określa koszt wysłania/prezentacji oferty klientowi, który z niej nie skorzysta – może to być na przykład koszt wysłania oferty pocztą lub rozmowy telefonicznej.

Po wybraniu opcji ***Analiza churn*** użytkownik ma do wyboru następujące parametry:



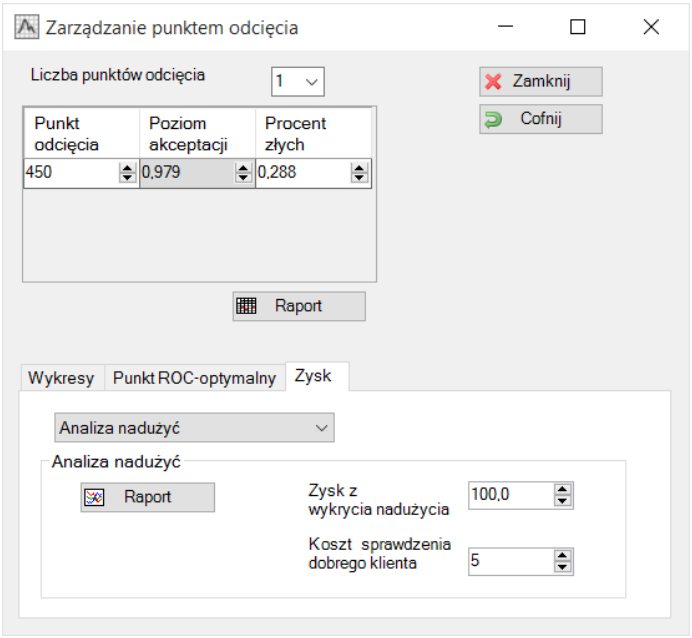
Raport – przycisk generuje raport zysku w zależności od punktu odcięcia.

Koszt kontaktu z klientem – określa całkowity koszt kontaktu z klientem.

Zysk z zatrzymanego klienta – określa zysk wynikający z zatrzymania klienta.

Procent zatrzymanych klientów – określa, jaki procent osób skłonnych do odejścia spodziewamy się zatrzymać dzięki planowanej akcji.

Po wybraniu opcji ***Analiza nadużyć*** użytkownik ma do wyboru następujące parametry:

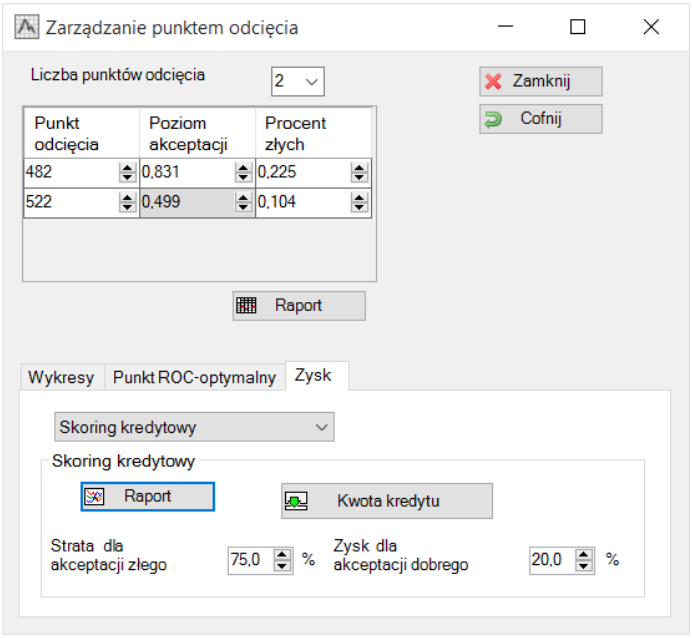


Raport – przycisk generuje raport zysku w zależności od punktu odcięcia.

Zysk z wykrycia nadużycia – określa kwotę, jaką zyskujemy dzięki wykrytemu przypadkowi nadużycia.

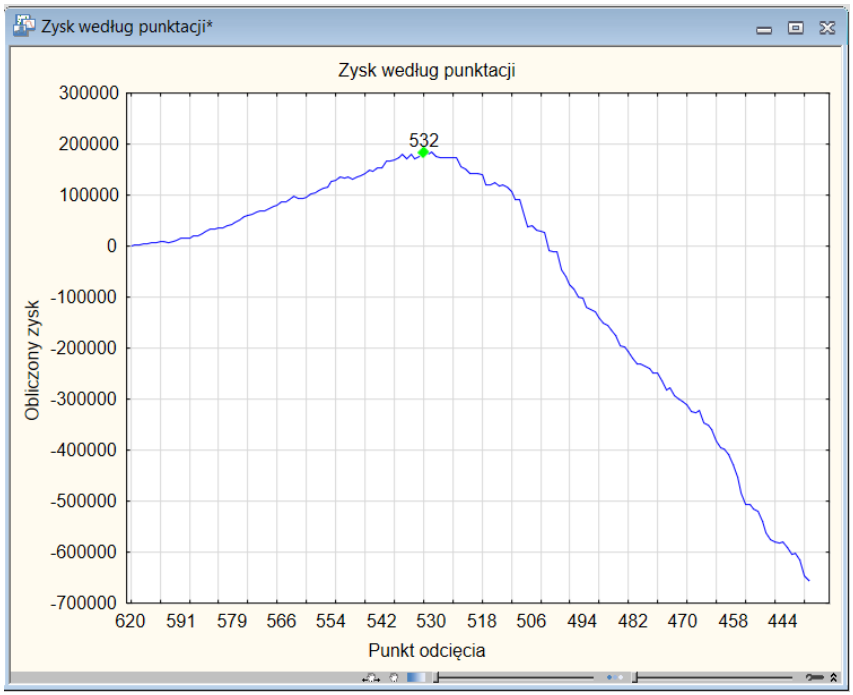
Koszt sprawdzenia dobrego klienta – określa koszt, jaki ponosimy sprawdzając przypadek legalny.

Ponieważ w naszym przypadku analizowany problem dotyczy skoringu kredytowego, by móc wykonać ten raport przechodzimy na kartę ***Zysk*** a następnie z listy rozwijalnej wybieramy opcję ***Skoring kredytowy***.



W pierwszym kroku po kliknięciu przycisku ***Kwota kredytu*** wskazujemy zmienną *Kwota* zawartą w analizowanym zbiorze danych. Następnie mamy możliwość podania ile procent kwoty kredytu bank zyskuje, gdy klient spłaci go w całości (***Zysk dla akceptacji dobrego***), oraz procentową stratę, czyli jakiej części udzielonego złego kredytu nie uda się odzyskać (***Strata dla akceptacji złego***).

Zakładając, że na każdym dobrym (spłaconym) kredycie zyskamy 20% kwoty, a na każdym złym (niespłaconym) tracimy 75% optymalny poziom odcięcia przesuwa się do 532 punktów.



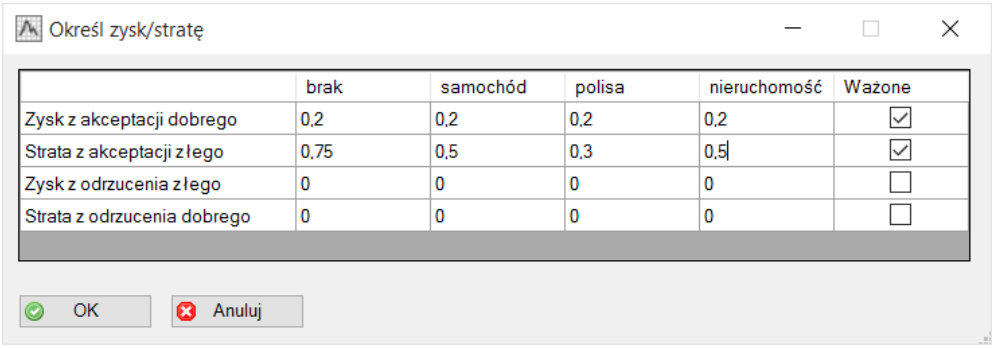
Widać więc, że dysponując odpowiednimi danymi na przykład z działu windykacji możemy uelastycznić ofertę kredytową i potencjalnie zwiększyć zyski banku poprzez akceptację większej liczby wniosków zachowując jednocześnie odpowiedni poziom bezpieczeństwa.

Przykład 17. Optymalny punkt odcięcia przy bardziej złożonej funkcji wypłaty

Rozważmy teraz modyfikację poprzedniego przykładu, w której wiemy dodatkowo, że w przypadku „złych" kredytów możliwe jest odzyskanie części pieniędzy gdy klient posiada zabezpieczenie. Przypuśćmy, że w zależności od zabezpieczenia strata związana z udzieleniem niespłaconego kredytu kształtuje się następująco:

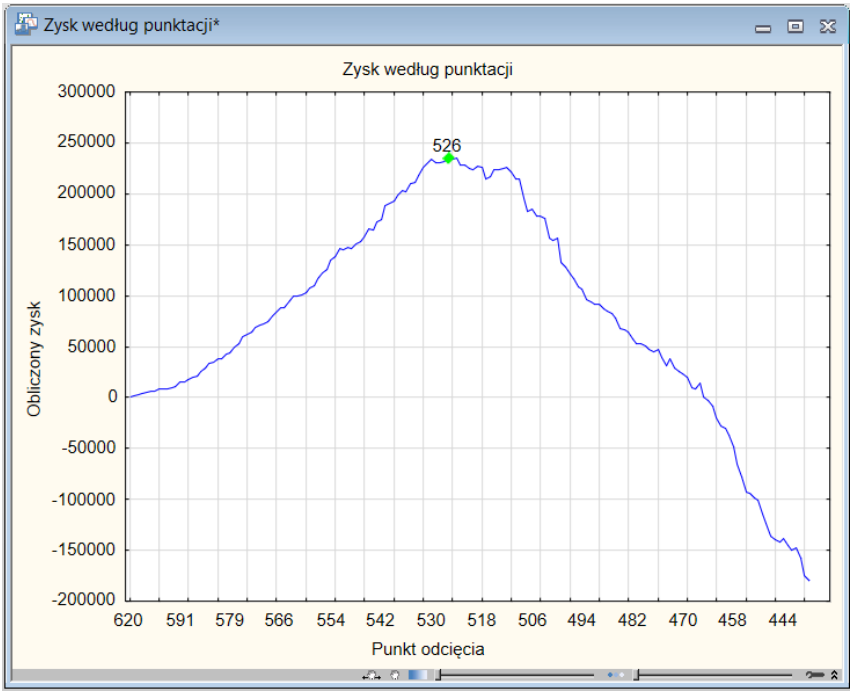
Zabezpieczenie	Strata (w % kwoty kredytu)
Brak	75
Polisa	30
Samochód	50
Nieruchomość	50

Analogicznie jak poprzednio wczytujemy model skoringowy i przechodzimy na kartę ***Zysk*** w oknie zarządzania pojedynczym punktem odcięcia. Z rozwijalnej listy wybieramy opcję ***Ogólna macierz zysków/strat***. Klikamy na przycisk ***Zmienne dodatkowe*** i wybieramy *Zabezpieczenie* jako zmienną grupującą oraz *Kwotę kredytu* jako *Wartość*. Po zatwierdzeniu wyboru zmiennych klikamy ***Ogólna macierz zysków/strat*** i wprowadzamy następujące wartości:



Wybierając opcję *Ważone* ustawiamy dla danego przypadku zysk/stratę proporcjonalną do kwoty kredytu. Gdy opcja ta zostaje niewybrana, zysk/strata ma stałą wartość (jak w opcjach *Skoring marketingowy* czy *Analiza churn*).

Klikamy ***OK*** i uruchamiamy analizę przyciskiem ***Raport***.



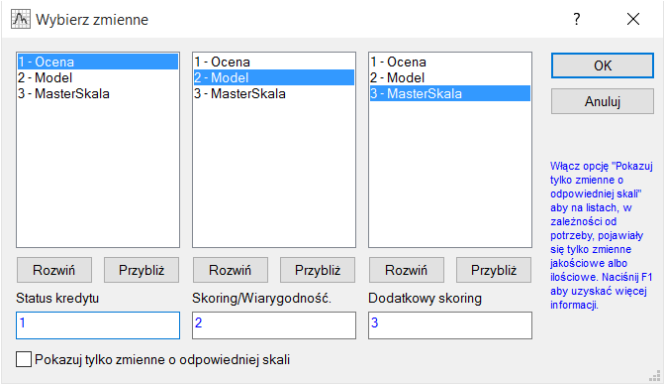
W porównaniu z wcześniejszym przykładem, w rozważnym – bardziej optymistycznym z punktu widzenia kredytodawcy – scenariuszu, optymalny punkt odcięcia przesuwa się nieznacznie – do wartości 526 – natomiast zwraca uwagę zmiana profilu wypłat oraz zakresu możliwych wartości.

Uwaga. *Poza skoringiem kredytowym oraz ogólną macierzą wypłat mamy możliwość uzyskania raportu zyskowności dla skoringu marketingowego, analizy lojalności klientów (churn) oraz fraudu.*

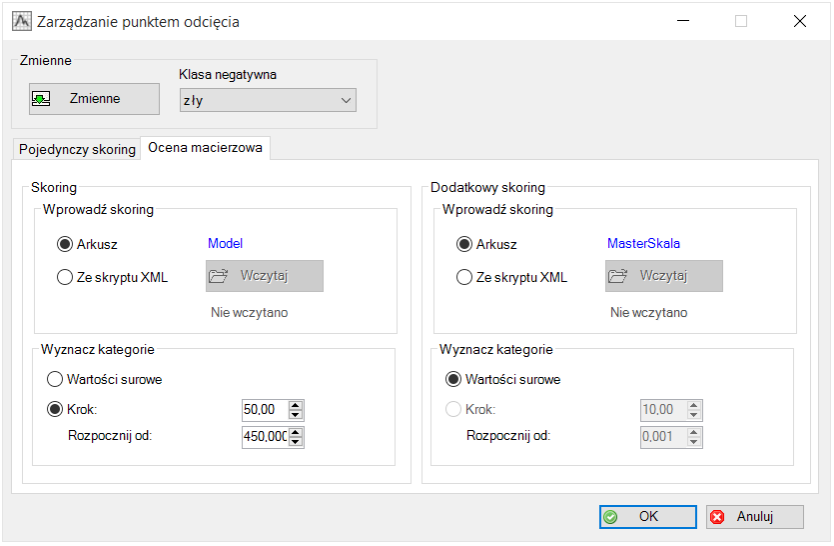
Wyznaczanie punktów odcięcia dla wyliczonego wcześniej skoringu można utożsamiać z podziałem wniosków na klasy wysokiego, średniego i niskiego ryzyka. Jeżeli zaś dysponujemy więcej niż jednym modelem, możemy utworzyć tabelę której wiersze odpowiadają segmentom ryzyka z pierwszego, a kolumny segmentom z drugiego modelu, zaś poszczególne komórki – nowym kategoriom, wyznaczonym na podstawie połączenia obydwu punktacji. W praktyce takie sytuacje mogą wystąpić np. gdy chcemy wykorzystać jednocześnie model empiryczny i ekspercki lub utworzony samodzielnie model i ocenę z zewnętrznej instytucji (BIK, agencja ratingowa itp.). Mówimy wtedy o *macierzowej ocenie* punktów odcięcia. Kolejny przykład ilustruje prowadzenie takiej kategoryzacji w Zestawie Skoringowym.

Przykład 18. Ocena macierzowa z uwzględnieniem skali wzorcowej

Otwieramy plik o nazwie *ModelAndRating.sta*, w którym znajduje się punktacja wyliczona z wewnętrznego modelu dla 1000 wniosków, a także kategoria ratingu w oparciu o skalę wzorcową. Aby przeprowadzić analizę macierzową tego zbioru danych, wybieramy ***Zarządzanie punktem odcięcia*** i po kliknięciu przycisku ***Zmienne*** wskazujemy *Ocenę* jako status kredytu, *Model* jako Skoring oraz *MasterSkala* jako drugi skoring.



Zwróćmy uwagę, że okno automatycznie przełączy się na kartę *Ocena macierzowa*. Ten rodzaj analizy wymaga kategoryzacji punktacji. Ratingi ze skali wzorcowej są już w takiej formie, natomiast dla skoringu z modelu musimy określić sposób dyskretyzacji. Zaznaczmy w polu *Skoring/Wyznacz kategorie* opcję *Krok* i wprowadzamy wartość 50 oraz 450 w polu *Rozpocznij od*.



Ogólnie przy wyznaczaniu kategorii do oceny macierzowej mamy do wyboru następujące możliwości:

- Wskazanie zmiennej w arkuszu. Jeśli jest to zmienna już skategoryzowana, możemy wybrać opcję *Wartości surowe* w grupie *Wyznacz kategorie*.
- Jeżeli zmienna z arkusza jest typu ilościowego (zawiera dokładne wartości punktacji), to dzielimy ją na przedziały o równej szerokości, którą ustawiamy w polu *Krok* po aktywowaniu tej opcji w grupie *Wyznacz kategorie*. Początek pierwszego przedziału wprowadzamy w polu *Rozpocznij od*, domyślnie jest on ustawiany jako najmniejsza wartość skoringu występująca w danych.
- Wyliczenie skoringu ze skryptu XML. W tym celu wybieramy opcję *Ze skryptu XML* i po kliknięciu *Wczytaj* wskazujemy na plik z modelem. Obliczony skoring należy również podzielić na przedziały za pomocą opcji *Krok* tak jak opisano powyżej.
- Wczytanie gotowego skryptu dyskretyzacji dla zmiennych z arkusza. Sposób tworzenia i korzystania z takich skryptów opisano dokładniej w sekcji 3.3.

Wracając do przykładu - po kliknięciu ***OK*** otrzymujemy okno analizy macierzowej, w którym możemy wyświetlić wybrane statystyki podgrup wyznaczonych przez kategorie pierwszej i drugiej skali. Dostępne są tu następujące opcje:

Liczba wniosków – ile wniosków ogółem należy do danej kategorii

Skum. liczba wniosków – jest to łączna liczba wniosków w kategorii bieżącej oraz we wszystkich kategoriach „wyższych" tzn. mających tę samą lub lepszą klasę ryzyka w obydwu badanych skalach. W praktyce wyliczana jako suma wartości w komórce tabeli ***Liczba wniosków*** odpowiadającej danej kategorii oraz wszystkich komórkach leżących od niej na prawo lub poniżej.

Liczba złych – ile wniosków należących do danej kategorii należy do klasy wskazanej w oknie definiowania analizy jako „negatywna"

Skum. liczba złych – łączna liczba wniosków z klasy negatywnej w kategorii bieżącej oraz „wyższych" (w sensie opisanym powyżej).

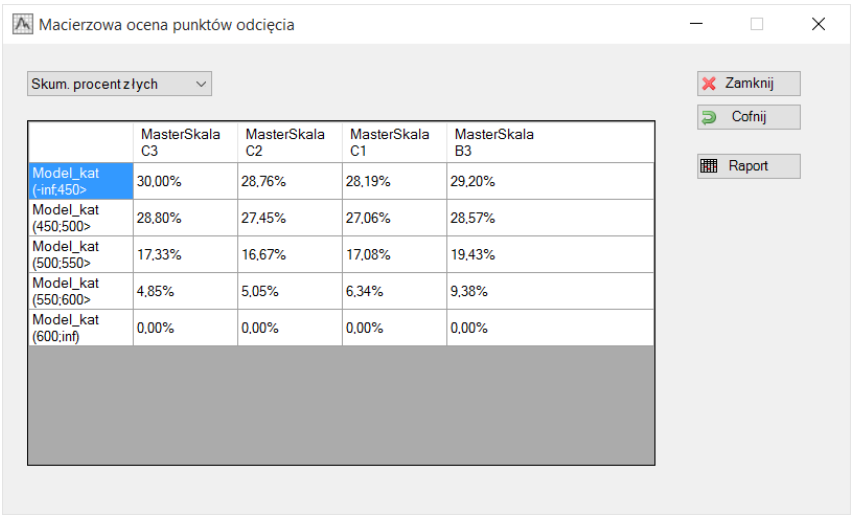
Procent złych – jaki odsetek wszystkich wniosków w danej kategorii stanowią złe; obliczany jako *(Liczba złych)/(Liczba wniosków)*

Skum. procent złych – łączny odsetek złych wniosków w kategorii bieżącej oraz wyższych; obliczany jako *(Skum. liczba złych)/(Skum. liczba wniosków)*.

Szansa złego wniosku – ile razy w danej kategorii większe jest prawdopodobieństwo *złego* niż *dobrego;* obliczane jako (*Procent złych)/(1-Procent złych*)

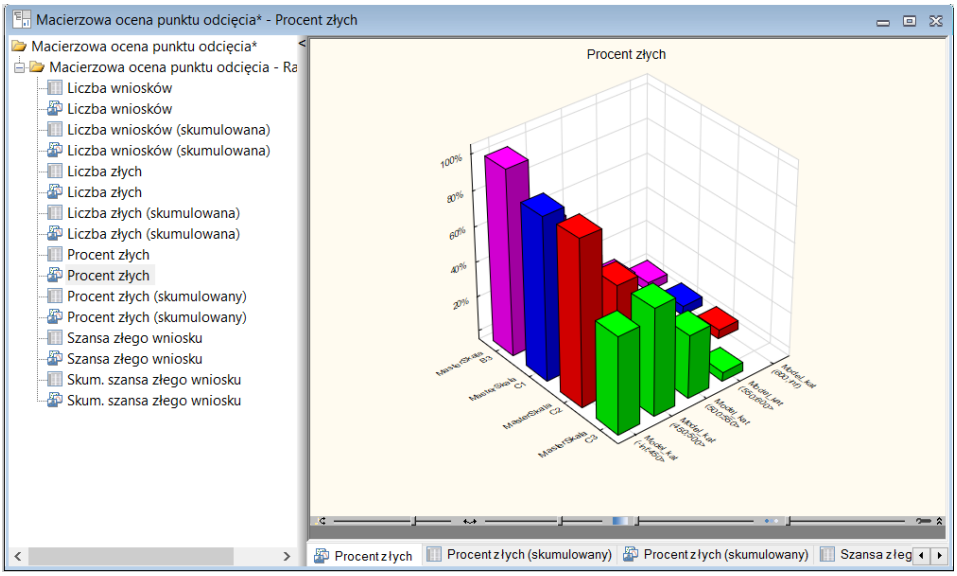
Skumulowana szansa złego – ile razy w kategorii bieżącej oraz wszystkich wyższych większe jest prawdopodobieństwo *złego* niż dobrego; obliczana jako (*Skum. procent złych)/(1-Skum. procent złych*).

Dla przykładu przedstawiono poniżej wyniki dostępne po wybraniu opcji ***Skum. procent złych***.



	MasterSkala C3	MasterSkala C2	MasterSkala C1	MasterSkala B3
Model_kat (-inf;450>	30,00%	28,76%	28,19%	29,20%
Model_kat (450;500>	28,80%	27,45%	27,06%	28,57%
Model_kat (500;550>	17,33%	16,67%	17,08%	19,43%
Model_kat (550;600>	4,85%	5,05%	6,34%	9,38%
Model_kat (600;inf)	0,00%	0,00%	0,00%	0,00%

Klikając ***Raport*** otrzymujemy skoroszyt z tabelami wszystkich przedstawionych statystyk oraz ich ilustracją w formie trójwymiarowych histogramów.



Dla lepszego wyjaśnienia znaczenia i sposobu obliczania poszczególnych statystyk, omówimy kolejne wyniki na przykładzie kategorii *Model_kat (500-550>-MasterSkala C2*:

Liczba wniosków **–** łącznie 150 wniosków otrzymało w wewnętrznym modelu punktację z przedziału 500-550 i jednocześnie klasę C2 w skali wzorcowej.

Skum. liczba wniosków – łącznie 630 wniosków otrzymało co najmniej 500 pkt. w modelu i rating C2 lub wyższy.

Liczba złych – spośród 150 wniosków w rozważanej kategorii 34 otrzymało negatywną ocenę.

Skum. liczba złych **–** spośród 630 wniosków w kategorii rozważanej oraz „wyższych” , 105 otrzymało negatywną ocene.

Procent złych **–** 34 negatywne spośród 150 wniosków to 22,67%.

Skum. procent złych - 105 negatywnych spośród wszystkich 630 wniosków to 16,67%

Szansa złego wniosku – prawdopodobieństwo *złego* wniosku przy punktacji pomiędzy 500 a 550 i ratingu C2 szacujemy na podstawie dostępnych danych jako 0,2267. Oszacowane prawdopodobieństwo wniosku *dobrego* wynosi zatem 1-0,2267=0,7733. Szansa obliczana jest jako 0,2267/0,7733=0,293. Możemy powiedzieć, że dla danej kombinacji punktacji oraz klasy ratingowej szansa *złego* wniosku jest w przybliżeniu jak 3 do 10.

Skumulowana szansa złego **–** prawdopodobieństwo *złego* wniosku przy punktacji powyżej 500 i ratingu C2 lub wyższych szacujemy na podstawie dostępnych danych jako 0,1667. Oszacowane prawdopodobieństwo wniosku *dobrego* wynosi zatem 1-0,1667=0,8333. Szansa obliczana jest jako 0,1667/0,8333=0,053. Możemy powiedzieć, że dla danej kombinacji punktacji oraz klasy ratingowej szansa *złego* wniosku jest w przybliżeniu jak 5 do 100.

5.4. Testy kalibracji

Testy kalibracji umożliwiają testowanie zgodności realizacji ryzyka w poszczególnych grupach ratingowych ze zdefiniowaną skalą wzorcową (masterskalą). W zależności od liczności kredytów w poszczególnych grupach ratingowych wykonywany jest test dwumianowy bądź test normalny – wyboru testu można dokonać ręcznie bądź skorzystać z zaimplementowanych wytycznych austriackiego nadzoru bankowego. W celu ułatwienia interpretacji uzyskanych wyników wprowadzono strategię *traffic light approach*.

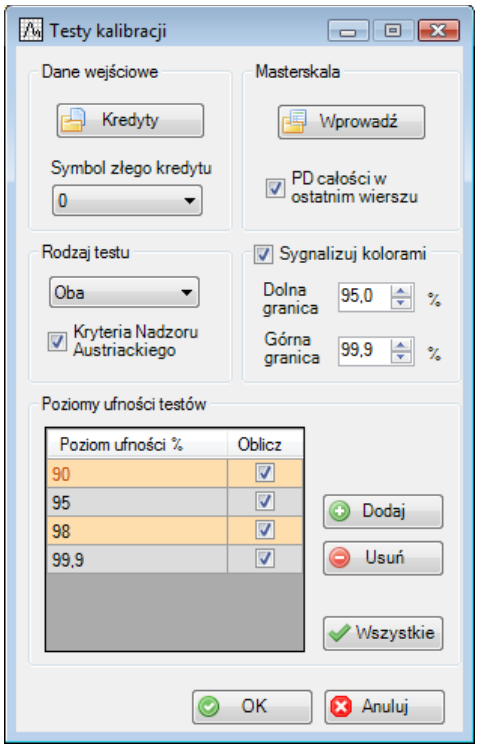
Wyboru odpowiedniego testu można dokonać na podstawie wytycznych austriackiego nadzoru bankowego (Oesterreichische Nationalbank, 2004). Jeżeli dana klasa spełnia poniższe kryteria wybierany jest test normalny, w przeciwnym przypadku test dwumianowy.

Maksymalna wartość PD	Minimalna liczność klasy
0.10%	9,010
0.25%	3,610
0.50%	1,810
1.00%	910
2.00%	460
3.00%	310
5.00%	190
10.00%	101
20.00%	57
50.00%	37

Strategia *traffic light approach* pozwala przypisać kolory: Czerwony, Żółty oraz Zielony poszczególnym poziomom ryzyka niedoszacowania parametru PD. Po określeniu granicznych poziomów ufności podejście to pozwala za pomocą kolorów poinformować użytkownika o stopniu ryzyka niedoszacowania.

Przykład 19. Wykonanie testów kalibracji

Aby wykonać testy kalibracji z katalogu *Testy kalibracji* otwieramy dwa zbiory danych: zbiór ze skalą wzorcową (masterskalą) – plik *MasterSkala.sta* oraz zbiór z kredytami – plik *Kredyty.sta*. Po otwarciu powyższych plików z menu ***Zestaw Skoringowy*** z grupy ***Ocena i kalibracja*** wybieramy polecenie ***Testy kalibracji***. Okno to udostępnia następujące opcje pozwalające parametryzować wykonanie analizy:



Przycisk ***Kredyty*** umożliwia wskazanie i wczytanie pliku zawierającego informację o ratingu danego kredytu oraz jego statusie (dobry/zły).

Symbol złego kredytu umożliwia wskazanie jednej z klas dychotomicznej zmiennej reprezentującej status jako symbol złego kredytu.

Przycisk ***Wprowadź*** znajdujący się w obszarze ***Masterskala*** pozwala na wskazanie i wczytanie pliku ze skalą wzorcową zawierającą informację o symbolu klasy ratingowej oraz oczekiwanym prawdopodobieństwie niezrealizowania zobowiązania PD (*Probability of default*) przypisanym do każdej klasy.

PD całości w ostatnim wierszu określa czy w pliku ze skalą wzorcową w ostatnim wierszu zamieszczono informacje o prawdopodobieństwie niezrealizowania zobowiązania kredytowego PD dla całego portfela kredytów.

Rodzaj testu pozwala na wybór jednego z dwóch testów: ***Normalnego*** lub ***Dwumianowego***. Opcja ***Oba*** wykona niezależnie obydwa testy.

> ***Kryteria Nadzoru Austriackiego*** dokona automatycznej selekcji odpowiedniego testu dla każdej z klas ratingowych na podstawie przedstawionych powyżej kryteriów. Opcja ta jest dostępna, jeżeli na liście rozwijalnej ***Rodzaj testu*** wybrano opcję ***Oba***.

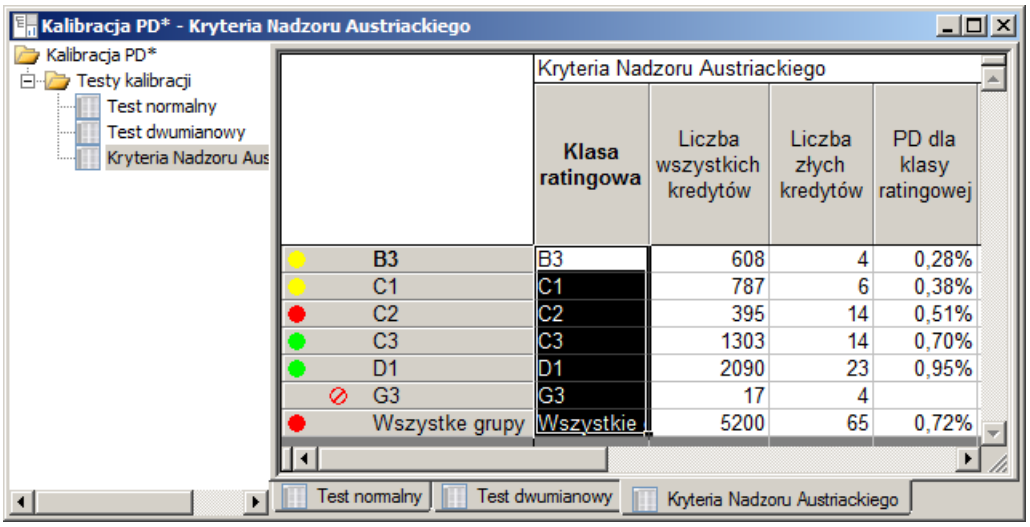
Sygnalizuj kolorami pozwala oceniać zgodność realizacji ryzyka w poszczególnych grupach ratingowych ze zdefiniowaną skalą wzorcową (masterskalą) zgodnie ze strategią opisaną powyżej.

> ***Dolna granica*** oraz ***Górna granica*** określają granice pomiędzy statusami ***Czerwonym***, ***Żółtym*** i ***Zielonym***. Wprowadzone wartości muszą być zdefiniowane w tablicy ***Poziomy ufności testów***.

Poziomy ufności testów umożliwiają zdefiniowanie, usuwanie oraz zmianę poziomów ufności testów wykonywanych podczas testowania kalibracji. Obliczenia zostaną wykonane dla każdego poziomu, dla którego zaznaczono opcję ***Oblicz***.

Po otwarciu okna ***Testy kalibracji*** klikamy przycisk ***Kredyty*** znajdujący się w obszarze ***Dane wejściowe***, wskazujemy zbiór *Kredyty.sta* oraz wybieramy zmienne zgodnie z opisem okna ***Wybierz zmienne***. Po wczytaniu pliku na liście rozwijalnej ***Symbol złego kredytu*** wybieramy symbol *1*. Następnie wprowadzamy plik z masterskalą (*MasterSkala.sta*) za pomocą przycisku ***Wprowadź*** z obszaru ***Masterskala***.

W obszarze ***Rodzaj testu*** wybieramy opcję ***Oba*** oraz zaznaczamy opcję ***Kryteria Nadzoru Austriackiego***, aby program automatycznie dobrał odpowiedni rodzaj testu dla poszczególnych grup ratingowych. Aby włączyć strategię *traffic light approach* zaznaczamy opcję ***Sygnalizuj kolorami***. Klikamy *OK* aby wygenerować zestaw raportów końcowych.



Na podstawie uzyskanych wyników możemy ocenić występowanie niedoszacowania wartości PD zapisanego w masterskali.

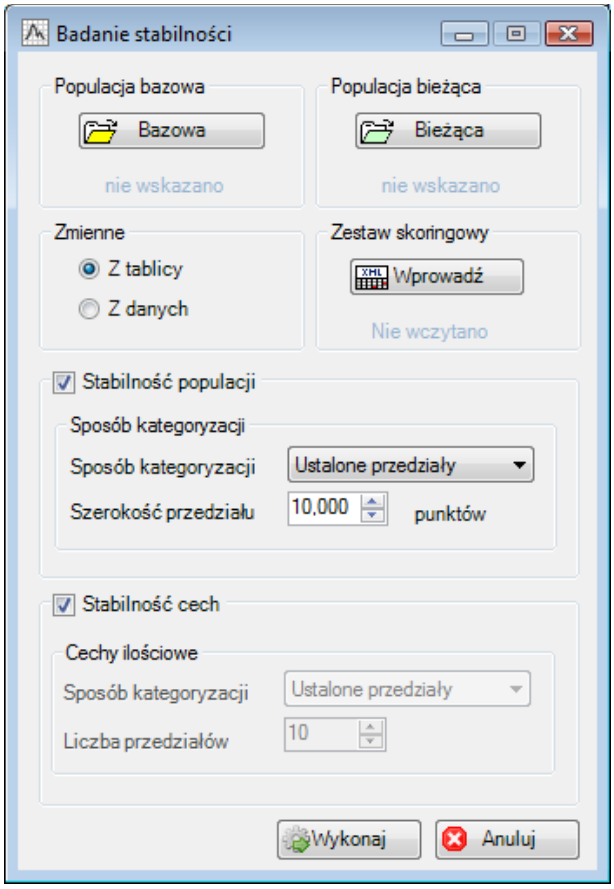
6. Monitoring

6.1. Stabilność populacji

Moduł ***Stabilność populacji*** umożliwia porównanie dwóch zbiorów danych (na przykład zbioru aktualnych i historycznych kredytobiorców) pod kątem różnic w strukturze wartości poszczególnych cech oraz w strukturze samego skoringu. Raport stabilności może pokazać, czy nastąpiła zmiana populacji przychodzącej i jak głęboka jest ta zmiana. Bardzo silne różnice w porównaniu z populacją bazową sugerują potrzebę ponownej estymacji modelu.

Przykład 20. – Badanie stabilności populacji

Moduł ***Stabilność populacji*** uruchamiamy poleceniem ***Monitoring*** | ***Stabilność populacji*** znajdującym się w menu ***Zestaw skoringowy***. Po wyborze tej opcji, wyświetlone zostanie okno ***Badanie stabilności***.



Uwaga. *Raport* ***Stabilność populacji*** *dostarcza ogólnych informacji o zmianach w populacji obecnie obsługiwanych klientów. Korzystając z raportu* ***Stabilność cech*** *można dokonać bardziej wnikliwej analizy identyfikując, które z atrybutów danych zmiennych mają największy wpływ na zmianę stabilności.*

W oknie tym możemy parametryzować analizę stabilności za pomocą następujących opcji:

Przycisk ***Bazowa*** pozwala na wskazanie arkusza zawierającego przypadki z populacji bazowej.

Przycisk ***Bieżąca*** pozwala na wskazanie arkusza zawierającego przypadki z populacji bieżącej.

Obszar ***Zmienne*** zawiera opcje pozwalające na określenie sposobu wyboru zmiennych do analizy.

> ***Z tablicy*** – zmienne, które będą brane pod uwagę w analizie zostaną zdefiniowane na podstawie modelu skoringowego zapisanego w pliku XML. Plik XML z modelem otwieramy za pomocą przycisku ***Wczytaj***. Wybranie tej opcji powoduje dezaktywację opcji dotyczących raportu stabilności cech. Wynika to z faktu, że kategorie poszczególnych zmiennych są jednoznacznie określone w modelu XML.
>
> ***Z danych*** – zmienne zawarte w arkuszu bazowym zostaną sparowane ze zmiennymi ze zbioru bieżącego. Za pomocą przycisku ***Zmienne*** konieczna będzie selekcja zmiennych.

Stabilność populacji – zaznaczenie tej opcji umożliwi wykonanie analizy stabilności populacji za pomocą następujących parametrów:

> ***Sposób kategoryzacji – Ustalone przedziały*** wybranie tej opcji powoduje podzielenie wartości skoringu na przedziały o szerokości wskazanej w opcji ***Szerokość przedziału***.
>
> ***Sposób kategoryzacji – Poszczególne wartości*** –wybranie tej opcji spowoduje, że wartości skoringu nie będą kategoryzowane przed wykonaniem obliczeń.

Szerokość przedziału oznacza szerokość przedziału przyjętą podczas kategoryzacji punktacji, gdy jako sposób kategoryzacji wybrano ***Ustalone przedziały***.

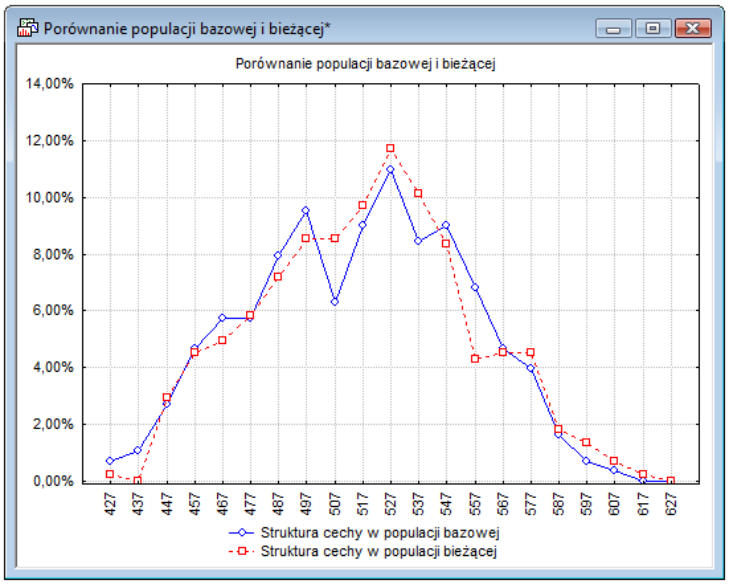
Stabilność cech – zaznaczenie tej opcji umożliwi wykonanie analizy stabilności cech. Jeżeli w obszarze ***Zmienne*** wybrano opcję ***Z danych*** dostępne są następujące parametry:

Sposób kategoryzacji – Ustalone przedziały wybranie tej opcji powoduje podzielenie wartości predyktorów ilościowych na równoliczne przedziały. Liczba przedziałów określana jest w opcji ***Liczba przedziałów***.

Sposób kategoryzacji – Poszczególne wartości – wybranie tej opcji spowoduje, że wartości predyktorów ilościowych nie będą kategoryzowane przed wykonaniem obliczeń.

Liczba przedziałów określa, na jaką liczbę przedziałów będą dzielone predyktory ilościowe przed wykonaniem obliczeń. Opcja ta ma odniesienie, gdy jako sposób kategoryzacji wybrano ***Ustalone przedziały***.

W celu wykonania analizy, w pierwszej kolejności wskazujemy zbiory danych z populacją bazową (*Bazowa.sta)* oraz bieżącą (*Obecna.sta)*. Następnie należy wskazać sposób wyboru zmiennych, które mają być analizowane. Wybieramy opcję ***Z tablicy***, a następnie wczytujemy plik XML z modelem – możemy wykorzystać model skoringowy *krokowa wsteczna.xml* (plik XML z modelem znajduje się w katalogu z plikami przykładowymi). Pozostałe opcje pozostawiamy na domyślnym poziomie. Zatwierdzamy wykonanie analizy za pomocą przycisku ***Wykonaj***. W wyniku działania programu otrzymujemy skoroszyt z raportem stabilności populacji i/lub cech.



Dane: Stabilność populacji (8 zmn. * 22 prz.)

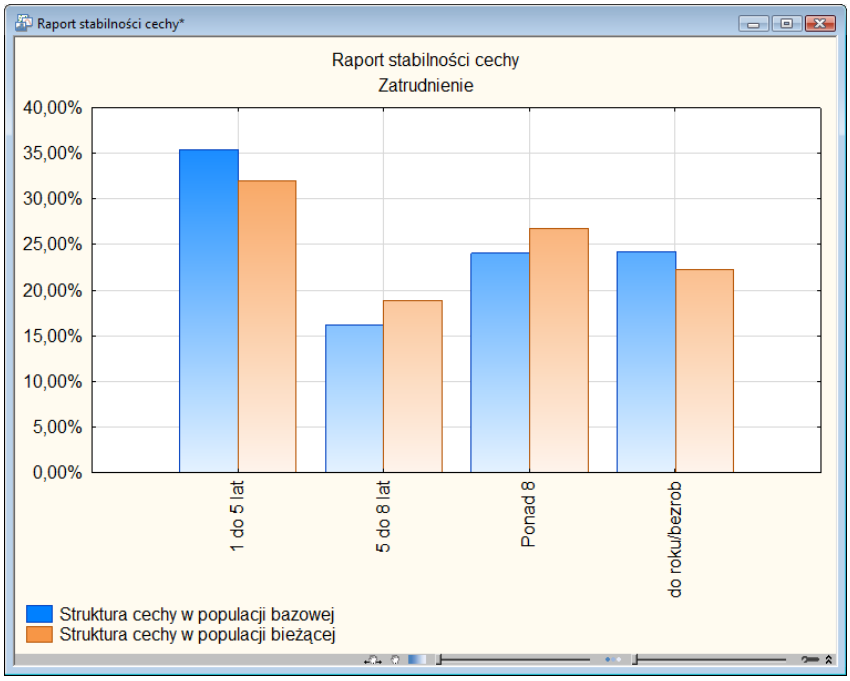
Od	Do	Liczba przypadków dla populacji bazowej	Liczba przypadków dla populacji bieżącej	Struktura cechy w populacji bazowej	Struktura cechy w populacji bieżącej	Zmiana struktury	Wskaźnik zmiany	Weight of evidence	Struktura wskaźnika stabilności populacji
427,0000<=x<437,0000		4	1	0,72%	0,23%	-0,494%	0,313	-1,161	0,006
437,0000<=x<447,0000		6	0	1,08%	0,00%	-1,079%	0	0	0
447,0000<=x<457,0000		15	13	2,70%	2,93%	0,23%	1,085	0,082	0
457,0000<=x<467,0000		26	20	4,68%	4,50%	-0,172%	0,963	-0,037	0
467,0000<=x<477,0000		32	22	5,76%	4,95%	-0,8%	0,861	-0,15	0,001
477,0000<=x<487,0000		32	26	5,76%	5,86%	0,1%	1,017	0,017	0
487,0000<=x<497,0000		44	32	7,91%	7,21%	-0,706%	0,911	-0,094	0,001
497,0000<=x<507,0000		53	38	9,53%	8,56%	-0,974%	0,898	-0,108	0,001
507,0000<=x<517,0000		35	38	6,29%	8,56%	2,264%	1,36	0,307	0,007
517,0000<=x<527,0000		50	43	8,99%	9,68%	0,692%	1,077	0,074	0,001
527,0000<=x<537,0000		61	52	10,97%	11,71%	0,74%	1,067	0,065	0
537,0000<=x<547,0000		47	45	8,45%	10,14%	1,682%	1,199	0,181	0,003
547,0000<=x<557,0000		50	37	8,99%	8,33%	-0,659%	0,927	-0,076	0,001
557,0000<=x<567,0000		38	19	6,83%	4,28%	-2,555%	0,626	-0,468	0,012
567,0000<=x<577,0000		26	20	4,68%	4,50%	-0,172%	0,963	-0,037	0
577,0000<=x<587,0000		22	20	3,96%	4,50%	0,548%	1,138	0,13	0,001
587,0000<=x<597,0000		9	8	1,62%	1,80%	0,183%	1,113	0,107	0
597,0000<=x<607,0000		4	6	0,72%	1,35%	0,632%	1,878	0,63	0,004
607,0000<=x<617,0000		2	3	0,36%	0,68%	0,316%	1,878	0,63	0,002
617,0000<=x<627,0000		0	1	0,00%	0,23%	0,225%	0	0	0
627,0000<=x<637,0000		0	0	0,00%	0,00%	0,%	0	0	0
Suma		556	444	100,00%	100,00%				**0,039**

Szczegóły obliczeniowe dotyczące wskaźników stabilności możemy znaleźć w dokumencie Scorecard Formula Guide. Możemy interpretować wskaźnik stabilności populacji w następujący sposób:

- Poniżej 0,1 – nieznaczące przesunięcie
- 0,1 – 0,25 – małe przesunięcie – do dalszego monitorowania
- Powyżej 0,25 – znaczące przesunięcie

Zarówno wykres jak i raport stabilności populacji sugerują, że zmiana jest niewielka. Obliczony wskaźnik stabilności populacji jest na poziomie 0,039.

Moduł stabilności populacji umożliwia także analizę stabilności zmiennych wykorzystanych przy budowie modelu. Dla każdej zmiennej tworzony jest zestaw raportów. Możemy uważać przesunięcie na poziomie poszczególnych cech za nieistotne, jeżeli wartość bezwzględna wskaźnika stabilności cech jest mniejsza od 5. Przykładowy wykres dla zmiennej *Zatrudnienie* prezentowany jest poniżej.



6.2. Analiza Vintage

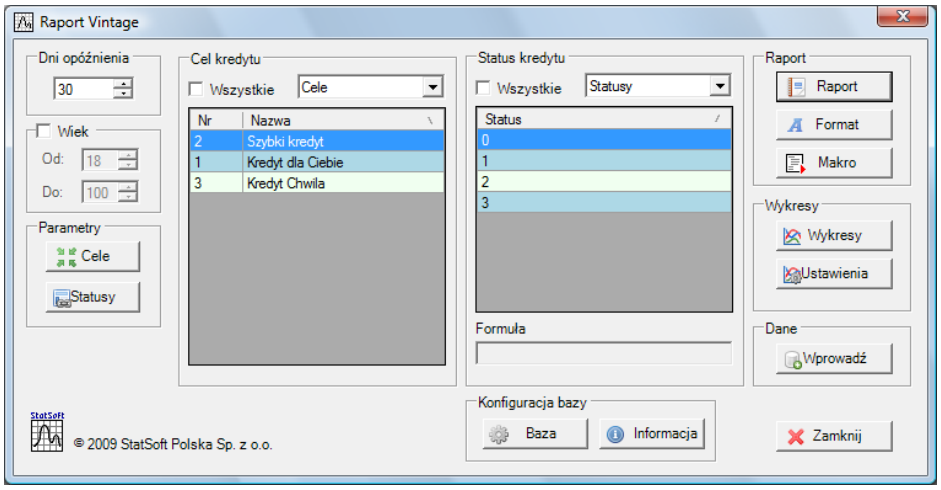
Moduł ***Analiza Vintage*** (wykorzystywany przede wszystkim w ryzyku kredytowym) pozwala na monitorowanie stanu portfela kredytów w kolejnych miesiącach spłaty. Umożliwia przygotowanie raportu w zależności od celu kredytów, ich statusu, liczby dni przeterminowania oraz wieku kredytobiorców. Raporty w postaci tabelarycznej są uzupełnione zestawem wykresów pozwalających na łatwiejszy monitoring portfela kredytów i interpretację zachodzących zmian.

Uwaga. *Do uruchomienia modułu Vintage wymagana jest baza danych zawierająca dane o stanie portfela kredytów w kolejnych miesiącach spłaty. Korzystanie z tego modułu wymaga dodatkowej usługi dostosowania procesu ETL do indywidualnych wymagań klienta (możliwa jest integracja modułu z dowolnym systemem bazodanowym).*

Poniższy przykład ilustruje możliwości modułu dla przygotowanej wcześniej bazy danych.

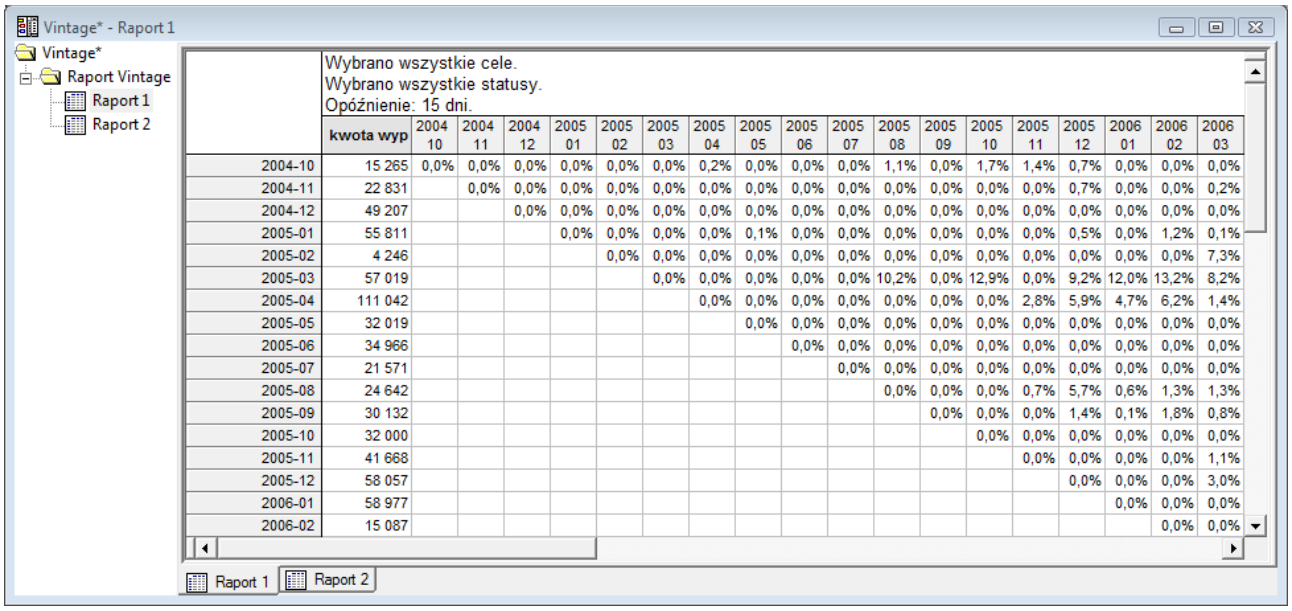
Przykład 21. – Analiza Vintage

Moduł ***Analiza Vintage*** uruchamiamy poleceniem ***Monitoring*** | ***Analiza Vintage*** znajdującym się w menu ***Zestaw skoringowy***.



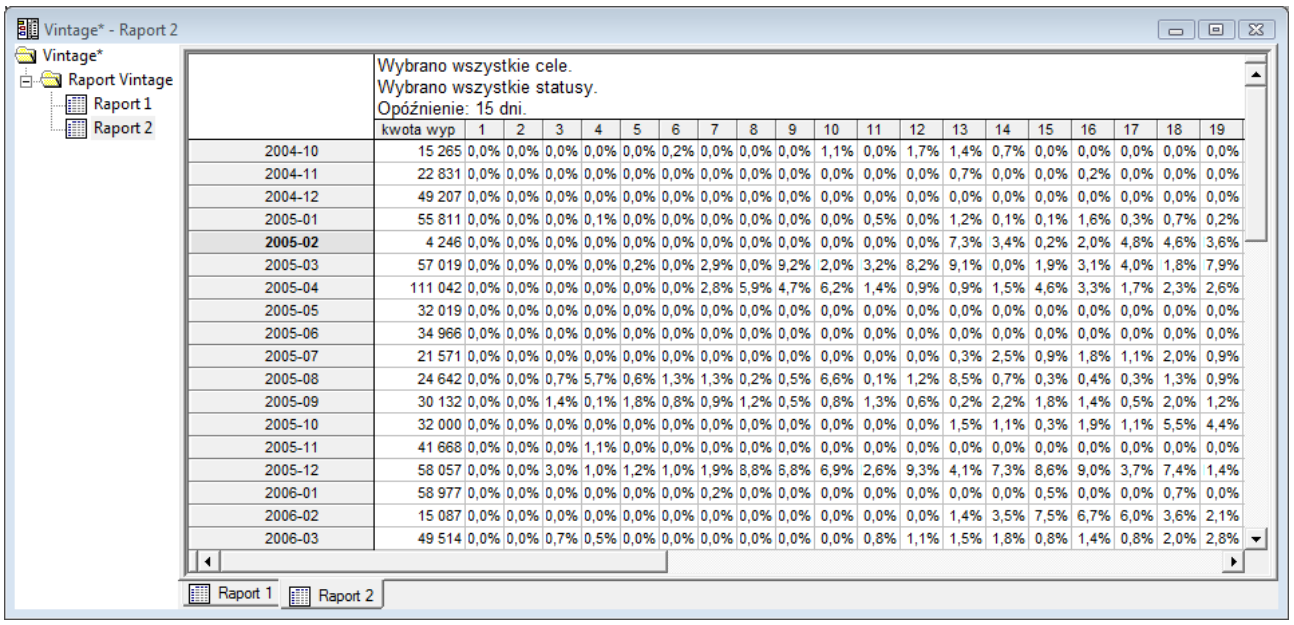
Uwaga. *W oknie Raport Vintage mamy możliwość parametryzacji analizowanych kredytów ze względu na cel kredytu oraz jego status (dodatkowo za pomocą przycisku Cele oraz Statusy możemy grupować poszczególne pozycje), a także określenia wieku kredytobiorcy oraz liczby dni opóźnienia w spłacie, jaką uznamy za odstępstwo od umowy.*

Zaznaczamy pole wyboru w obszarze ***Wiek***, a następnie klikamy przycisk ***Raport*** tworząc tym samym dwa raporty *Vintage*. W raporcie pierwszym w zmiennej *kwota wyp* zawarte są sumaryczne kwoty kredytów wypłacone w kolejnych miesiącach. Wartości w wypełnionych komórkach informują jaki procent wypłaconej kwoty stanowią kredyty przeterminowane (według przyjętej wartości opóźnienia) w danym miesiącu kalendarzowym.



	kwota wyp	2004 10	2004 11	2004 12	2005 01	2005 02	2005 03	2005 04	2005 05	2005 06	2005 07	2005 08	2005 09	2005 10	2005 11	2005 12	2006 01	2006 02	2006 03
2004-10	15 265	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,2%	0,0%	0,0%	0,0%	1,1%	0,0%	1,7%	1,4%	0,7%	0,0%	0,0%	0,0%
2004-11	22 831		0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,7%	0,0%	0,0%	0,2%
2004-12	49 207			0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2005-01	55 811				0,0%	0,0%	0,0%	0,0%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,5%	0,0%	1,2%	0,1%
2005-02	4 246					0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	7,3%
2005-03	57 019						0,0%	0,0%	0,0%	0,0%	0,0%	10,2%	0,0%	12,9%	0,0%	9,2%	12,0%	13,2%	8,2%
2005-04	111 042							0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	2,8%	5,9%	4,7%	6,2%	1,4%
2005-05	32 019								0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2005-06	34 966									0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2005-07	21 571										0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2005-08	24 642											0,0%	0,0%	0,0%	0,7%	5,7%	0,6%	1,3%	1,3%
2005-09	30 132												0,0%	0,0%	0,0%	1,4%	0,1%	1,8%	0,8%
2005-10	32 000													0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2005-11	41 668														0,0%	0,0%	0,0%	0,0%	1,1%
2005-12	58 057															0,0%	0,0%	0,0%	3,0%
2006-01	58 977																0,0%	0,0%	0,0%
2006-02	15 087																	0,0%	0,0%

Raport 2 zawiera analogiczne zestawienie z tą różnicą, że wartości w wypełnionych komórkach informują, jaki procent wypłaconej kwoty stanowią kredyty przeterminowane w kolejnych miesiącach począwszy od daty wypłacenia kredytu.



	kwota wyp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2004-10	15 265	0,0%	0,0%	0,0%	0,0%	0,0%	0,2%	0,0%	0,0%	0,0%	1,1%	0,0%	1,7%	1,4%	0,7%	0,0%	0,0%	0,0%	0,0%	0,0%
2004-11	22 831	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,7%	0,0%	0,0%	0,2%	0,0%	0,0%	0,0%
2004-12	49 207	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2005-01	55 811	0,0%	0,0%	0,0%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,5%	0,0%	1,2%	0,1%	0,1%	1,6%	0,3%	0,7%	0,2%
2005-02	4 246	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	7,3%	3,4%	0,2%	2,0%	4,8%	4,6%	3,6%
2005-03	57 019	0,0%	0,0%	0,0%	0,0%	0,2%	0,0%	2,9%	0,0%	9,2%	2,0%	3,2%	8,2%	9,1%	0,0%	1,9%	3,1%	4,0%	1,8%	7,9%
2005-04	111 042	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	2,8%	5,9%	4,7%	6,2%	1,4%	0,9%	0,9%	1,5%	4,6%	3,3%	1,7%	2,3%	2,6%
2005-05	32 019	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2005-06	34 966	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2005-07	21 571	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,3%	2,5%	0,9%	1,8%	1,1%	2,0%	0,9%
2005-08	24 642	0,0%	0,0%	0,7%	5,7%	0,6%	1,3%	1,3%	0,2%	0,5%	6,6%	0,1%	1,2%	8,5%	0,7%	0,3%	0,4%	0,3%	1,3%	0,9%
2005-09	30 132	0,0%	0,0%	1,4%	0,1%	1,8%	0,8%	0,9%	1,2%	0,5%	0,8%	1,3%	0,6%	0,2%	2,2%	1,8%	1,4%	0,5%	2,0%	1,2%
2005-10	32 000	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	1,5%	1,1%	0,3%	1,9%	1,1%	5,5%	4,4%
2005-11	41 668	0,0%	0,0%	0,0%	1,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
2005-12	58 057	0,0%	0,0%	3,0%	1,0%	1,2%	1,0%	1,9%	8,8%	8,8%	6,9%	2,6%	9,3%	4,1%	7,3%	8,6%	9,0%	3,7%	7,4%	1,4%
2006-01	58 977	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,2%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,5%	0,0%	0,0%	0,7%	0,0%
2006-02	15 087	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	1,4%	3,5%	7,5%	6,7%	6,0%	3,6%	2,1%
2006-03	49 514	0,0%	0,0%	0,7%	0,5%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,8%	1,1%	1,5%	1,8%	0,8%	1,4%	0,8%	2,0%	2,8%

Dodatkowo klikając przycisk ***Wykresy*** możemy wygenerować wykresy opisujące wskaźnik przeterminowania kredytów o określonym okresie wypłaty, bądź w danym miesiącu spłaty, a także średni wskaźnik przeterminowania. Za pomocą przycisku ***Ustawienia*** możemy określać parametry generowanych wykresów.

6.3. Macierze migracji

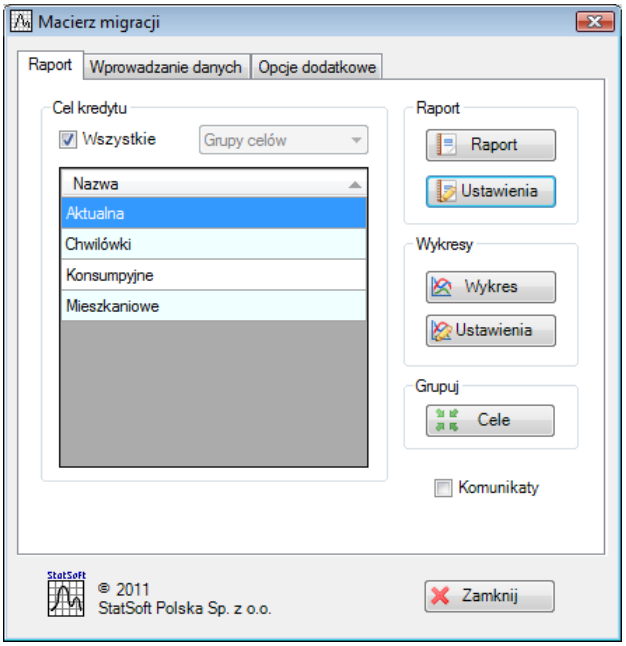
Moduł ***Macierze migracji*** pozwala na utworzenie raportów opisujących strukturę portfela oraz macierzy migracji dla wskazanego punktu startowego. Raporty w postaci tabelarycznej są uzupełnione zestawem wykresów przedstawiających zmiany przeterminowania w zależności od miesiąca obserwacji oraz portfela.

Uwaga. *Do uruchomienia modułu Macierze migracji wymagana jest baza danych zawierająca dane o stanie portfela kredytów w kolejnych miesiącach spłaty. Korzystanie z tego modułu wymaga dodatkowej usługi dostosowania procesu ETL do indywidualnych wymagań klienta. (możliwa jest integracja modułu z dowolnym systemem bazodanowym).*

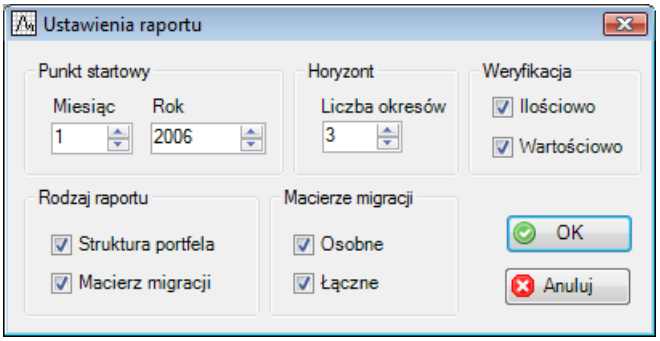
Poniższy przykład ilustruje możliwości modułu dla przygotowanej wcześniej bazy danych.

Przykład 22. – Macierze migracji

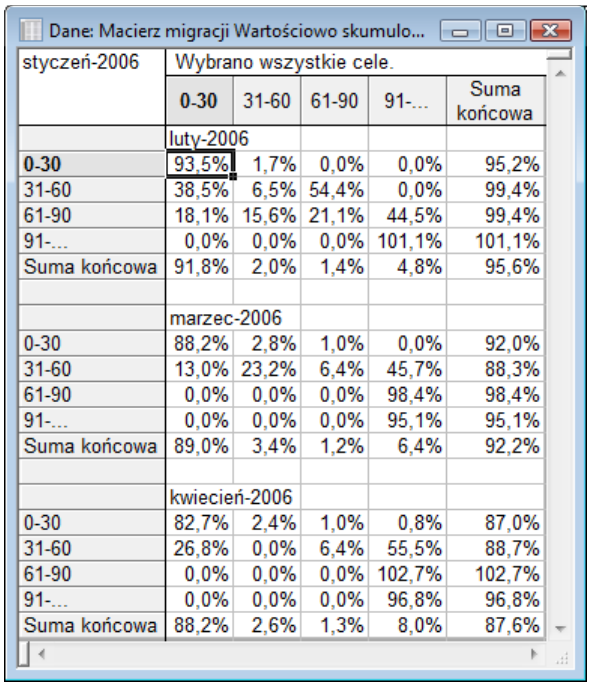
Moduł ***Macierze migracji*** uruchamiamy poleceniem ***Monitoring*** | ***Macierze migracji*** znajdującym się w menu ***Zestaw skoringowy***.



Na karcie ***Raport*** mamy możliwość wybrania celu kredytów, dla którego chcemy obliczyć macierze migracji, dodatkowe parametry raportu określamy klikając przycisk ***Ustawienia***, w oknie ***Ustawienia raportu***.



Mamy możliwość wygenerowania zarówno raportu migracji jak i przedstawić strukturę portfela kredytów, zarówno w ujęciu ilościowym jak i wartościowym. Poniższy rysunek przedstawia przykładową macierz migracji.



Dane: Macierz migracji Wartościowo skumulo...

styczeń-2006	Wybrano wszystkie cele.				
	0-30	31-60	61-90	91-...	Suma końcowa
	luty-2006				
0-30	93,5%	1,7%	0,0%	0,0%	95,2%
31-60	38,5%	6,5%	54,4%	0,0%	99,4%
61-90	18,1%	15,6%	21,1%	44,5%	99,4%
91-...	0,0%	0,0%	0,0%	101,1%	101,1%
Suma końcowa	91,8%	2,0%	1,4%	4,8%	95,6%
	marzec-2006				
0-30	88,2%	2,8%	1,0%	0,0%	92,0%
31-60	13,0%	23,2%	6,4%	45,7%	88,3%
61-90	0,0%	0,0%	0,0%	98,4%	98,4%
91-...	0,0%	0,0%	0,0%	95,1%	95,1%
Suma końcowa	89,0%	3,4%	1,2%	6,4%	92,2%
	kwiecień-2006				
0-30	82,7%	2,4%	1,0%	0,8%	87,0%
31-60	26,8%	0,0%	6,4%	55,5%	88,7%
61-90	0,0%	0,0%	0,0%	102,7%	102,7%
91-...	0,0%	0,0%	0,0%	96,8%	96,8%
Suma końcowa	88,2%	2,6%	1,3%	8,0%	87,6%

Dodatkowo klikając przycisk ***Wykres*** możemy wygenerować wykresy opisujące zmiany przeterminowania w zależności od miesiąca obserwacji oraz w zależności od portfela w wybranym miesiącu. Za pomocą przycisku ***Ustawienia*** możemy określać parametry generowanych wykresów.

7. Wykorzystanie przestrzeni roboczych Statistica Data Miner

W środowisku Statistica Data Miner pracujemy w specjalnej tzw. **przestrzeni roboczej** - cały projekt i przepływ danych między kolejnymi etapami analizy reprezentowany jest graficznie.

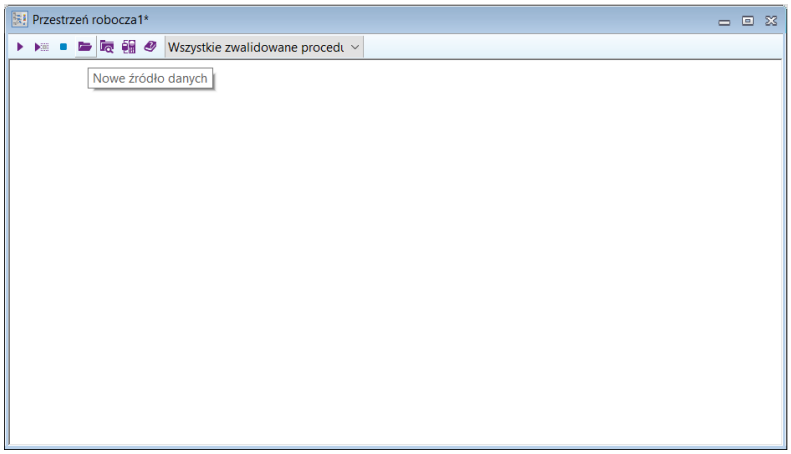
Każda procedura przetwarzająca dane reprezentowana jest przez ikonę (tzw. *węzeł*). Przepływ danych obrazują strzałki łączące poszczególne węzły. Węzły zaprojektowane są tak, że dane wypływające z jednego z węzłów mogą stanowić wejście dla innych węzłów. Umożliwia to składanie projektu analizy z poszczególnych węzłów niczym z klocków.

W postaci węzłów dostępne jest wiele procedur przekształcania, analizy i wizualizacji danych. Wszystkie węzły dostępne są w *Przeglądarce węzłów*, która ma uporządkowaną hierarchiczną strukturę. W systemie zdefiniowano wiele przeglądarek, zawierających tylko te węzły, które są potrzebne do wykonywania określonych zadań. Użytkownik może stworzyć własną przeglądarkę, z tymi narzędziami, z których najczęściej korzysta.

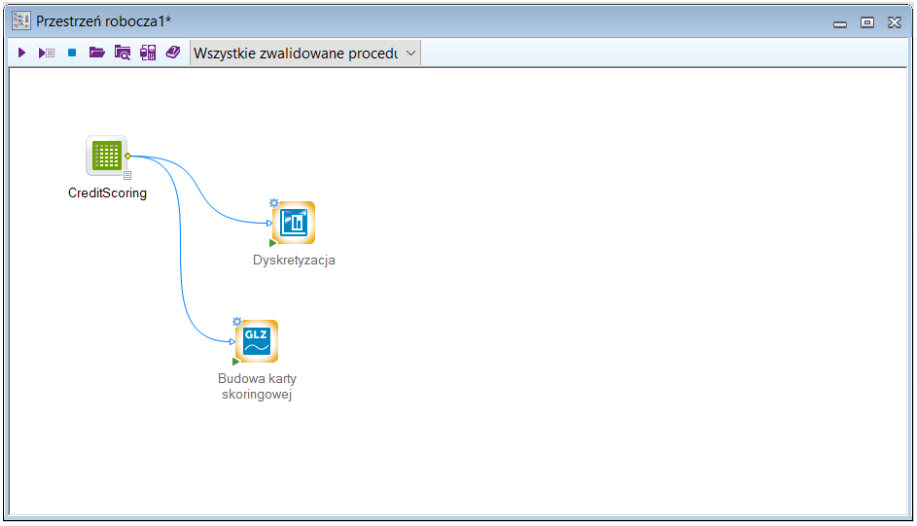
Przeciągając węzły z *Przeglądarki węzłów* do przestrzeni roboczej i łącząc je ze sobą strzałkami, otrzymujemy kompletną ścieżkę przygotowania i analizy danych. Dzięki takiemu trybowi pracy nawet skomplikowaną, wieloetapową analizę możemy łatwo zbudować i modyfikować, przeciągając obiekty myszką. Ponadto łatwo jest zorientować się w strukturze projektu.

Przykład 23. – Wykorzystanie przestrzeni roboczej

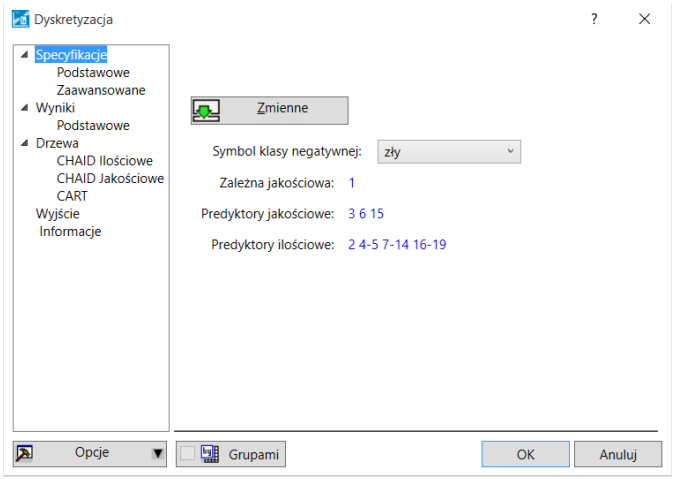
Odtworzymy analizę i przygotowanie danych z przykładów 4 oraz 7 za pomocą węzłów Data Miner. W tym celu ponownie wykorzystamy plik *CreditScoring.sta*. Z głównego menu Statistica wybieramy ***Nowy/Przestrzeń robocza*** i klikamy ***OK*** aby utworzyć pustą przestrzeń roboczą. W górnej części okna z przestrzenią wybieramy ***Nowe źródło danych*** i wskazujemy plik *CreditScoring*. W przestrzeni pojawia się symbol źródła danych.



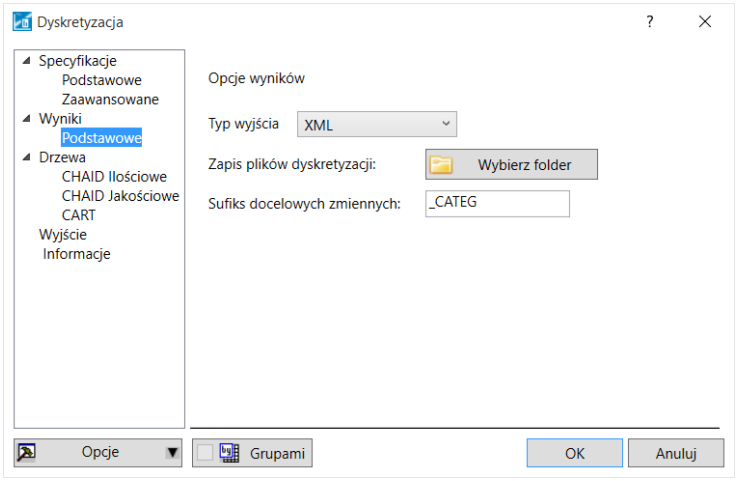
Obok opcji ***Nowe źródło danych*** znajduje się przycisk ***Przeglądarka węzłów***. Klikamy go aby uzyskać dostęp do możliwych do wykorzystania węzłów Data Miner. Z grupy ***Zestaw Skoringowy*** wybieramy węzły ***Dyskretyzacja*** i ***Budowa karty skoringowej***, które umieszczamy w przestrzeni klikając przycisk ***Wstaw*** przeglądarki. Łączymy źródło danych z oboma węzłami, otrzymując przestrzeń jak na rysunku:



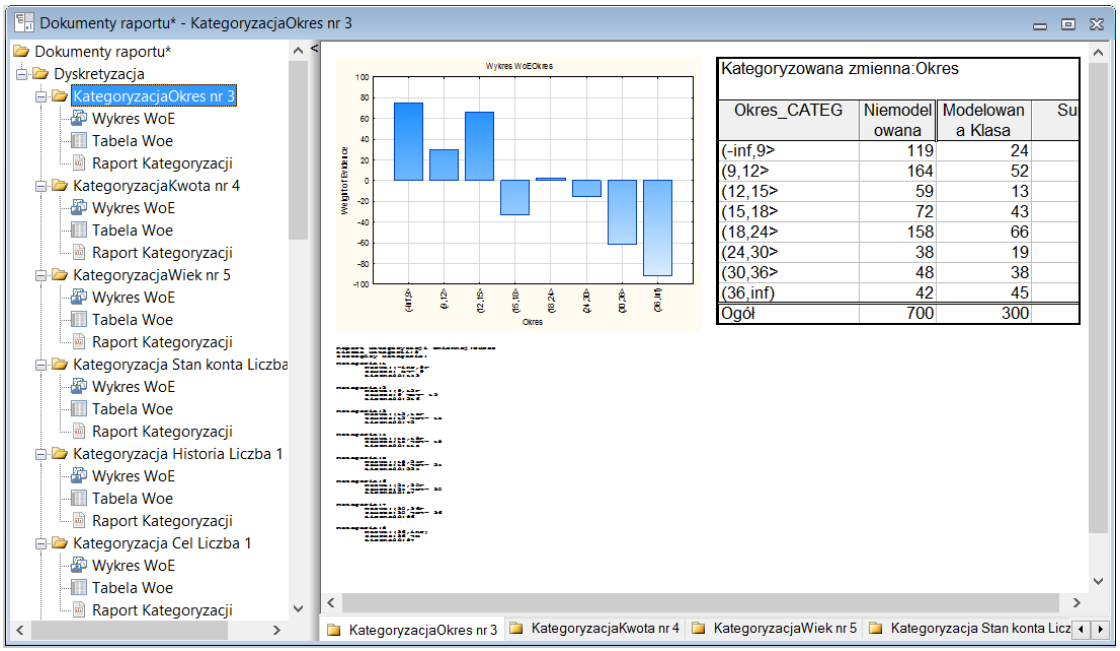
Klikamy na symbol w lewym górnym rogu węzła ***Dyskretyzacja*** aby wejść do jego ustawień. Wybieramy – analogicznie jak w innych oknach – zmienne: *Ocena* jako zależną, a pozostałe jako predyktory odpowiedniego typu.



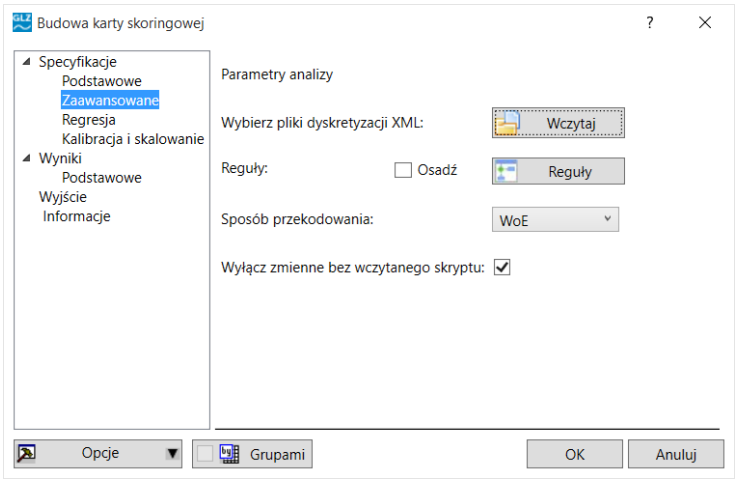
W panelu po lewej stronie znajdują się pozycje odpowiadające poszczególnym grupom ustawień. W wyniku działania węzła zostaną utworzone skrypty XML z regułami dyskretyzacji oraz skoroszyt z raportem dyskretyzacji. W zakładce ***Wyniki/Podstawowe*** wybierając przycisk ***Wybierz folder*** a następnie wskazujemy lokalizację do zapisu skryptów.



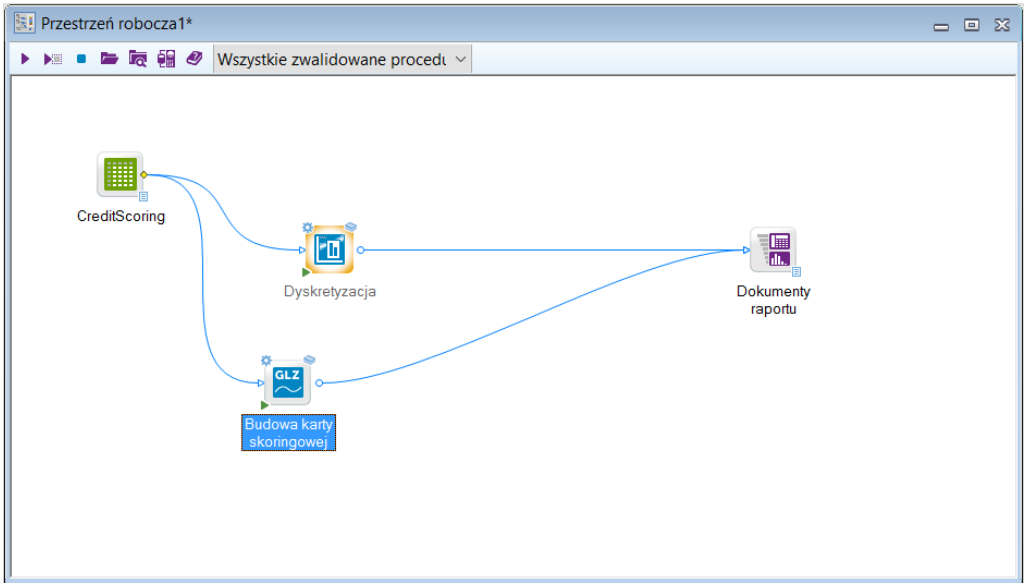
Po wprowadzeniu odpowiednich ustawień klikamy ***OK*** i uruchamiamy węzeł klikając symbol ▶ w lewym dolnym rogu. Po chwili pracy programu w przestrzeni pojawi się nowy element – skoroszyt ***Dokumenty raportu*** z wynikami dyskretyzacji, a na dysku we wskazanym miejscu zostaną zapisane pliki XML.



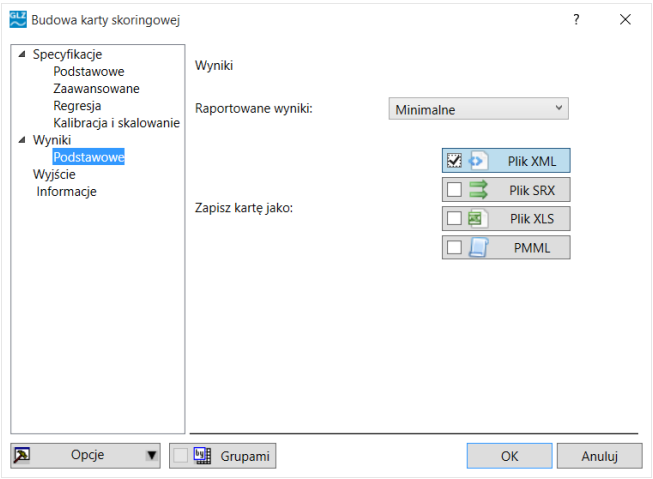
Przygotowane skrypty możemy od razu wykorzystać w przestrzeni. W tym celu otwieramy okno z ustawieniami węzła ***Budowa karty skoringowej.*** Na karcie ***Specyfikacje/Podstawowe*** wybieramy, tak jak poprzednio, zmienną zależną i predyktory, a następnie wchodzimy w ***Zaawansowane,*** naciskamy przycisk ***Wczytaj*** i wskazujemy wszystkie zapisane uprzednio skrypty dyskretyzacji.



W karce ***Regresja*** wybieramy opcję *Krokowa wsteczna* w polu ***Rodzaj regresji***. Klikamy ***OK*** i uruchamiamy węzeł. Po wykonaniu obliczeń wyniki trafiają również do skoroszytu ***Dokumenty raportu*** – oba węzły są z nim połączone tak jak przedstawiono poniżej:



Podstawowy zakres wyników węzła ***Budowa karty skoringowej*** obejmuje arkusz Statistica z utworzoną kartą skoringową oraz arkusze z oceną parametrów regresji logistycznej zastosowanej do budowy karty. Więcej wyników możemy wygenerować zmieniając ustawienia opcji *Raportowane wyniki* w zakładce ***Wyniki*/Podstawowe.**



Jeśli zaznaczymy tam opcję *Plik XML* i wskażemy miejsce na dysku, to oprócz wyników w skoroszycie zostanie wygenerowany skrypt XML pozwalający na szybkie wdrożenie modelu, które można zrealizować jak w przykładzie 11, ale także za pomocą kolejnego węzła Data Miner.