

STATISTICA Zestaw Skoringowy

Spis treści

1. Wstęp.....	2
2. Informacje techniczne	3
2.1. Wybór predyktorów.....	3
2.2. Reguły i interakcje.....	4
2.3. Dyskretyzacja zmiennych – konstrukcja atrybutów	4
2.4. Tworzenie tablicy skoringowej na podstawie przygotowanych danych.....	5
2.5. Tworzenie modelu skoringowego typu SURVIVAL.....	6
2.6. Analiza wniosków odrzuconych	6
2.7. Ocena modeli	7
2.8. Zarządzanie punktem odcięcia	8
2.9. Obliczanie skoringu.....	9
2.10. Testy kalibracji	9
2.11. Badanie stabilności populacji i cech	9
2.12. Analiza Vintage.....	10
2.13. Macierze migracji	10
3. Co nowego w wersji 4.0 programu STATISTICA Zestaw Skoringowy.....	11
3.1. Nowe moduły.....	11
3.2. Udoskonalenia	13

1. Wstęp

STATISTICA Zestaw Skoringowy pozwala łatwo zamienić dane w gotową tablicę skoringową. Do zbudowania tablicy skoringowej nie jest wymagana specjalistyczna wiedza - interfejs został zaprojektowany tak, aby poprowadzić użytkownika krok po kroku przez poszczególne etapy.

Rozwiązanie składa się z odpowiedniego zestawu STATISTICA (np. STATISTICA + Trees lub STATISTICA Data Miner) oraz dziesięciu dedykowanych modułów wspierających budowę, ocenę i pielęgnację tablicy:

- Przygotowanie danych
 - Wybór predyktorów
 - Reguły i interakcje
 - Dyskretyzacja zmiennych
- Modelowanie
 - Budowa tablicy skoringowej (model logistyczny)
 - Model skoringowy typu SURVIVAL
 - Analiza wniosków odrzuconych
- Ocena
 - Ocena modeli
 - Zarządzanie punktem odcięcia
 - Obliczanie skoringu
 - Testy kalibracji
- Monitoring
 - Badanie stabilności populacji i cech
 - Analiza Vintage
 - Macierze migracji

Za ich pomocą analityk w wygodny sposób może przeprowadzić wstępną ocenę danych, dokonać dyskretyzacji bądź kategoryzacji predyktorów z wykorzystaniem zaawansowanych metod analitycznych (np. algorytm CHAID) oraz dokonać oceny jakości podziału i mocy predykcyjnej przekształcanej zmiennej na podstawie wskaźnika WoE oraz IV.

W kolejnym etapie na podstawie wybranych przez użytkownika predyktorów i określonych parametrów skali budowany jest model skoringowy, który można następnie przekształcić do postaci tablicy skoringowej (*scorecard*). Budowa modelu skoringowego może przebiegać w sposób automatyczny, ale zaawansowani użytkownicy mają możliwość wyboru trybu eksperckiego, w którym mogą modyfikować wszystkie parametry modelu. Tablicę skoringową można zapisać w postaci odpowiedniego skryptu (Visual Basic, XML) oraz zachować w postaci dokumentu

tekstowego, bądź pliku Excela. System można rozbudować tak, aby generował tablicę skoringową w innych ustalonych formatach.

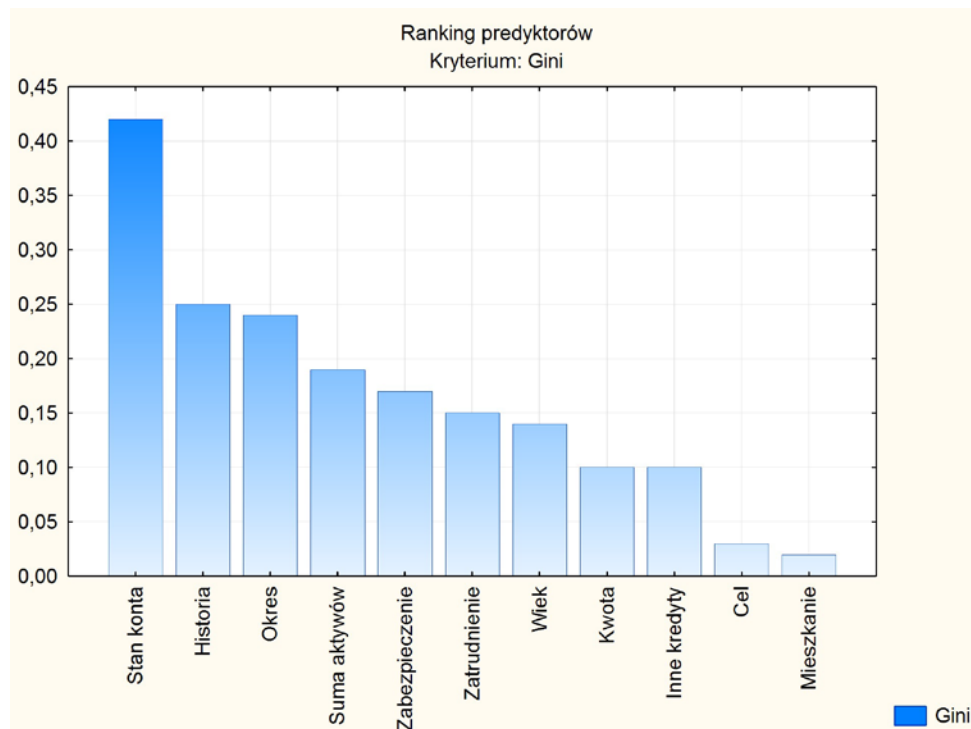
Funkcjonalność w zakresie budowy tablic skoringowych uzupełniają specjalnie przygotowane raporty przydatne zarówno podczas budowania, strojenia jak i utrzymania tablic skoringowych, np. do oceny zbudowanej tablicy, do wyboru optymalnego punktu odcięcia a także do badania stabilności populacji. Ponadto system zawiera moduł do analizy wniosków odrzuconych pozwalający na analizę wniosków odrzuconych przez bank i uwzględnienie ich przy budowie tablicy skoringowej poprzez uzupełnienie brakującej informacji o typie kredytu: „dobry/zły” z wykorzystaniem metod probabilistycznych.

2. Informacje techniczne

Funkcjonalność programu *STATISTICA Zestaw Skoringowy* w każdym z proponowanych wariantów obejmuje:

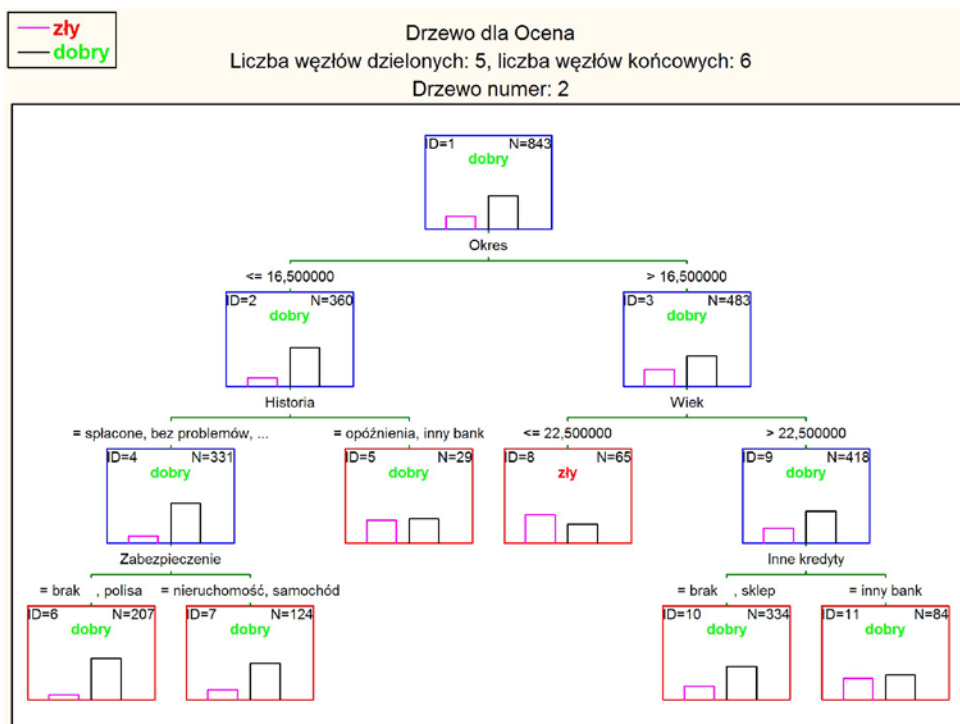
2.1. Wybór predyktorów

- przygotowanie rankingu predyktorów na podstawie miar Information Value, Gini oraz V Cramera
- wygodna eliminacja nieistotnych predyktorów
- wybór reprezentantów skupisk skorelowanych zmiennych ilościowych za pomocą analizy głównych składowych przydatne zwłaszcza w skoringach behawioralnych



2.2. Reguły i interakcje

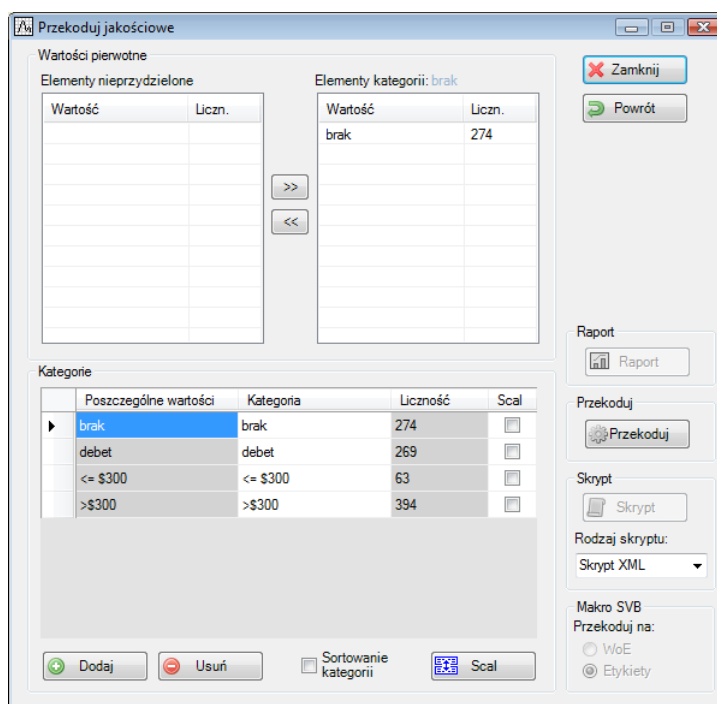
- o wyszukiwanie reguł umożliwiających identyfikację podgrup wysokiego ryzyka
- o wykorzystanie metody Losowy Las (*Random Forest*) do identyfikacji reguł



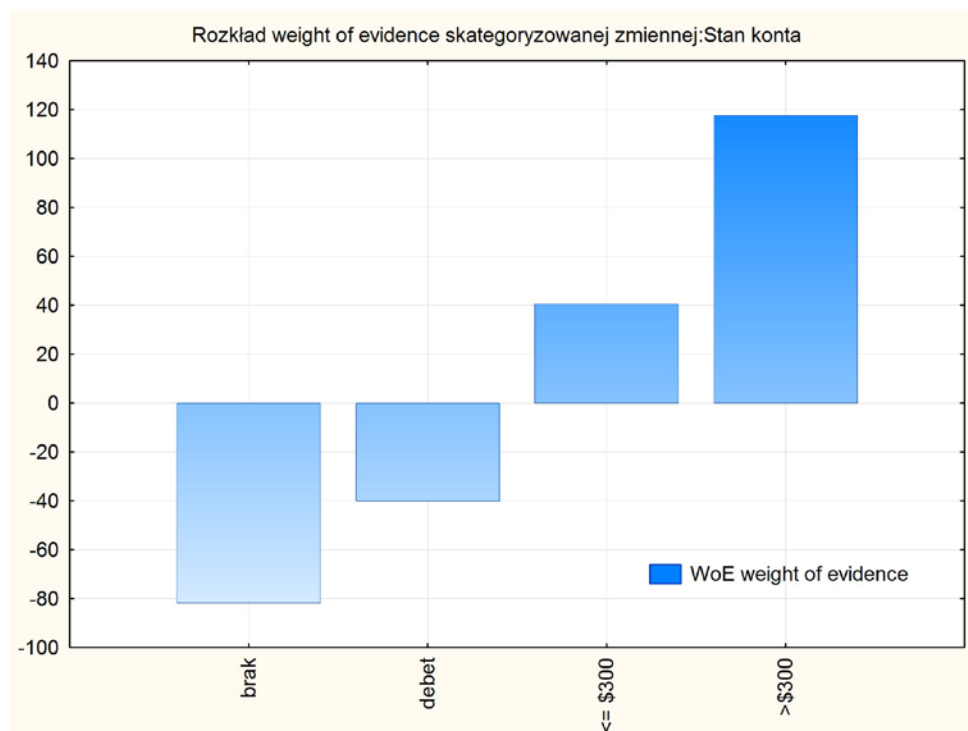
- o generowanie rankingu interakcji pomiędzy parami zmiennych przy użyciu regresji logistycznej

2.3. Dyskretyzacja zmiennych – konstrukcja atrybutów

- o manualne definiowanie przedziałów dla zmiennej ciągłej,
- o manualne grupowanie dla zmiennej dyskretnej



- o automatyczne tworzenie przedziałów dla zmiennej ciągłej według zadanych parametrów dotyczących liczebności przypadków w poszczególnych przedziałach
- o automatyczne tworzenie przedziałów dla zmiennej dyskretnej na podstawie minimalnej liczności
- o automatyczne tworzenie przedziałów dla zmiennej ciągłej lub dyskretnej za pomocą algorytmu CHAID
- o obsługa wartości nietypowych
- o diagnozowanie jakości podziału na przedziały na podstawie weight of evidence, wskaźnika information value oraz odpowiednich wykresów
- o możliwość wczytania skryptu dyskretyzacji i reedycja zdefiniowanych przedziałów



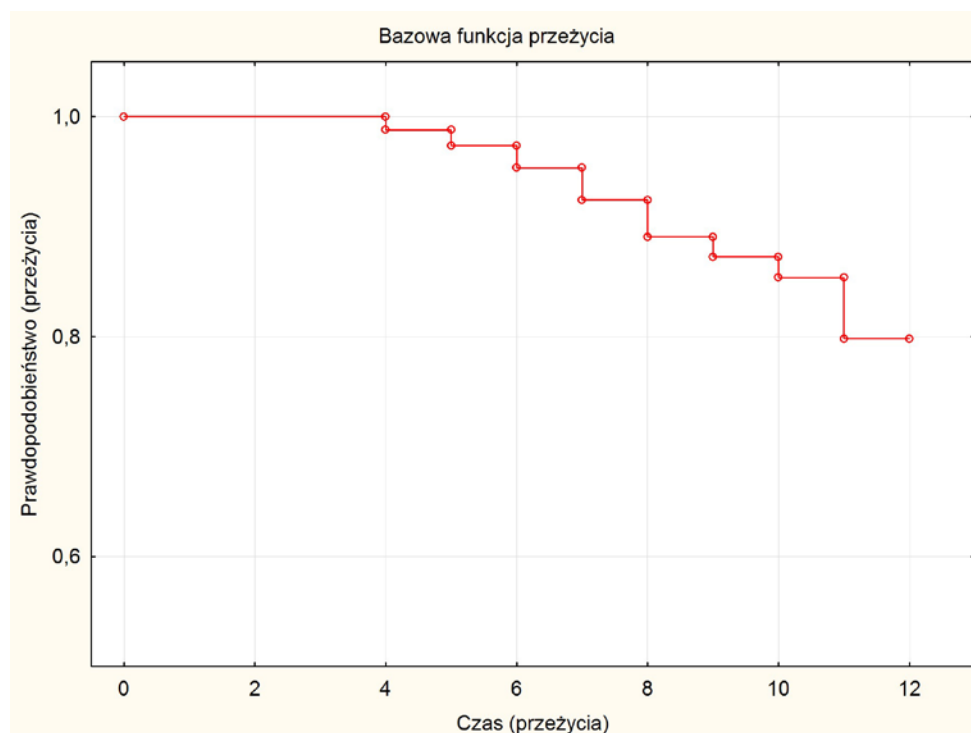
2.4. Tworzenie tablicy skoringowej na podstawie przygotowanych danych

- o tworzenie modelu skoringowego za pomocą regresji logistycznej - zaawansowane strategie doboru zmiennych do modelu
- o budowa modelu logistycznego na podstawie prób bootstrapowych
- o podział na próbę uczącą i testową
- o tworzenie wyskalowanej tablicy skoringowej (również typu *weight of evidence*) na podstawie modelu regresji logistycznej
- o zapis tablicy skoringowej w postaci kodu Visual Basic oraz XML
- o zapis tablicy skoringowej w postaci pliku Excela
- o możliwość zapisu tablicy skoringowej w postaci kodu w dowolnym języku (c, php, java itp.) na podstawie specyfikacji klienta
- o raport opisujący powstałą tablicę skoringową
- o tworzenie modelu skoringowego a pomocą drzew klasyfikacyjnych CART
- o tworzenie modelu skoringowego za pomocą wzmocnianych drzew klasyfikacyjnych (*boosted trees*)

Zmienna	Zakres	WoE	Ocena	s. Walda	p	Punkcja	Skoring zaokr.
Mieszkanie	własne	-43,600	0,00443	2,16262	0,14140	30,981	31
Mieszkanie	Wartość n...	-	-	-	-	36,750	37
Okres	(-inf;9>	75,377	0,00804	17,72377	0,00003	54,041	54
Okres	(9;15>	38,549	0,00804	17,72377	0,00003	45,497	45
Okres	(15;30>	-10,834	0,00804	17,72377	0,00003	34,041	34
Okres	(30;36>	-61,368	0,00804	17,72377	0,00003	22,318	22
Okres	(36;inf]	-91,629	0,00804	17,72377	0,00003	15,298	15
Okres	Wartość n...	-	-	-	-	37,189	37
Płeć	m	11,303	-0,01104	1,06505	0,30207	32,954	33
Płeć	k	-23,534	-0,01104	1,06505	0,30207	44,051	44
Płeć	Wartość n...	-	-	-	-	36,394	36
Rata	> 35	25,131	0,01806	11,92720	0,00055	49,650	50
Rata	25-35	15,547	0,01806	11,92720	0,00055	44,656	45
Rata	15- 25	6,454	0,01806	11,92720	0,00055	39,917	40
Rata	< 15	-15,730	0,01806	11,92720	0,00055	28,357	28
Rata	Wartość n...	-	-	-	-	36,468	36
Stan	MS	16,164	0,01692	4,08200	0,04334	44,446	44
Stan	MZW	16,164	0,01692	4,08200	0,04334	44,446	44
Stan	MR	-26,469	0,01692	4,08200	0,04334	23,632	24
Stan	K	-26,469	0,01692	4,08200	0,04334	23,632	24
Stan	Wartość n...	-	-	-	-	36,953	37
Stan konta	brak	-81,810	0,00767	53,33824	0,00000	18,449	18
Stan konta	debet	-40,139	0,00767	53,33824	0,00000	27,671	28

2.5. Tworzenie modelu skoringowego typu SURVIVAL

- o budowa modeli skoringowych za pomocą proporcjonalnego hazardu Coxa
- o symulacja przebiegu funkcji przeżycia dla różnych wartości parametrów wejściowych

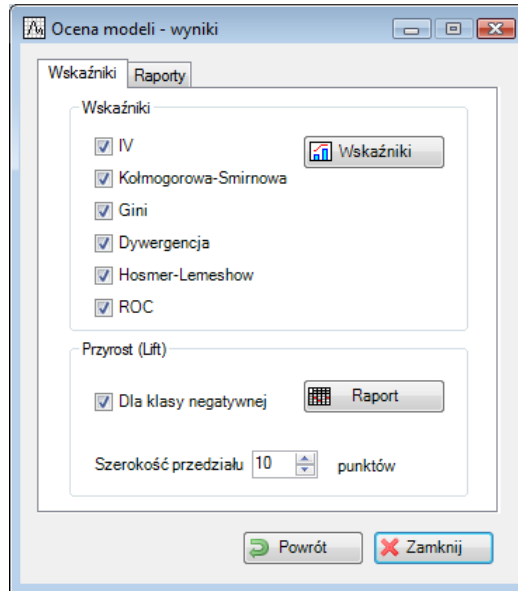


2.6. Analiza wniosków odrzuconych

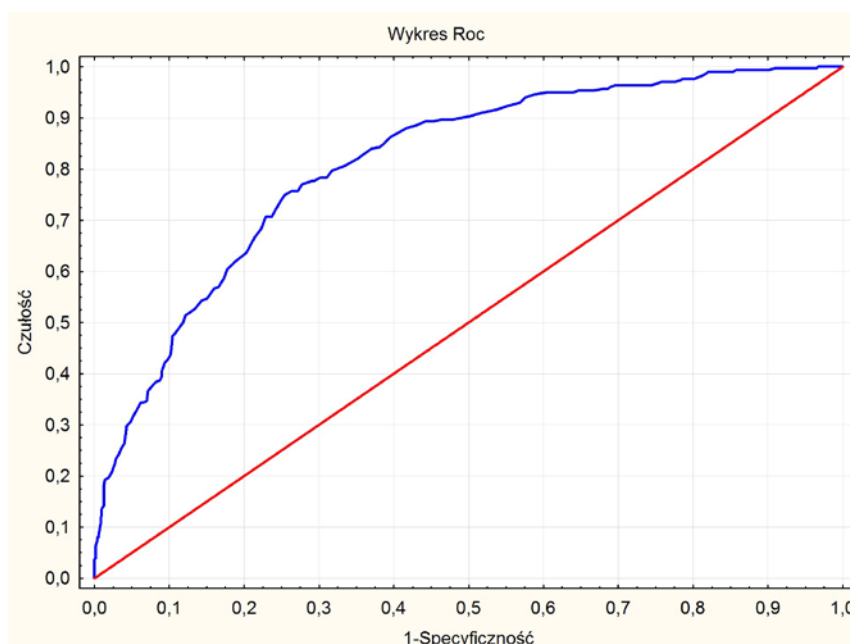
- o Parceling (połączenie metod statystycznych z wiedzą ekspercką)
- o Metoda k-najbliższych sąsiadów

2.7. Ocena modeli

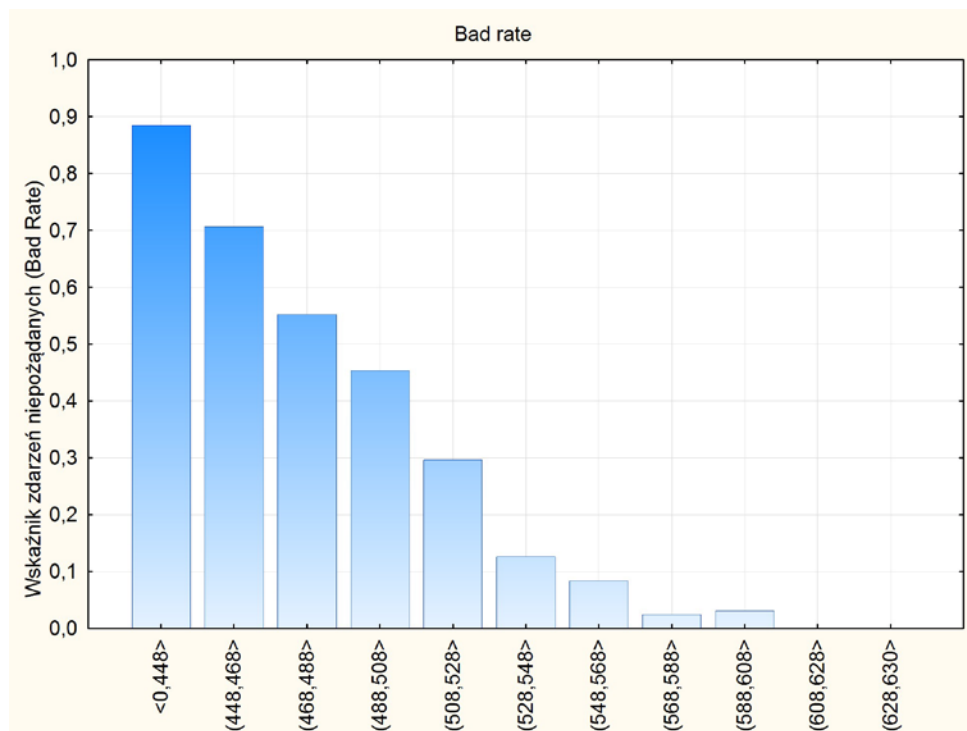
- Ocena tablic zapisanych w postaci XML (tablica skoringowa bądź model SURVIVAL)
- Ocena modeli na podstawie skoringu bądź prawdopodobieństwa zapisanego w arkuszu danych (dowolna metoda analityczna)



- Ocena jakości zbudowanych modeli na podstawie miar:
 - IV (*Information Value*)
 - KS (Kolmogorowa-Smirnowa) - dodatkowo wartość prawdopodobieństwa testowego p
 - Hosmera-Lemeshowa - dodatkowo wartość prawdopodobieństwa testowego p
 - Dywergencji
 - Giniego
 - Pola pod krzywą ROC

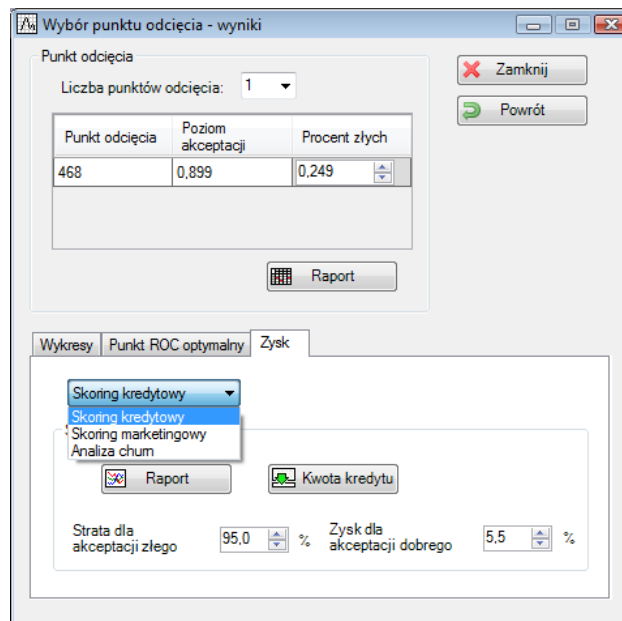


- Analiza lift
 - Wykres lift
 - Wykres gain
 - Raport wartości lift
- Raporty
 - Cech (*Characteristic report*)
 - Końcowej punktacji (*Final score report*)
 - Wykresy Bad rate oraz Odds



2.8. Zarządzanie punktem odcięcia

- możliwość wskazania od 1 do 3 punktów odcięcia
- zestaw narzędzi i raportów pozwalających ocenić trafność odcięcia
- wybór punktu odcięcia na podstawie analizy ROC dla zadanych kosztów błędnych klasyfikacji i wskazanej frakcji złych przypadków
- symulacja zyskowności modelu dla skoringu kredytowego, marketingowego oraz modeli lojalnościowych (*churn*)



2.9. Obliczanie skoringu

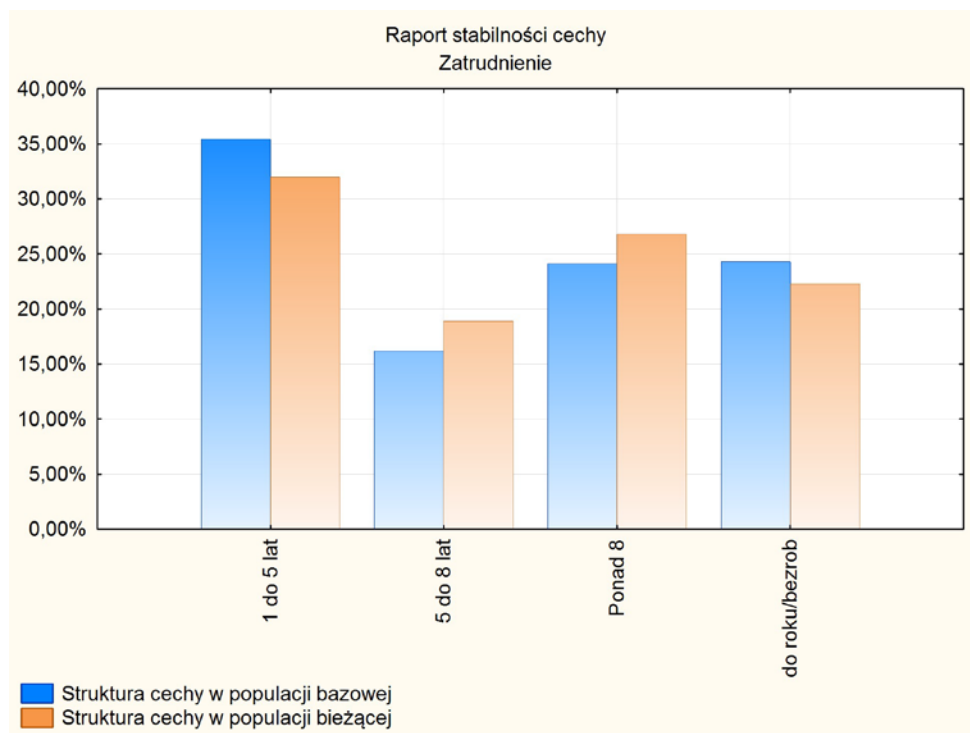
- o obliczanie skoringu dla nowych danych na podstawie wybranego modelu
- o możliwość wyliczania PD (default probability)
- o skalowanie wartości PD dla modeli budowanych na zbilansowanym zbiorze danych
- o wyliczanie skoringu dla modeli typu SURVIVAL

2.10. Testy kalibracji

- o Testowanie zgodności realizacji ryzyka w poszczególnych grupach ratingowych ze zdefiniowaną masterskalą
- o Test dwumianowy i normalny
- o Zaimplementowana strategia *traffic light approach*

2.11. Badanie stabilności populacji i cech

- o raport stabilności populacji
- o raporty stabilności cech
- o zestaw dodatkowych wykresów



2.12. Analiza Vintage

- monitorowanie stanu portfela kredytów w kolejnych miesiącach spłaty
- przygotowanie raportu w zależności od celu, statusu kredytów, liczby dni przeterminowania oraz wieku kredytobiorców
- zestaw wykresów pozwalających na łatwiejszy monitoring portfela kredytów i interpretację zachodzących zmian

2.13. Macierze migracji

- Obliczanie raportów struktury portfela oraz migracji kredytów
- Ujęcie ilościowe i kwotowe
- Zestaw wykresów opisujących zmiany przeterminowania

3. Co nowego w wersji 4.0 programu STATISTICA Zestaw Skoringowy

Przy projektowaniu nowej wersji programu STATISTICA Zestaw Skoringowy szczególny nacisk położyliśmy na uwzględnienie sugestii dotychczasowych użytkowników rozwiązania, tak aby jeszcze bardziej uprościć i przyspieszyć proces budowy i oceny kart skoringowych oraz dostarczyć dodatkowe narzędzia zgodnie z oczekiwaniami użytkowników systemu.

Wprowadzone rozwiązania powinny zadowolić zarówno użytkowników używających programu dla celów analizy ryzyka kredytowego i operacyjnego jak również osoby wykorzystujące nasz program w marketingu (*skoring marketingowy*), medycynie (*skoring medyczny*) i innych dziedzinach biznesu i nauki.

Wersja 4.0 Zestawu Skoringowego składa się obecnie z 13 modułów.

3.1. Nowe moduły

W nowej wersji dodano trzy moduły:

- 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 mamy możliwość przygotowania dwustanowych zmiennych pochodnych. Dodatkowo moduł pozwala na utworzenie rankingu interakcji pomiędzy parami zmiennych. Moduł dla każdej możliwej pary predyktorów buduje model logistyczny zawierający parę zmiennych oraz ich interakcję. Użytkownik ma możliwość oceny siły interakcji za pomocą testu LR.

Losowy las - kreator reguł

Reguła
(Okres > 15,5) AND (Mieszkanie <> wynajem AND Mieszkanie <> własne)

Drzewo	Zmienna 1	Zmienna 2	Liczność	Bad rate	Przyrost (złe)	Przyrost (dobre)	Wybierz
26	Stan konta	Historia	63	0,73	2,574	0,377	<input type="checkbox"/>
31	Kwota	Kwota	29	0,724	2,373	0,397	<input type="checkbox"/>
56	Stan konta	Historia	65	0,723	2,308	0,403	<input type="checkbox"/>
44	Stan konta	Historia	54	0,722	2,453	0,394	<input type="checkbox"/>
73	Kwota	Rata	57	0,719	2,66	0,385	<input type="checkbox"/>
5	Suma aktywów	Kwota	74	0,716	2,468	0,4	<input type="checkbox"/>
74	Zatrudnienie	Historia	42	0,714	2,292	0,415	<input type="checkbox"/>
17	Suma aktywów	Historia	55	0,709	2,214	0,428	<input type="checkbox"/>
94	Cel	Okres	95	0,705	2,395	0,418	<input type="checkbox"/>
34	Okres	Mieszkanie	77	0,701	2,292	0,43	<input type="checkbox"/>
95	Historia	Stan konta	63	0,698	2,32	0,432	<input type="checkbox"/>
3	Historia		75	0,693	2,45	0,428	<input type="checkbox"/>
53	Historia		26	0,692	2,347	0,436	<input type="checkbox"/>
62	Kwota	Okres	78	0,692	2,285	0,441	<input type="checkbox"/>
81	Suma aktywów	Kwota	63	0,683	2,314	0,45	<input type="checkbox"/>
24	Kwota	Cel	61	0,672	2,35	0,459	<input type="checkbox"/>

Zaznacz\odznacz
Przyrost (dobre) [dropdown]
Więcej od: 1,30 [input]
[OK] [Cancel]

Wyniki
Raport [button]
Reguły [button]

Skrypt
Skrypt [button]
Makro [dropdown]

Powrót [button]
Zamknij [button]

- Moduł *Testy kalibracji* umożliwia testowanie zgodności realizacji ryzyka w poszczególnych grupach ratingowych ze zdefiniowaną 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*.

Testy kalibracji

Dane wejściowe
Kredyty [button]
Symbol złego kredytu [dropdown]

Masterskala
Wprowadź [button]
 PD całości w ostatnim wierszu
 Sygnalizuj kolorami

Rodzaj testu
Oba [dropdown]
 Kryteria Nadzoru Austriackiego

Dołna granica 95,0 %
Górna granica 99,9 %

Poziomy ufnoci testów

Poziomy ufnoci %	Oblicz
90	<input checked="" type="checkbox"/>
95	<input checked="" type="checkbox"/>
98	<input checked="" type="checkbox"/>
99,9	<input checked="" type="checkbox"/>

[Dodaj] [Usuń] [Wszystkie]

[OK] [Anuluj]

- Moduł *Macierze migracji* 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.

3.2. Udoskonalenia

Dodatkowo wprowadzono szereg zmian i udoskonalień do pozostałych ośmiu modułów, z których najważniejsze to:

- Wybór predyktorów
 - Możliwość automatycznego wyboru czynników przy wyborze reprezentantów
 - Możliwość uwzględnienia zmiennych jakościowych przy wyborze reprezentantów
 - Miara GINI w rankingu predyktorów
- Budowa tablicy skoringowej
 - Dobór zmiennych do modelu wsparty za pomocą estymacji bootstrap
 - Możliwość wskazania zmiennej identyfikującej próbę uczącą i testową
- Ocena modeli
 - Możliwość oceny modeli małej próby za pomocą v-krotnego sprawdzianu krzyżowego oraz bootstrap
 - Wykres wartości KS
 - Możliwość utworzenia raportu walidacji modeli w przestrzeni roboczej programu *STATISTICA Data Miner*
 - Możliwość utworzenia raportu walidacji modeli w programie *STATISTICA Enterprise*
- Obliczanie Skoringu
 - Możliwość włączenia opcji kalibracji prawdopodobieństwa do oczekiwanej wartości PD
- Zarządzanie punktem odcięcia
 - Możliwość wyboru punktu odcięcia na podstawie zyskowności modelu dla skoringu kredytowego, skoringu marketingowego oraz skoringu *churn*