PRAKTYCZNY SKORING - NIE TYLKO KREDYTOWY

Piotr Wójtowicz, Grzegorz Migut
StatSoft Polska

Jakie są różnice pomiędzy osobami prawidłowo regulującymi swoje zobowiązania a niechętnie spłacającymi swoje długi, czy też pomiędzy klientami lojalnymi a skłonnymi odejść do konkurencji? Znajomość odpowiedzi na te i podobne pytania pozwala kształtować skuteczniejszą strategię działania instytucji finansowych oraz lepiej wykorzystać dostępne zasoby i potencjał rynku. Odpowiedzi na te pytania najlepiej jest szukać wewnątrz własnej organizacji, ukryte są one w olbrzymich ilościach informacji zgromadzonych w systemach informatycznych. Najlepszym sposobem ich wydobycia jest wykorzystanie metod statystycznych oraz data mining.

Jednym z najpopularniejszych rodzajów modeli statystycznych wykorzystywanych w bankowości są modele skoringowe. Parametry modeli skoringowych po oszacowaniu na historycznym zbiorze obserwacji możemy wykorzystywać do klasyfikowania nowych i obecnych klientów. Modele te wykorzystywane są jako narzędzia wspierające proces kredytowy, oceniają wiarygodność kredytową klientów, bądź skłonność do nadużyć. Warto zauważyć, że modele skoringowe wykorzystywane są nie tylko w obszarze zarządzania ryzykiem, ale np. również w procesie utrzymania klienta, gdzie wskazują osoby najbardziej zagrożone odejściem. Kolejny obszar to wsparcie procesu sprzedaży - za ich pomocą można optymalizować ofertę sprzedaży dla konkretnej grupy klientów. Model skoringowy potrafi wskazać osoby, które z największym prawdopodobieństwem odpowiedzą na ofertę poszczególnych produktów. Krótko mówiąc, skoring sprawdza się, kiedy chcemy podzielić naszych klientów na dwie kategorie: spłaci kredyt/nie spłaci, odpowie na ofertę/nie odpowie, przyniesie zysk/nie będzie dochodowy, zagrożony odejściem/pozostanie klientem.

Ogólnie mówiąc, na podstawie cech klienta, np. demograficznych, behawioralnych itp., budujemy model skoringowy, który przewiduje prawdopodobieństwo przynależności do pożądanej przez nas kategorii.

Zazwyczaj, zwłaszcza w skoringu kredytowym, model taki należy zamienić na kartę skoringową, w której dla każdej z cech analizowana osoba otrzymuje punkty, które są sumowane w jedną syntetyczną ocenę. Do niedawna najczęściej stosowana metoda typowo ekspercka oznaczała z reguły zakup kart od zewnętrznego dostawcy. Od pewnego czasu, ze względu na coraz lepszą dostępność dobrej jakości danych w poszczególnych instytucjach finansowych, jakość uzyskiwanych przewidywań oraz koszty, stosowana jest przede wszystkim metoda statystyczno-ekspercka – punktacja uzyskiwana jest w wyniku działania

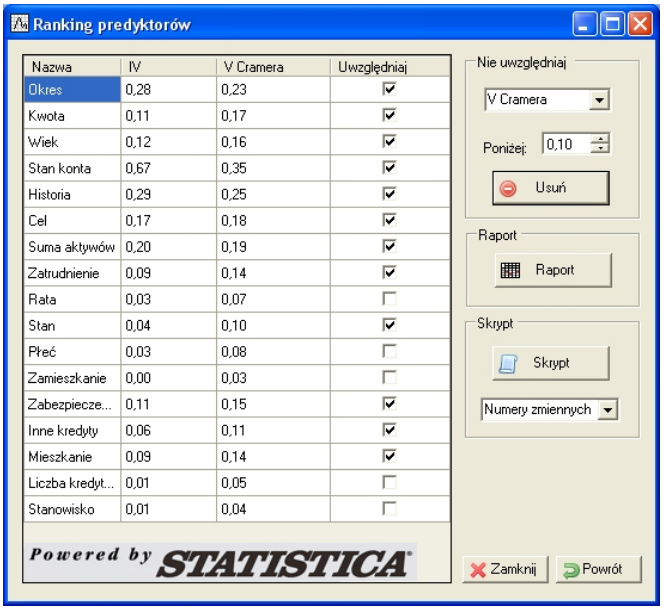
metod statystycznych pod nadzorem specjalistów branżowych z wykorzystaniem własnych danych.

Modele skoringowe można budować przy użyciu szeregu mniej lub bardziej zaawansowanych metod statystycznych i data mining. Do najpopularniejszych należy zaliczyć regresję logistyczną, drzewa klasyfikacyjne oraz różne „odmiany" drzew – drzewa wzmacniane (*boosting trees*) czy losowy las (*random forest*). Wykorzystywać można również sieci neuronowe, metodę wektorów nośnych (*suport vector machines*) czy k-najbliższych sąsiadów i szereg innych. Dla wielu osób już same nazwy tych procedur analitycznych brzmią dziwnie i egzotycznie, jak zatem zastosować je w praktyce? Na szczęście są odpowiednie narzędzia, które pozwolą nawet mniej doświadczonym analitykom szybko i efektywnie budować, oceniać i wdrażać modele i karty skoringowe. Prześledźmy najważniejsze etapy budowy karty skoringowej z wykorzystaniem popularnego w polskich instytucjach finansowych *Zestawu Skoringowego STATISTICA*.

Oczywiście w pierwszej kolejności musimy się zatroszczyć o dane historyczne potrzebne do budowy modelu. Z reguły dane o wcześniejszych kredytach, czy o zachowaniach klientów są dostępne w różnych bazach danych, bądź przechowywane w postaci plików *Excela*. Bez problemu można je więc wczytać do naszej aplikacji. Należy tylko zadbać, aby docelowy zbiór danych miał postać tabeli, w której w wierszach znajdą się poszczególni klienci (w analizie danych często wiersze nazywa się przypadkami), a w kolumnach poszczególne cechy naszych klientów (nazywane zmiennymi), np. wiek, przeznaczenie kredytu, produkty z jakich korzysta, liczba kontaktów z *call center* i inne w zależności od celu analizy.

To jeszcze nie koniec przygotowania danych, kolejnym krokiem jest odpowiednie przekodowanie danych (oczywiście mamy pod ręką wygodne narzędzia do tego), aby modelowana zmienna miała dwie wartości (kupił/nie kupił, spłacił/nie spłacił itp.). Oczywiście należy podjąć odpowiednie decyzje biznesowe co oznaczają odpowiednie kategorie, np. po jakim czasie należy uznać, że klient nie kupił produktu, albo nie spłacił kredytu.

Po przekodowaniu danych i ich wyczyszczeniu (w *STATISTICA* dostępnych jest szereg narzędzi do tego celu) możemy przeprowadzić wstępną selekcję cech klienta, jakie uwzględnimy na karcie skoringowej za pomocą modułu *Wybór predyktorów*. Budujemy w nim ranking zmiennych i wybieramy jedynie te zmienne, które są w sposób istotny powiązane z modelowanym zjawiskiem. Jeśli budowany model wykorzystuje cechy opisujące zachowanie klientów (skoring behawioralny), bardzo często mamy do czynienia z grupami podobnych do siebie zmiennych. Ponieważ nadmiarowość informacji jest często równie niebezpieczna, jak jej brak, to w takiej sytuacji wybieramy reprezentantów grup skorelowanych zmiennych ilościowych za pomocą analizy głównych składowych. Na szczęście nie musimy wiedzieć, co to są te składowe główne – *STATISTICA* wykona analizę automatycznie.



Przykładowy ranking predyktorów

Kolejnym etapem przygotowania danych, szczególnie jeśli w wyniku analizy chcemy otrzymać klasyczną kartę skoringową, jest dyskretyzacja zmiennych, która pozwala podzielić wybrane zmienne na przedziały. Przedziały te będą odpowiadały później poszczególnym kategoriom na karcie skoringowej. Wstępny podział zostanie zrobiony automatycznie (np. algorytmem CHAID), ale w tym momencie wymagana jest także praca analityka. W *STATISTICA* podczas podziału zmiennych na przedziały można jednocześnie uwzględnić wiedzę ekspercką oraz statystyczną, a także uwarunkowania biznesowe, a zatem poprawiamy „automat", dopasowując granice przedziałów karty skoringowej do bieżących potrzeb. Moduł dyskretyzacji zmiennych pozwala również na obsługę braków danych, które traktujemy, jako osobną kategorię uwzględniając możliwość istotnego wpływu braku danych na badane zjawisko. Pamiętajmy więc – brak danych to też informacja.

W tym momencie mamy za sobą najbardziej pracochłonny etap, czyli przygotowanie danych, i możemy przejść do kolejnego etapu – budowy modeli analitycznych.

Najpopularniejszą metodą budowy modeli jest zdecydowanie regresja logistyczna. Popularność ta wynika zarówno z przyczyn historycznych, jak i praktycznych - postać modelu jest czytelna i łatwa do interpretacji. Zwykle model regresji po oszacowaniu parametrów przekształca się do tablicy skoringowej, w której poszczególnym poziomom (zdefiniowanym na etapie dyskretyzacji przedziałom czy atrybutom) każdej z analizowanych cech klienta przypisywana jest odpowiednia liczba punktów, określająca wpływ danej kategorii na szanse wystąpienia danego zjawiska.

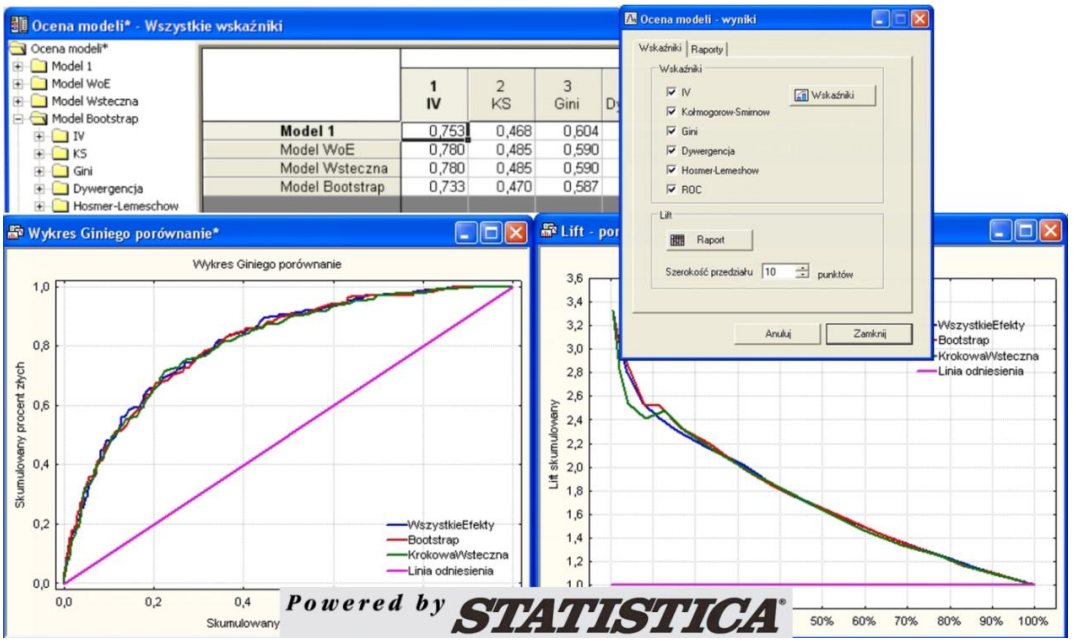
Regresja logistyczna – zaczyna to brzmieć trochę groźnie. Ale nie ma obawy. *Zestaw Skoringowy STATISTICA* ma ustawione wszystkie parametry na odpowiednie wartości domyślne i każdy jest w stanie wykonać regresję logistyczną. Oczywiście zaawansowany analityk może sam modyfikować poszczególne opcje budowy modelu, zatem zachęcamy do pogłębiania swojej wiedzy w tym obszarze i podjęcia własnych prób.

Zbudowany model można następnie przekształcić do karty (tablicy) skoringowej, którą można zachować, np. w postaci dokumentu tekstowego lub arkusza *Excel*. Tak przygotowaną kartę skoringową można uruchomić wewnątrz systemu albo zintegrować lub przenieść do systemu obsługi wniosków. Przykładowo karta skoringowa może wyglądać w następujący sposób (fragment):

Fragment przykładowej karty skoringowej

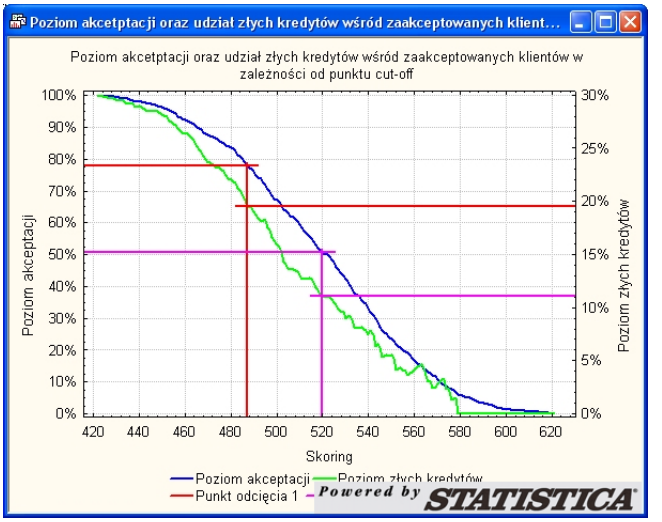
Wiek	Punkty
Do 35	20
Od 35 do 60	25
Pow. 60	15
Okres bycia klientem	Punkty
Do 1 roku	10
Od 1 roku do 3 lat	12
Powyżej 3 lat	15
Liczba produktów	Punkty
1	12
2-4	15
Powyżej 4	10

Doświadczeni analitycy mogą ocenić jakość modelu regresji logistycznej na podstawie różnych dostępnych wskaźników statystycznych. Jednak dla większości osób wygodniejsze i bardziej czytelne będą miary dostępne w module *Ocena modeli*, który umożliwia ocenę i porównanie zbudowanych modeli skoringowych. W skoringu wykorzystuje się 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* oraz *Lift*. Nie będziemy tu opisywali szczegółowo poszczególnych wskaźników - dla każdego tworzony jest szczegółowy raport z prezentacją graficzną, więc łatwo można ocenić uzyskane modele/karty skoringowe.



Przykładowa ocena modelu

To jeszcze nie koniec pracy. Mając zbudowany model skoringowy, możemy przystąpić do oceny poszczególnych klientów pod kątem prawdopodobieństwa przynależności do pożądanej przez nas grupy. Dla każdego klienta sumujemy punkty uzyskane za przynależność do poszczególnych kategorii, ale gdzie się kończą „dobrzy" klienci, a zaczynają „źli"? Musimy wskazać graniczną wartość, tzw. punkt odcięcia. I znowu doświadczenie i wiedzę analityka można wesprzeć metodami statystycznymi, a to wszystko oczywiście z uwzględnieniem strategii banku. Można zastosować wybór punktu odcięcia na podstawie analizy ROC dla zadanych kosztów błędnych klasyfikacji lub wskazanej frakcji złych kredytów.



Przykładowy raport zależności od punktów odcięcia

Po ustaleniu odpowiedniego punktu odcięcia w końcu możemy zacząć stosować uzyskany model. Należy jednak mieć świadomość, że nawet najlepszy model zdezaktualizuje się z czasem. Dlatego należy regularnie korzystać z raportów oceniających stabilność populacji i cech. Z czasem, gdy klienci się zmienią, należy przebudować model – raport stabilności pozwoli uchwycić odpowiedni moment.

Na potrzeby skoringu kredytowego opisaną powyżej ścieżkę budowy modeli można jeszcze uzupełnić o dostępną w *STATISTICA* analizę wniosków odrzuconych, pozwalającą na analizę wniosków odrzuconych przez bank i uwzględnienie ich przy budowie tablicy skoringowej. W tym celu należy uzupełnić brakującą informację o typie kredytu: „dobry/zły" z wykorzystaniem dostępnych metod probabilistycznych.

Podsumowując, warto zwrócić uwagę, że taki proces budowy modeli skoringowych można przeprowadzić w dowolnej instytucji finansowej, zarówno w dużym banku (np. PKO BP), jak i w znacznie mniejszych instytucjach (np. SKOK) – przykłady na www.StatSoft.pl. Co więcej, jest to metoda uniwersalna, łatwa do wdrożenia i relatywnie niedroga.

Do najważniejszych zalet karty skoringowej należą:

- czytelna i zrozumiała dla każdego postać modelu,
- sprawdzony i uznany standard,
- obiektywna, porównywalna ocena,
- szybkość oceny.

Zachęcamy do korzystania z tego efektywnego narzędzia, jakim jest karta skoringowa, nie tylko do zarządzania ryzykiem, ale także w obszarze CRM analitycznego i wszędzie tam, gdzie należy rozdzielić klientów na dwie grupy w zależności od ich oczekiwanego zachowania.