



WYKORZYSTANIE SKORINGU DO PRZEWIDYWANIA WYŁUDZEŃ KREDYTÓW W INVEST-BANKU

Bartosz Wójcicki

Naczelnik Wydziału Analiz i Prewencji Przestępstw, Invest-Bank S.A.

Grzegorz Migut

StatSoft Polska Sp. z o.o.

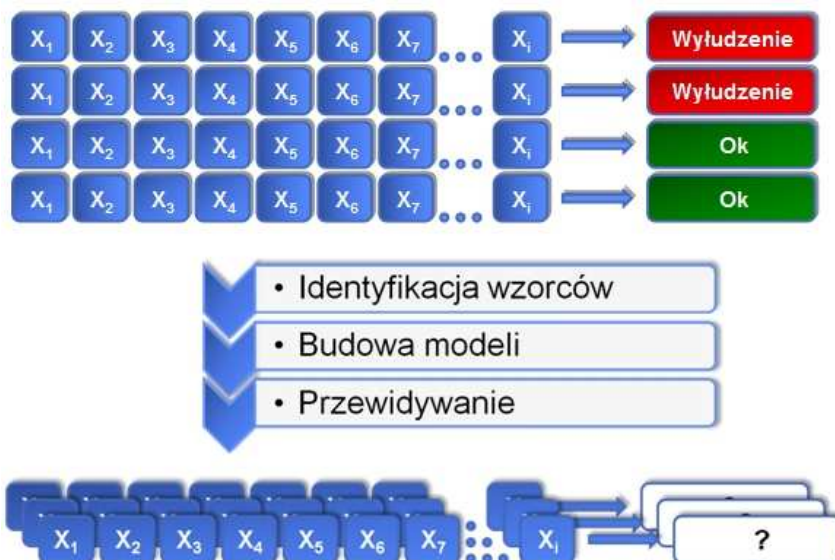
Problem nadużyć, zwłaszcza na polu działalności kredytowej, jest dla każdej instytucji finansowej poważnym wyzwaniem wpływającym na rentowność jej działalności. Dlatego też Invest-Bank S.A. podjął decyzję o wdrożeniu dedykowanego systemu zarządzania procesem weryfikacji wniosków kredytowych, którego jednym z kluczowych zadań jest ocena ryzyka popełnienia nadużycia przez potencjalnych kredytobiorców. Ważnym elementem tego systemu są modele skoringowe i zestaw reguł oceniających skłonność klientów do wyłudzenia kredytu.

Dyktowany przez rynek wymóg prostoty i elastyczności procedur bankowych, a z drugiej strony rosnące globalne ryzyko występowania problemów w spłacalności i wyłudzeń kredytów spowodowało, że oparcie procesu przyznawania kredytów jedynie na modelach generycznych oraz wiedzy weryfikatorów stało się niewystarczające do zapewnienia akceptowalnego dla Banku poziomu ryzyka.

Ponieważ Bank dysponował bogatą bazą historycznych kredytów, możliwe stało się przygotowanie na ich podstawie modeli skoringowych. Dzięki temu odnalezione reguły i wzorce mogły być idealnie dopasowane do rzeczywistego profilu klientów Banku.

Modele skoringowe to modele statystyczne lub data mining budowane na podstawie historycznych danych dotyczących cech kredytobiorców oraz informacji, czy dany kredyt był kredytem wyłudzonego czy też nie. Na podstawie takich danych model określa wzorce zachowań kredytobiorców. Jeśli wzorce wychwycone przez model okażą się wartościowe, możemy je następnie zastosować dla nowych klientów. Przyjmujemy tutaj niejawnie założenie, że podobne wzorce będą pojawiały się w przyszłości.

Modele te określamy mianem modeli skoringowych, ponieważ rezultatem ich działania jest ocena (*scoring*) ryzyka wyłudzenia przez klienta kredytu. Ocena ta może zostać wyrażona w formie prawdopodobieństwa bądź punktacji – im niższa ocena, tym większe ryzyko popełnienia nadużycia. Ogólny schemat budowy tego typu modeli przedstawia poniższy rysunek.



W niniejszym artykule zaprezentowane zostaną najważniejsze etapy budowy modeli skoringowych w Invest-Banku oraz korzyści, jakie przyniosło ich wdrożenie.

Założenia projektowe

INVEST-BANK SA. jest bankiem detalicznym realizującym działalność bankową na terenie całego kraju. Przy sprzedaży produktów kredytowych wykorzystywana jest zarówno sieć Oddziałów Banku, jak również sieć Partnerów Handlowych. Bank podjął działania mające na celu zmniejszenia ryzyka udzielenia kredytów wyłudzonych. W ramach prowadzonych analiz nie stwierdzono istotnych (prostych) powiązań danych zawartych w systemach informatycznych, mogących wykryć próbę wyłudzenia kredytu przed jego uruchomieniem. Ponieważ Bank nie dysponował wiedzą i narzędziami informatycznymi pozwalającymi przeprowadzić dogłębną analizę i zbudować model określający prawdopodobieństwo wyłudzenia produktu kredytowego przed jego uruchomieniem, zdecydowano się na budowę systemu informatycznego wspierającego pracę pracowników Banku weryfikujących dane zawarte na wnioskach kredytowych. System miał za zadanie poprawę jakości portfela kredytowego, poprzez znaczną eliminację kredytów wyłudzonych, bazując na danych historycznych zgromadzonych w bankowych bazach danych.

W ramach przetargu wyłoniono firmę dostarczającą taki system, a zbudowanie modeli statystycznych powierzono firmie StatSoft Polska.

Przebieg procesu modelowania

Podczas realizacji projektu wykorzystane zostały sprawdzone metodyki data mining pozwalające na standaryzację procesu modelowania, odpowiednią koordynację prac analitycznych oraz klarowne dokumentowanie uzyskanych wyników analiz. Podczas projektu



korzystano z metodyki *Virtuous Cycle of Data Mining*. Korzystanie z tego standardowego podejścia pozwoliło na sprawowanie kontroli nad poszczególnymi etapami procesu analizy, sprawne wyłapywanie i korektę ewentualnych błędów popełnionych zarówno podczas gromadzenia danych, jak i realizacji samego projektu oraz uzyskanie spójnych, powtarzalnych wyników.

Proces wydobywania wiedzy z danych jest procesem złożonym, składa się na niego kilka krytycznych etapów, bez realizacji których uzyskanie pozytywnych rezultatów jest praktycznie niemożliwe. Zgodnie z metodyką *Virtuous Cycle of Data Mining* cały proces analizy podzielono na cztery odrębne, następujące po sobie etapy:

- ◆ analizę biznesową,
- ◆ zrozumienie i przygotowanie danych,
- ◆ modelowanie,
- ◆ ocenę uzyskanych wyników,

Specyficzne działania podejmowane na kolejnych etapach procesu analizy oraz pojawiające się podczas ich realizacji problemy były na bieżąco konsultowane z ekspertami Banku, których wiedza pozwoliła uniknąć szeregu niebezpieczeństw pojawiających się na kolejnych etapach projektu. Umożliwiła także w odpowiedni sposób ukierunkować działania analityków biorących udział w projekcie oraz ustalić krytyczne dla całego projektu parametry.

Analiza biznesowa

Przed przystąpieniem do budowy modeli duży nacisk położono na zapoznanie się z problemem biznesowym zdefiniowanym przez Bank. Szczegółowo zweryfikowane zostały wymagania Banku odnośnie planowanych analiz, jego potrzeb oraz oczekiwań stawianych w kontekście budowy i działania modeli. Zapoznano się także szczegółowo z procesem weryfikacji wniosków kredytowych.

Celem tego etapu było określenie krytycznych parametrów projektu, takich jak:

- ◆ produkty kredytowe, dla których mają zostać zbudowane modele,
- ◆ zakres danych, jakie będą używane podczas modelowania,
- ◆ poziom szczegółowości danych,
- ◆ definicja wyłudzenia.

Po konsultacjach ustalono, że na pierwszym etapie modelowania wykorzystane zostaną jedynie dane z wniosków kredytowych, które podczas kolejnych aktualizacji modeli zostaną uzupełnione o dane pochodzące z BIK oraz dane geograficzne.

Modele miały zostać zbudowane dla trzech produktów kredytowych:

- ◆ kredytu gotówkowego,
- ◆ kredytu ratalnego,
- ◆ kredytu samochodowego.



Najtrudniejszą na tym etapie kwestią było określenie, czym jest nadużycie na poziomie operacyjnym, oraz ustalenie okoliczności, jakie muszą być spełnione, aby można lub nie można było powiedzieć, że dany kredyt został wyludzony. Definicja, którą należało przyjąć musiała być jednoznaczna i niezmienna na pozostałych etapach analizy, ponieważ na jej podstawie planowano wyszukiwanie wzorców oraz przeprowadzenie oceny uzyskanych wyników. Określając kryteria, na podstawie których została przygotowana definicja „złego” kredytu, pod uwagę wzięto jej zgodność z przyjętym celem biznesowym oraz uwarunkowaniami technicznymi.

Po konsultacjach przyjęto definicję kredytu opartą na obserwacji spłacalności kolejnych rat kredytu połączoną z informacjami o wystąpieniu nadużycia napływającymi z poszczególnych Oddziałów Banku. Szczegóły definicji ze względu na poufność nie zostaną przedstawione.

Cały ten etap był najbardziej krytycznym elementem projektu, ponieważ decyzje dotyczące kluczowych parametrów projektu, jakie zostały podjęte na tym etapie, miały zasadniczy wpływ na przebieg realizacji kolejnych etapów analizy.

Zrozumienie i przygotowanie danych

Po określeniu krytycznych parametrów projektu rozpoczęto etap, którego celem było jak najlepsze zrozumienie charakteru danych i występujących w nich problemów oraz takie ich przygotowanie, aby zawarte w nich wzorce mogły być w jak najprostszy sposób zidentyfikowane przez algorytmy, jakie miały być użyte do budowy modeli skoringowych.

Przed przystąpieniem do analizy Bank przekazał konsultantom StatSoft przygotowane ekstrakty danych.

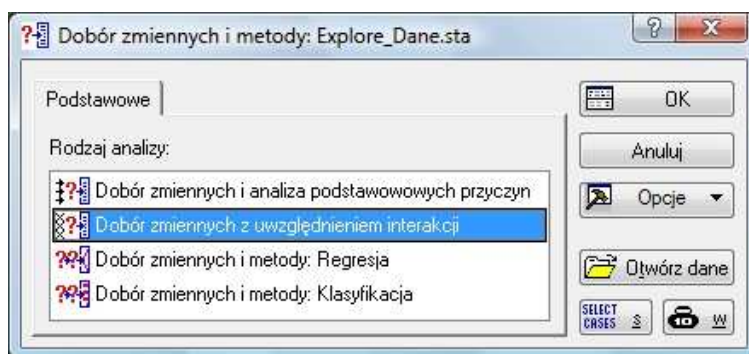
Wstępna analiza danych miała na celu odrzucenie z grupy potencjalnych predyktorów grupy zmiennych bezużytecznych dla modelowania. Zmienne wykluczano z grupy potencjalnych predyktorów, jeśli:

- ◆ nie miały wartości lub były wypełnione w znikomym stopniu,
- ◆ wszystkie wartości były takie same (brak zmienności),
- ◆ były zmiennymi anachronicznymi – informacje w tych zmiennych były zapisane po przyznaniu kredytu – ich wartości nie były znane w trakcie weryfikacji,
- ◆ były identyfikatorami,
- ◆ występował brak związku pomiędzy badanym zjawiskiem a analizowaną zmienną – wartość IV (*Information Value*) = 0,
- ◆ wystąpił znikomy związek pomiędzy badanym zjawiskiem a analizowaną zmienną, połączony ze złym uwarunkowaniem zmiennej, np. zmienna rzadka (*sparse data*) – zawierająca znikomą liczbę wypełnionych przypadków,
- ◆ zmienna była zmienną tekstową wypełniana dodatkowymi informacjami w sposób ręczny.

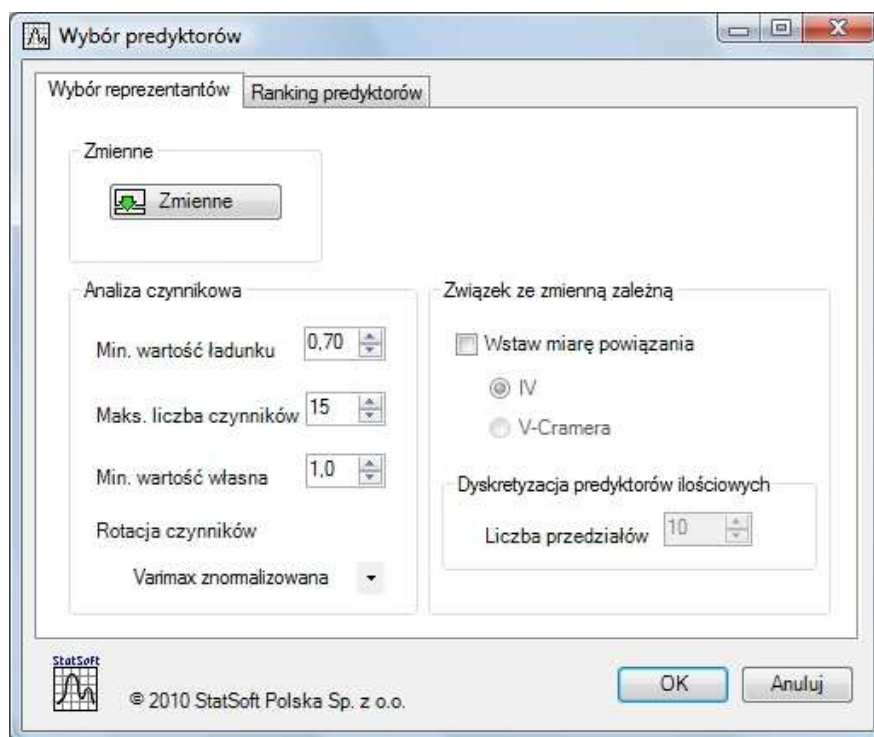


Jeśli dana zmienna wykazywała nawet słaby związek z badanym zjawiskiem lub jej znaczenie merytoryczne sugerowało, iż może on się pojawić w przyszłości, zmienna pozostawała w grupie potencjalnych predyktorów.

Kolejny krok analizy polegał na uzupełnieniu zestawu zmiennych pierwotnych dodatkowymi zmiennymi pochodnymi, które mogłyby wprowadzić do modelu pewną dodatkową informację. Zestaw zmiennych pochodnych określony został z jednej strony na podstawie wiedzy merytorycznej ekspertów Banku oraz analityków, z drugiej strony dodatkowe zmienne pochodne określono za pomocą dedykowanego modułu *Dobór zmiennych i metody*, pozwalającego wyszukiwać istotne interakcje zawarte w zbiorze danych i na tej podstawie definiować zmienne pochodne. Dzięki temu modułowi udało się zidentyfikować istotne zmienne pochodne, mające dodatkowo intuicyjną interpretację merytoryczną. Zmienne te zostały dołączone do pierwotnego zestawu danych.



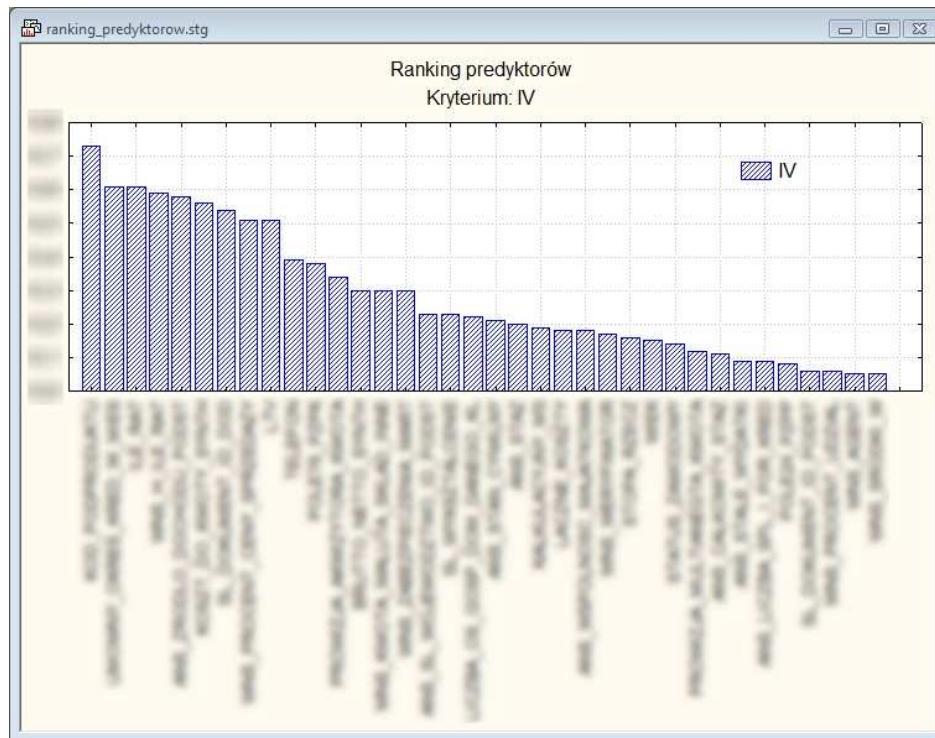
Celem kolejnego etapu była redukcja pierwotnego zestawu potencjalnych predyktorów i wyeliminowanie z niego tych cech, które były ze sobą nadmiernie skorelowane, powielając informację zawartą w innych zmiennych.



Za pomocą opcji *Wybór reprezentantów* modułu *Wybór predyktorów* zawartego w programie *STATISTICA Zestaw Skoringowy* (zob. rys. powyżej) przeprowadzono grupowanie analizowanych zmiennych w zestawy podobnych (w sensie korelacji) cech. Moduł ten oparty jest na analizie czynnikowej z rotacją czynników, która umożliwiła identyfikację grup zwykle mocno skorelowanych ze sobą predyktorów. Ponieważ nadmierna korelacja zmiennych mogła mieć negatywny wpływ na proces szacowania parametrów modelu, dodatkowo skorelowane zmienne nie wносиły żadnej nowej informacji, dokonano ich selekcji, wybierając z grup zmiennych te, które miały najbardziej intuicyjną interpretację oraz w których występowały najmniejsze problemy z jakością.

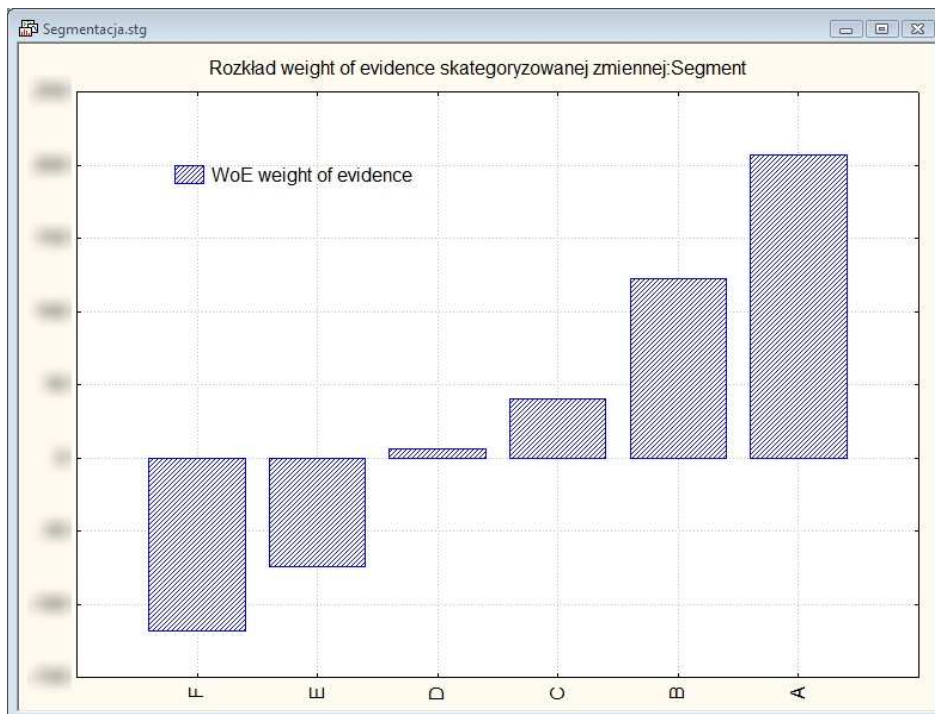
Kolejnym etapem analizy było stworzenie rankingu predyktorów na podstawie ich związku ze zmienną zależną. Przygotowany ranking predyktorów pozwolił odrzucić kolejną grupę zmiennych, których wpływ na analizowane zjawisko był znikomy.

Ze względu na poufność prezentowanych wyników wrażliwe informacje w zamieszczonych w artykule wykresach zostały „zamglone”.



Analiza uzyskanego rankingu pozwoliła na wyróżnienie zmiennej, której wpływ na badane zjawisko znacząco przekraczał graniczną wartość 0,5. Zmienna ta została więc wykorzystana jako zmienna segmentująca zbiór danych. Analizując jakość portfela kredytów w poszczególnych kategoriach analizowanej zmiennej, wyróżniono sześć segmentów ryzyka.

Segmenty A i B charakteryzujące się najniższym odsetkiem złych kredytów okazały się praktycznie wolne od ryzyka. Ponieważ liczba wyłudzonych kredytów znajdujących się w tych segmentach była niewystarczająca, aby na jej podstawie budować modele, a z drugiej strony stosowanie zaawansowanych narzędzi oceny ryzyka dla tych segmentów było bezcelowe, podjęto decyzję o nieuwzględnianiu ich w modelowaniu.

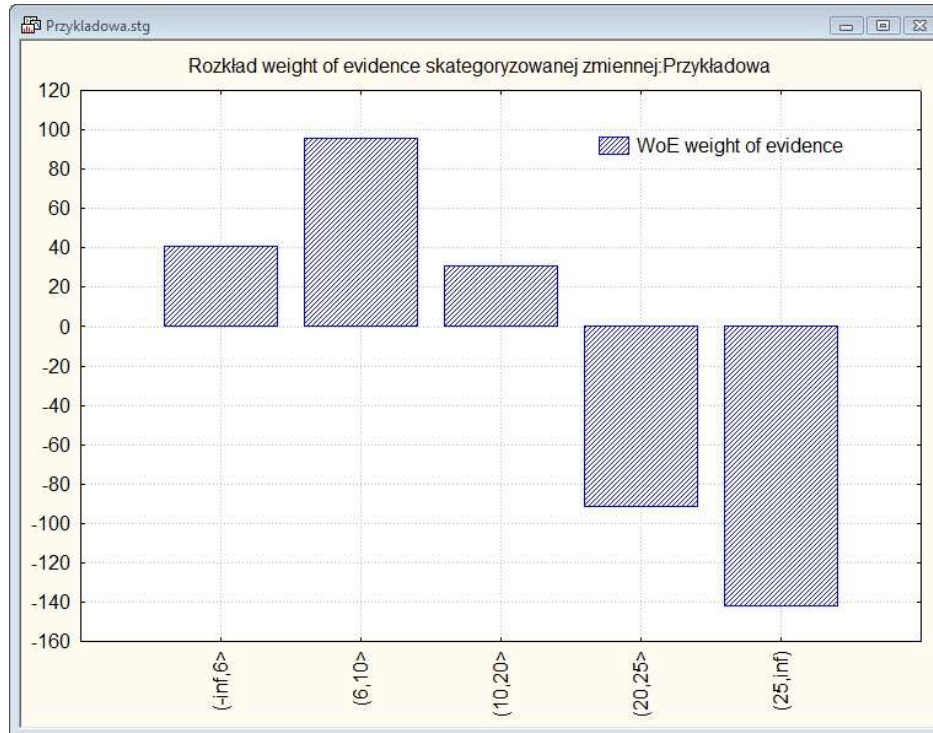


Cztery pozostałe segmenty zostały scalone do dwóch segmentów na podstawie podobieństwa merytorycznego łączonych elementów. Ostatecznie więc zbudowane zostały dwa odrębne modele. Dla każdego z analizowanych segmentów wykonano odrębny ranking predyktorów w kolejnym etapie selekcji zmiennych.

Wyselekcjonowane zmienne zostały następnie poddane procesowi dyskretyzacji, czyli przekształcenia polegającego na przygotowaniu przedziałów zmiennych o stałym wpływie na skłonność do popełniania nadużyć. Przekształcenia te przeprowadzono w module *Dyskretyzacja zmiennych* zawartym w *Zestawie Skoringowym STATISTICA*. Dla każdej z analizowanych zmiennych został przygotowany profil ryzyka oddający specyfikę wpływu poszczególnych kategorii na skłonność klientów do wyłudzenia kredytu.

Dyskretyzacja zmiennych była po segmentacji kolejnym krytycznym elementem procesu przygotowania danych. Przygotowane kategorie z jednej strony pozwalały wygładzić szumy zawarte w danych, redukując tym samym ryzyko przeuczenia modelu, z drugiej jednak strony nieprawidłowo przeprowadzony proces dyskretyzacji mógł wprowadzić dodatkowe zaburzenia, obniżając tym samym jakość budowanego modelu. Dlatego też ostateczny kształt profilów ryzyka wymagał konsultacji z ekspertami Banku.

Poniżej przedstawiono przykładowy profil ryzyka dla jednego z predyktorów. Możemy zauważyć, że przykładowa zmienna została podzielona na pięć przedziałów ryzyka. Po przejściu z przedziału $(-\infty, 6>$ do przedziału $(6, 10>$ nastąpił spadek ryzyka wyłudzenia kredytu, natomiast w kolejnych przedziałach ryzyko wyłudzenia sukcesywnie wzrastało.



Dzięki zabiegowi budowy przedziałów w naturalny sposób rozwiązany został problem wartości odstających – skrajne wartości trafiły do odpowiednich przedziałów, nie wpływając tym samym negatywnie na proces budowy modelu. W przypadku cech zawierających braki danych zostały one zamienione w osobną kategorię, dzięki czemu w naturalny sposób uwzględniono w modelowaniu możliwość istotnego wpływu braku danych na badane zjawisko.

Dyskretyzacja jest wręcz obowiązkowym przekształceniem w przypadku budowy modeli skoringowych. Ogólne kategorie znacznie redukują ryzyko, że modelowane będą losowe wahania w danych zamiast rzeczywistych zależności, które mają odbicie w zmienności analizowanego zbioru. Modelowanie na podstawie odpowiednio przygotowanych atrybutów skutkować powinno stabilnym i elastycznym modelem (mało wrażliwym na pewne wahania stabilności cech). W związku z tym model będzie się cechował większą „żywością”.

Zestaw istotnych predyktorów oraz przekształceń danych został opisany w postaci raportu z etapu przygotowania danych. Na jego podstawie w systemie Banku zostały przygotowane odpowiednie struktury danych umożliwiające późniejsze stosowanie modelu.

Wykonanie analiz

Przygotowane dane były podstawą budowy modeli przewidujących skłonność klientów do nadużyć. Dla obydwu segmentów modele skoringowe zostały przygotowane w module *Budowa tablicy skoringowej* zawartym w *Zestawie Skoringowym STATISTICA*. Modele przewidujące wyłudzenia zostały zbudowane za pomocą regresji logistycznej. Ze względu na istotne ryzyko nadmiernego dopasowania się modelu do danych, jak również w trosce

o odwzorowanie profilów ryzyka w modelu wartości pierwotne predyktorów zostały przekodowane na odpowiadające im wartości WoE.

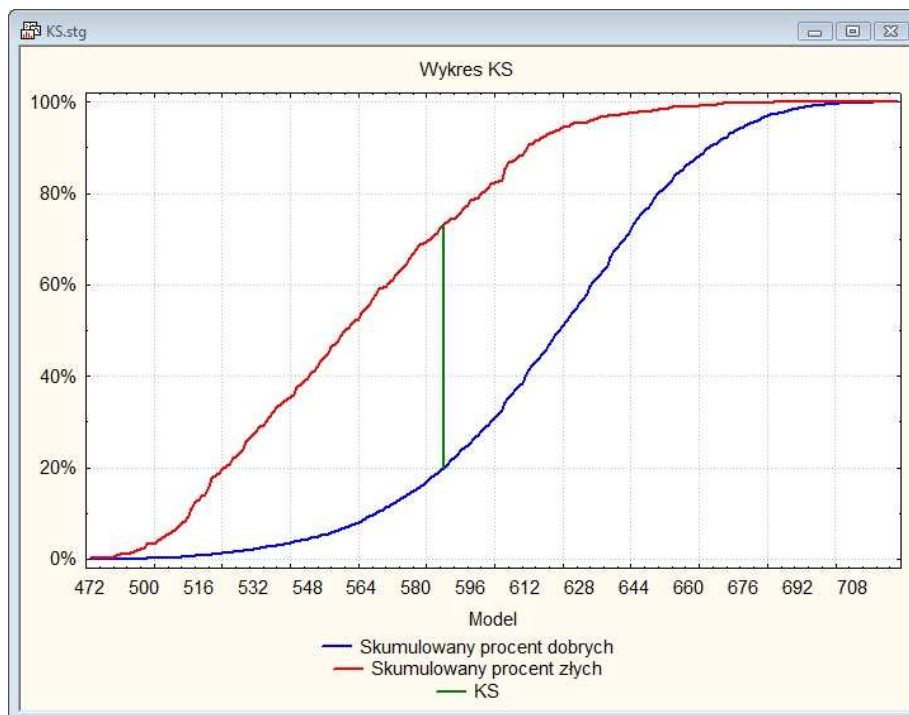
Następnie podzielono zbiór przypadków na dwa podzbiory: uczący i testowy oraz przeprowadzono ostateczną selekcję zmiennych, wspierając się metodą regresji krokowej wstecznej. W zbudowanych modelach uwzględniono kilkanaście zmiennych. Przygotowane modele zostały następnie przeskalowane do postaci tablicy skoringowej, a następnie przekazane w postaci pliku XML pracownikom Banku.

Zmienna	Zakres	WoE	Ocena	s. Walda	p	Skoring	Skoring zaokr.
...	...	42,434	0,00327	7,50403	0,00616
...	...	42,434	0,00327	7,50403	0,00616
...	...	42,434	0,00327	7,50403	0,00616
...	...	42,434	0,00327	7,50403	0,00616
...	...	-	-	-	-
...	...	3,110	0,00791	31,69831	0,00000
...	...	-59,351	0,00791	31,69831	0,00000
...	...	53,810	0,00791	31,69831	0,00000
...	...	-	-	-	-
...	...	40,558	0,00376	26,43063	0,00000
...	...	95,598	0,00376	26,43063	0,00000
...	...	31,071	0,00376	26,43063	0,00000
...	...	-91,046	0,00376	26,43063	0,00000
...	...	-142,120	0,00376	26,43063	0,00000
...	...	-	-	-	-
...	...	60,665	0,00577	102,25281	0,00000
...	...	-15,203	0,00577	102,25281	0,00000
...	...	74,399	0,00577	102,25281	0,00000
...	...	-106,970	0,00577	102,25281	0,00000
...	...	-125,210	0,00577	102,25281	0,00000
...	...	-125,210	0,00577	102,25281	0,00000
...	...	27,341	0,00577	102,25281	0,00000
...	...	27,341	0,00577	102,25281	0,00000
...	...	27,341	0,00577	102,25281	0,00000

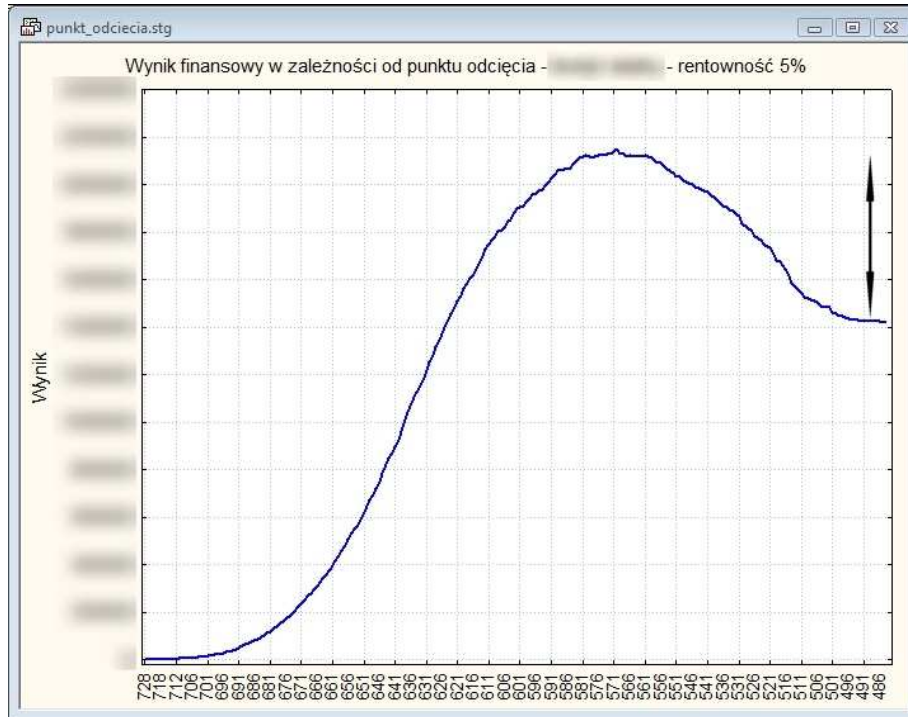
Przyjęty format modelu umożliwił łatwą jego implementację w systemie informatycznym Banku, dodatkowo pozwalając uwzględnić sytuację pojawienia się wśród cech klientów kategorii nieuwzględnionych podczas budowy karty. Zbudowane modele zostały uzupełnione zestawem reguł logicznych pozwalających na dodatkowe wsparcie procesu decyzyjnego.

Ocena uzyskanych wyników

Jakość zbudowanych modeli oceniono za pomocą standardowych miar wykorzystywanych podczas oceny modeli skoringowych, takich jak: wskaźnik Giniego, statystyka Kołmogorowa-Smirnowa oraz Lift. Ze względu na poufność wyników ich szczegółowe wartości nie zostaną przedstawione. Poniżej zamieszczono przykładowe wykresy utworzone w module *Ocena modeli w Zestawie Skoringowym STATISITCA*.



Ostatnim elementem etapu oceny była analiza polegająca na wskazaniu optymalnego punktu odcięcia. Podczas analizy przeprowadzono szereg symulacji dla kilku przyjętych poziomów rentowności kredytów. Poniżej zamieszczono wykres dla jednego z poziomów rentowności.



Wartości na osi X oznaczają poziom punktacji uzyskanej na podstawie modelu skoringowego, natomiast na osi Y przedstawiony został wyrażony kwotowo zysk płynący z portfela przyznaných kredytów. Analizując powyższy wykres, możemy zauważyć znaczącą różnicę w potencjalnej dochodowości analizowanego portfela kredytów po wprowadzeniu modelu skoringowego. Poziom zysk bez działania modelu można odczytać po prawej stronie wykresu. Stosowanie modelu dało potencjał znaczącego wzrostu zysku. Różnicę tę oddaje zamieszczona na wykresie strzałka.

Podsumowanie

Bank przyjął założenie, że w pierwszym etapie po wdrożeniu systemu niski skoring nie determinuje odrzucenia wniosku. W takich przypadkach przeprowadzane są dodatkowe weryfikacje, mające potwierdzić lub zaprzeczyć podejrzeniom wyłudzenia kredytu. Ze względu na poufność danych poziom odrzucenia wniosków będący efektem działania modeli statystycznych nie zostanie przedstawiony. W wyniku przeprowadzonych w Banku analiz potwierdzono skuteczność działania zbudowanych przez StatSoft Polska modeli, które zmniejszają ryzyko udzielenia kredytów wyłudzonych.

Dodatkowo stwierdzono, że w przypadku przeprowadzenia dodatkowych weryfikacji w około 2,5% przypadków (w stosunku do wszystkich wpływających wniosków) dochodzi do uruchomienia kredytów przy niskim współczynniku skoringu. Z tego około 2% jest kredytami dobrze spłacalnymi, jedynie w około 0,5% kredytów uruchomionych przy niskim skoringu dochodzi do zaległości w spłatach.