



BUDOWA MODELU SCORINGOWEGO DO E-POŻYCZKI Z WYKORZYSTANIEM NARZĘDZI *STATISTICA*

Kamila Karnowska i Katarzyna Cioch, SKOK im. Franciszka Stefczyka

Wykorzystanie metod scoringowych do oceny punktowej klientów ubiegających się o kredyt jest obecnie praktykowane w większości instytucji finansowych. Tę metodę szacowania ryzyka niewypłacalności potencjalnego kredytobiorcy stosuje się od ponad dwóch lat, również w Spółdzielczej Kasie Oszczędnościowo-Kredytowej im. F. Stefczyka.

Poniższe opracowanie prezentuje poszczególne etapy budowy karty scoringowej, stworzonej na potrzeby oceny klientów, ubiegających się o pożyczkę internetową w SKOK Stefczyka. Celem autorów jest przedstawienie kolejnych kroków budowy modelu, od zdefiniowania założeń produktu, poprzez wybór właściwych zmiennych do analizy, określenie sposobów utworzenia próby uczącej i testowej, aż do budowy tablicy scoringowej i oceny modelu. Ponieważ omawiany przykład dotyczy rzeczywistych danych, zwrócono szczególną uwagę na problemy, jakie wystąpiły przy pracy nad projektem. W opracowaniu omówiono problem kategoryzacji zmiennych, jakości wskaźników oceny mocy predykcyjnej zmiennych, wybór sposobu dyskretyzacji (drzewa klasyfikacyjne CHAID) oraz ostateczny wybór metody budowy modelu predykcyjnego. Podczas analizy konstrukcji karty, zaprezentowano uzyskane wyniki miar jakości modelu, w postaci statystyki KS oraz krzywych ROC.

Do utworzenia karty scoringowej wykorzystano program *STATISTICA*, ze szczególnym uwzględnieniem dodatku *Zestaw Scoringowy*.

Projekt e-pożyczka

Celem projektu e-pożyczka było zbudowanie modelu do oceny punktowej klienta, ubiegającego się w SKOK o pożyczkę konsumpcyjną za pośrednictwem Internetu.

Formuła produktu była następująca: produkt przeznaczony dla osób fizycznych na cele konsumpcyjne, kwota pożyczki nie większa niż 20 000 zł, okres kredytowania do 36 miesięcy, brak zabezpieczeń.

Budowany model opierał się na cechach najbardziej znaczących dla tej grupy klientów, którzy w swojej historii kredytowej w SKOK korzystali już z pożyczek o zbliżonych parametrach.



Model, z uwagi na specyfikę produktu, miał spełniać następujące wymagania:

- ◆ jak najmniejsza ilość informacji pozyskiwanych bezpośrednio od klienta (klient podaje swoje dane samodzielnie, dane wymagają szybkiej weryfikacji),
- ◆ automatyczna i jednoznaczna decyzja kredytowa.

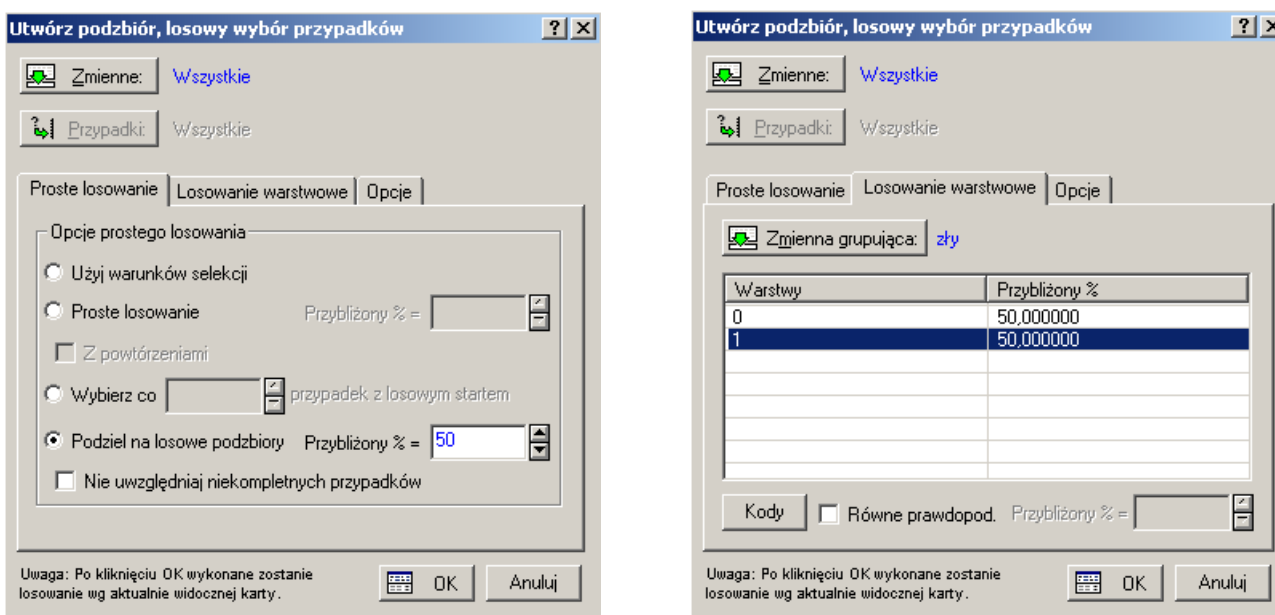
Model do e-pożyczki został wdrożony wiosną 2008 r., wraz z rozpoczęciem kampanii sprzedaży pożyczki drogą elektroniczną. Produkt e-pożyczka jest wciąż dostępny w ofercie SKOK Stefczyka (<https://www.stefczykonline.pl/>), a model, o którym mowa w opracowaniu, nadal funkcjonuje. W związku z tym, dla zachowania klauzuli poufności obejmującej dokumentację dotyczącą modelu, większość podawanych informacji jest nie uszczegółowiona lub niektóre wyniki są pominięte.

Dane do modelu

Do budowy modelu wykorzystano dane zarejestrowane w wewnętrznym systemie SKOK Stefczyka i w aplikacji archiwizującej informacje z wniosków kredytowych.

Przyjęto, że okres próby będzie wynosić 24 miesiące. Natomiast okres obserwacji kredytu powinien obejmować pierwsze 12 miesięcy od daty uruchomienia. W rozważaniach uwzględniono wnioski, dla których wydano pozytywną decyzję kredytową. Rozpatrywano pożyczki otwarte i zamknięte. Ze zbioru danych usunięto puste rekordy (braki danych). Ostatecznie w analizie uwzględniono kilkadziesiąt tysięcy danych.

Klienta „złego” zdefiniowano wg przyjętych w SKOK Stefczyka standardów. Przyjęto, że okres obserwacji wystąpienia złego zdarzenia to 12 miesięcy od momentu uruchomienia rachunku kredytowego.



Rys. 1. Tworzenie zbioru uczącego i testowego.



Zbiory uczący i testowy utworzono z zachowaniem właściwej proporcji udziału przypadków „złych” w całej próbie. Podziału dokonano, wykorzystując pakiet *STATISTICA*, w którym wykorzystano możliwość losowego podziału zbioru na dwa odrębne podzbiory z utrzymaniem stosownego udziału „złych”.

Wybór zmiennych do modelu uzależniono od wartości wskaźników predykcyjnych, wskazujących na dobre rozróżnianie klientów „złych” i „dobrych”. Sugerując się jakością mierników, odrzucono kilkanaście cech, które nie weszły do modelu. Ostatecznie do modelu przyjęto zmienne jakościowe, ilościowe oraz podwójnej precyzji.

Budowa modelu

Do budowy modelu scoringowego wykorzystano *Zestaw Skoringowy*, dostępny w programie *STATISTICA*. Program składa się z czterech modułów:

- ◆ moduł do dyskretyzacji zmiennych,
- ◆ moduł do budowy tablicy skoringowej,
- ◆ moduł do oceny i porównania modeli,
- ◆ moduł badający stabilność populacji i cech.

Dyskretyzacja zmiennych

Ponieważ tablica skoringowa jest narzędziem stosowanym dla danych dyskretnych (każda z cech uwzględnianych w modelu jest podzielona na przedziały), pierwszym etapem procesu przygotowania tablicy była odpowiednia modyfikacja danych, czyli dyskretyzacja zmiennych ciągłych i kategoryzacja zmiennych jakościowych. W związku z tym w programie *STATISTICA* w pierwszej kolejności określono typ danej zmiennej: ilościowa, jakościowa lub podwójnej precyzji. Dla zmiennych typu jakościowego oraz dla wartości nietypowych zmiennych typu podwójnej precyzji - przypisano etykiety.

Podstawowe algorytmy i wskaźniki wykorzystywane przy dyskretyzacji zmiennych

W celu określenia optymalnego sposobu podziału zmiennych na przedziały wykorzystano *metodę drzew klasyfikacyjnych CHAID*. Algorytm ten nadaje się zwłaszcza do analizy dużych zbiorów danych. W tym przypadku wykorzystano zbiór danych rzędu kilkudziesięciu tysięcy rekordów. W *Zestawie Skoringowym* analiza CHAID pozwala automatycznie określić liczbę przedziałów, które są odpowiednio liczne w sensie ilości złych kredytów.

Proces dyskretyzacji wykonywano poprzez przekształcenia interakcyjne, bazując na analizie stosownych raportów informujących o wartościach *współczynników WOE (Weight of Evidence)* i *IV (Information Value)* oraz na podstawie wykresu WOE.

Miary WOE i IV wykorzystano do oceny dobroci podziału i mocy predykcyjnej zmiennych. Wskaźnik WOE pozwolił ocenić siłę predykcyjną każdego z atrybutów analizowanej

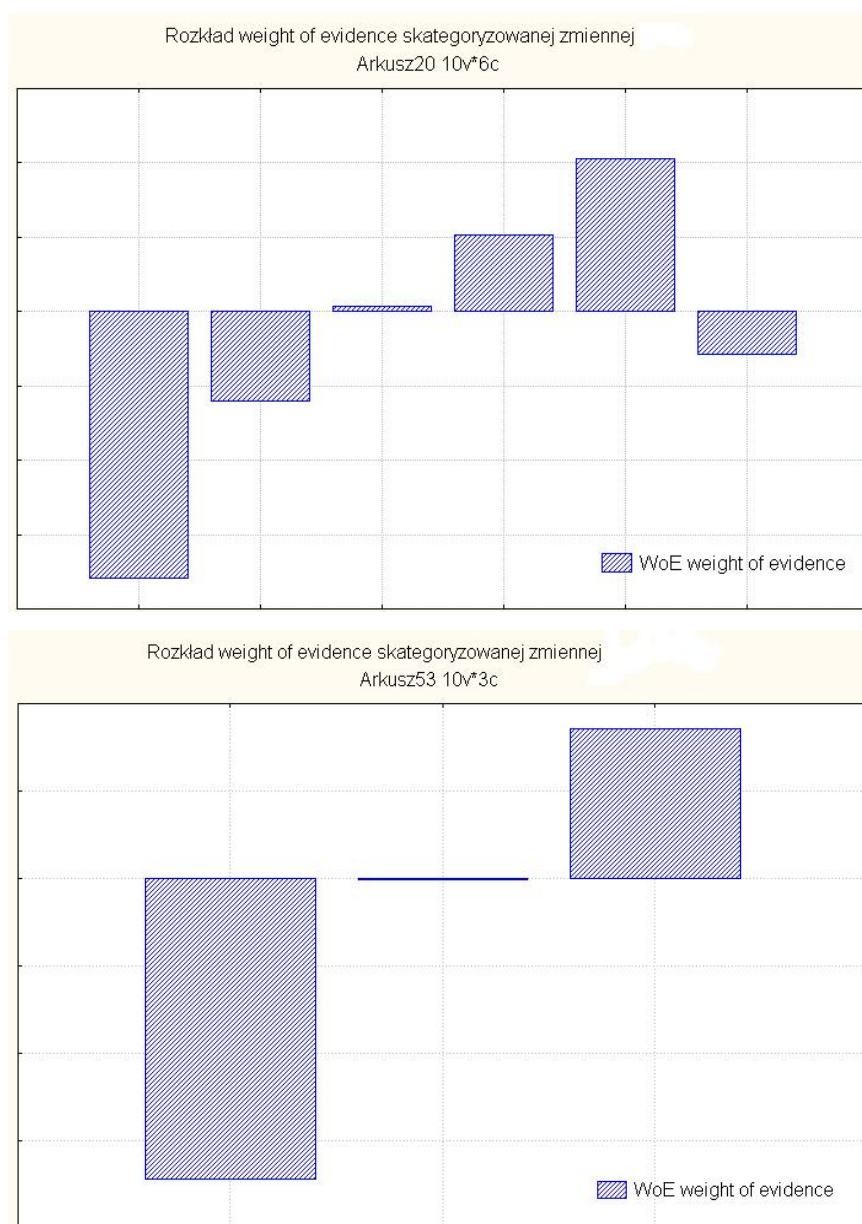


cechy, natomiast za pomocą miernika *Information Value* zbadano siłę predykcyjną całej badanej zmiennej.

Do budowy modelu użyto zmiennych o najwyższym współczynniku IV, ale zwrócono uwagę, aby wskaźnik nie przekroczył wartości 0,5. Przymuszczalnie taka zmienna o zbyt dużej sile predykcyjnej mogłaby mieć znaczny wpływ na jakość modelu (mogłaby zdominować model), gdyż niesie ze sobą zbyt duże ryzyko spadku stabilności modelu.

Raport z dyskretyzacji zmiennych

Raport z dyskretyzacji zmiennych zawierał ostateczną wersję podziału na właściwe przedziały dla odpowiednich zmiennych. Poniższe wykresy przedstawiają raport ilustrujący rozkład współczynnika WOE dla przykładowych zmiennych.



Rys. 2. Wykres WOE dla przykładowych zmiennych.



Poniższa tabela przedstawia raport z kategoryzacji przykładowej zmiennej. Podczas budowy modelu badano, ile procent obserwacji liczy każdy przedział oraz jaki jest udział procentowy „złych” klientów w każdym przedziale. Weryfikowano również, czy wykres WOE układa się w logiczny trend.

Tabela 1. Raport z dyskretyzacji przykładowej zmiennej.

Kategoryzowana zmienna									
Zmienna_kat	Dobry	Zły	Suma	Portfel	Procent złych	Procent dobrych	Procent wszystkie	IV	WoE weight
(-inf,1>	11607	668	12275	5,44%	51,15%	30,05%	30,74%	0,11	-53,18
(1,2>	7331	240	7571	3,17%	18,38%	18,98%	18,96%	0,00	3,24
(2,3>	5093	132	5225	2,53%	10,11%	13,19%	13,09%	0,01	26,60
(3,16>	11961	234	12195	1,92%	17,92%	30,97%	30,54%	0,07	54,72
(16,inf)	2630	32	2662	1,20%	2,45%	6,81%	6,67%	0,04	102,22
Ogół grp	38622	1306	39928	3,27%	100,00%	100,00%	100,00%	0,24	

Wśród badanych zmiennych rozpatrywano na przykład cechę „dobry klient”, która jest składową między innymi stażu członkowskiego w SKOK oraz jakości i długości trwania historii kredytowej. Okazała się ona zmienną dobrej jakości w sensie mocy predykcyjnej oraz pożądaną zmienną do modelu, gdyż nie wymagała pozyskiwania dodatkowych informacji od klienta.

Zweryfikowano również, jaką siłą predykcyjną wyróżnia się zmienna BIK - ocena Biura Informacji Kredytowej.

Moduł do dyskretyzacji zmiennych w *Zestawie Skoringowym* początkowo potraktował tę zmienną jako jakościową, nadając każdej z ocen osobną kategorię. To z kolei wiązało się z trudnością wykorzystania drzew CHAID i błędną dyskretyzacją zmiennej BIK. Problem został rozwiązany przez producenta *Zestawu Skoringowego* – StatSoft Polska. Ostatecznie, poprzez właściwe zdefiniowanie zmiennej (jako podwójnej precyzji) oraz wybór wartości nietypowych dla tej cechy, istniała możliwość poprawnej kategoryzacji zmiennej BIK. Otrzymano prawidłowy podział oceny punktowej oraz ocen uzupełniających, a tym samym dobrze wyznaczone wskaźniki WOE oraz IV.

Budowa modelu

W celu zbudowania modelu scoringowego z menu *Zestaw Skoringowy* wybrano moduł *Budowa tablicy scoringowej*. Wygenerowano kilka modeli, stosując różne kombinacje wcześniej ustalonych zmiennych.

W procesie budowy tablicy scoringowej wykorzystano zmienne po dyskretyzacji, na podstawie których zbudowano model logitowy. Zanim wybrano strategię budowy modelu, przekodowano zmienne na typ jakościowy. Ostatecznie model zbudowano, bazując na podstawowej **regresji logistycznej**. Na tym etapie budowania modelu jest możliwość wyboru parametrów skali punktacji.

Model matematyczny regresji logistycznej pozwala dość dokładnie opisać wpływ (prawdopodobieństwo zajścia pewnego zdarzenia) kilku zmiennych objaśniających dowolnego typu na zmienną objaśnianą, w przypadku gdy jest ona dwuwartościowa („dobry” lub „zły” klient).



W budowanym modelu skoringowym zmienna objaśniana przyjmuje wartość 0 w przypadku „złego klienta” lub 1 w przypadku „dobrego klienta”. Oznacza to, że w omawianym przypadku modelowano prawdopodobieństwo bycia „dobrym klientem”.

Weryfikacja modelu i jego parametrów - podsumowanie

Po zbudowaniu modelu przeprowadzono ocenę jego istotności oraz istotności parametrów. W tym celu wykorzystano test ilorazu wiarygodności służący do wstępnej oceny istotności całego modelu. Do zweryfikowania statystycznej istotności parametrów użyto wyników *testu Walda*.

Jakość modelu, czyli dopasowanie modelu do danych empirycznych, oceniano między innymi za pomocą miar *AIC (Akaike information criterion)* oraz *BIC (Bayesian information criterion)*.

Tablica scoringowa

Ostatecznie dla każdego z modeli uzyskano tablicę scoringową. Zawiera ona punktację scoringową dla każdego atrybutu zmiennych uwzględnionych do budowy modelu. Tablica zawiera również kategorie braku danych lub innych wartości, które nie wystąpiły w próbie, a mogą się pojawić w przyszłości.

Poniższa tabela przedstawia fragment tablicy scoringowej.

Tabela 2. Tablica scoringowa dla wszystkich kategorii przykładowej zmiennej modelu.

	1 Nazwa zmiennej	2 Zakres wartości	3 WOE	4 Parametr regresji	5 Statystyka Walda	6 Wartość p	7 Wartość score	8 Score zaokr.
...
28	ZM1	0	-22,589	0,00721	234,8337	0	88,262	88
29	ZM1	1	184,2508	0,00721	234,8337	0	131,293	131
30	ZM1	Brak danych					97,555	98
...

Probability of default (prawdopodobieństwo niewypłacalności)

Dla każdego rekordu dodatkowo wyznaczona została wartość *PD (Probability of default – prawdopodobieństwo niewypłacalności)*. Wielkość ta wyraża prawdopodobieństwo niespłacenia kredytu wynikające z modelu logitowego.

Poniższa tabela przedstawia fragment przykładowych danych, dla których wyznaczono wartość PD.

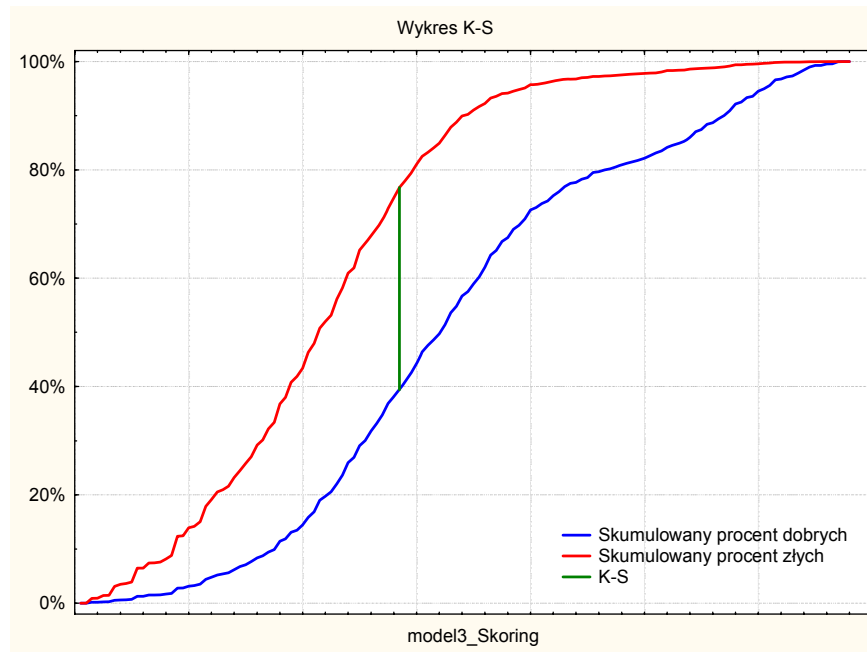
Tabela 3. Wartości zmiennych skategoryzowanych i PD – przykład.

ID	ZM1 kat	ZM2 kat	ZM3 kat	ZM4 kat	Probability of default
123456	-71,66725	37,61922	84,52684	-22,58899	0,092375

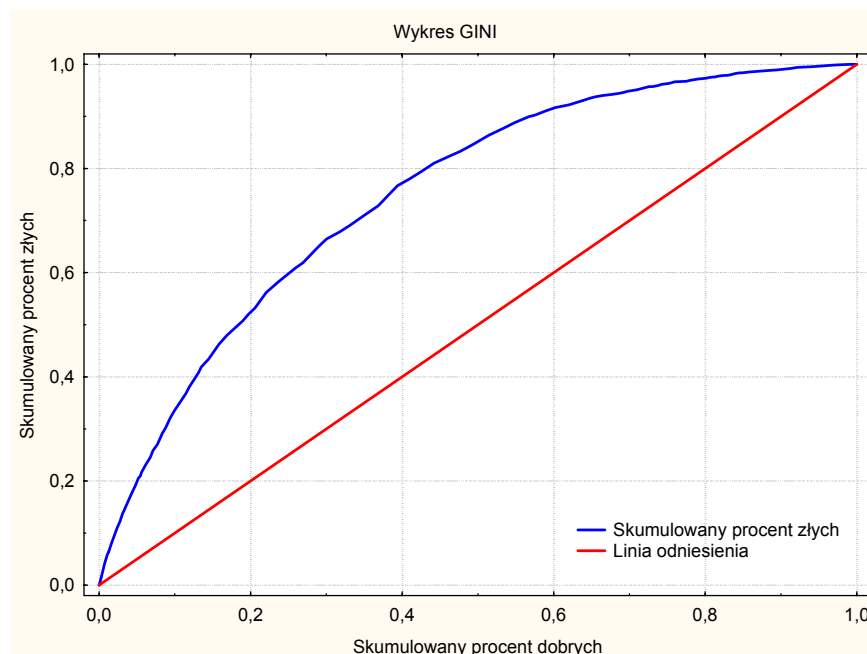


Wybór i ocena modelu

Początkowo zbudowano trzy różne modele, których jakość można było porównać w module służącym do oceny modelu. Wyboru modelu dokonano na podstawie interpretacji wskaźników: *IV*, *statystyki Kołogorowa-Smirnowa (K-S)*, *wskaźnika Giniego*, *krzywej ROC*.



Rys. 3. Wykres statystyki K-S dla badanych modeli.



Rys. 4. Wykres statystyki Giniego dla badanych modeli.

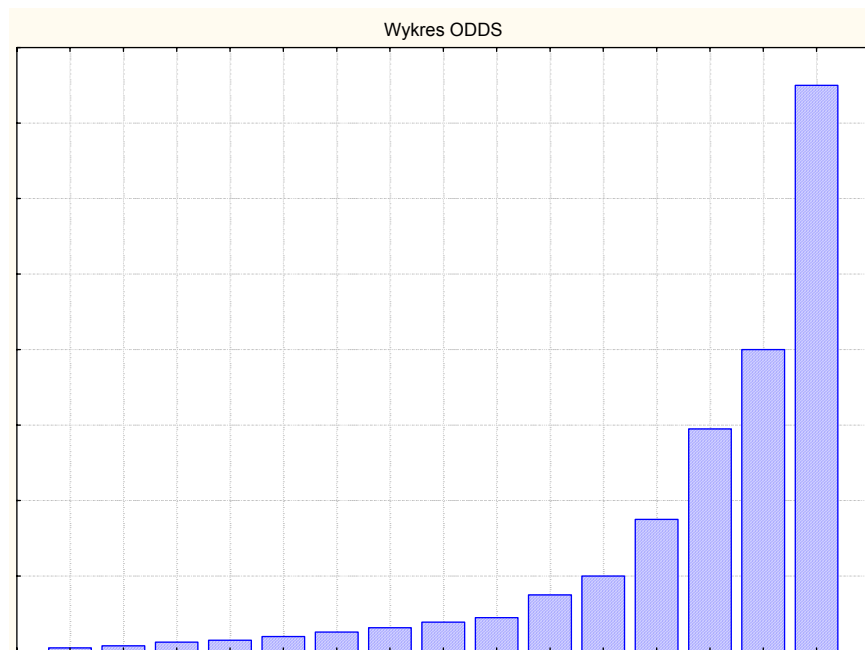


Ocena modelu - podsumowanie

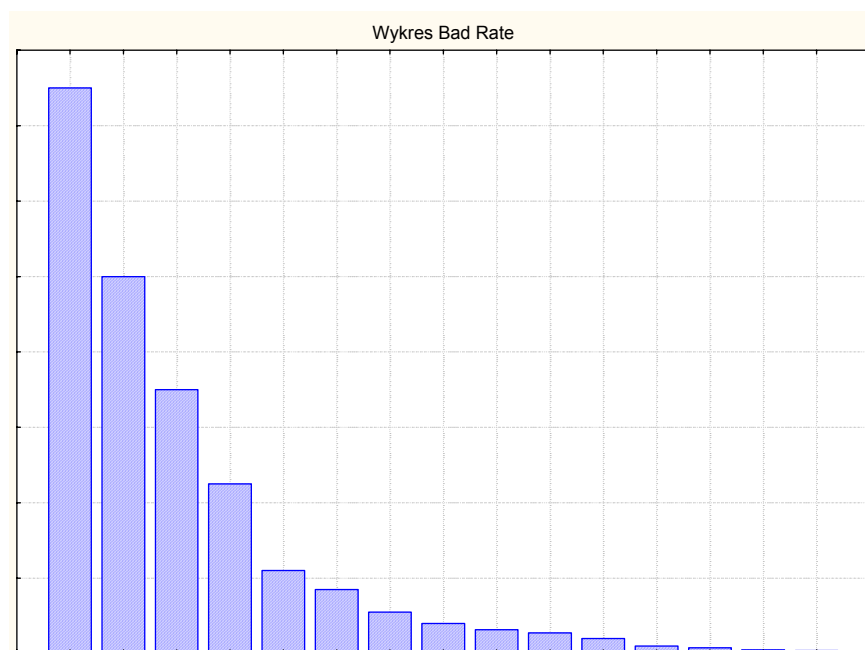
W celu ilustracji oceny modelu wykorzystano wykresy **ODDS** oraz **Bad rate**.

ODDS (odds ratio – iloraz szans) w danym przedziale wyraża stosunek ilości „dobrych klientów” do ilości „złych klientów” w tym przedziale.

Wskaźnik **Bad rate** w danym przedziale oznacza stosunek ilości „złych klientów” do ilości wszystkich klientów w danym przedziale.



Rys. 5. Wykres *ODDS*.



Rys. 6. Wykres *Bad Rate*.



Tabela punktacji scoringowej

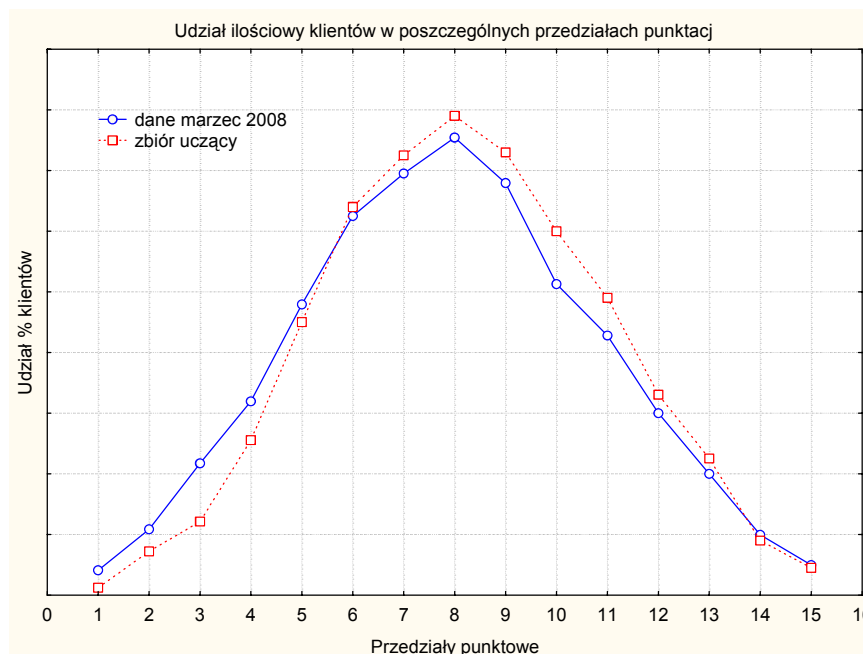
Efektem końcowym pracy była tabela, zawierająca informacje o wartościach punktowych, przypisanych przez model, oraz o wartościach PD dla wszystkich rekordów zbioru uczącego i testowego.

Tabela 4. Punktacja scoringowa modelu dla przykładowego rekordu – fragment.

LP	ID	ZM1	ZM2	ZM3	...	ZLY	ZM1_kat	ZM2_kat	ZM3_kat	...	score_ZM1	score_ZM2	score_ZM3	...	score_model	Default Probability
2	12345	1	5	2,69	...	0	-71,667	37,619	84,527	...	86	102	81	...	553	0,541

Stabilność populacji i cech

Z uwagi na fakt, że stabilność populacji bada się w dłuższym przedziale czasowym, sprawdzono jedynie i porównano udział procentowy ilości wniosków w poszczególnych przedziałach punktacji dla zbioru uczącego i danych przetworzonych przez program analizujący wnioski kredytowe w marcu 2008 r. Nie stwierdzono znacznych odchyień, co oznacza, że populacja aktualnych klientów SKOK i klientów, na podstawie których budowano model, są podobne.



Rys. 7. Udział ilościowy klientów w poszczególnych przedziałach punktacji.

W przyszłości analiza stabilności modelu scoringowego będzie obejmować:

1. Badanie stabilności procenta „defaultów” (osób, które nie dotrzymały warunków kredytowych).



- ◆ Badanie, czy procent złych kredytów maleje wraz ze wzrostem wartości punktacji (scoru) w badanych okresach czasu.
 - ◆ Badanie, czy rośnie procent złych kredytów w kolejnych okresach czasu, przy jednocześnie malejącym procencie złych kredytów dla najwyższych wartości punktacji (scoru).
2. Badanie stabilności rozkładu punktacji (scoru) - czy rozkład scoru w czasie wykazuje tendencje do zwiększania udziału kredytów o niskich wartościach punktacji scoringowej oraz zmniejszania udziału kredytów o najwyższych wartościach scoru.
 3. Badanie stabilności populacji - indeks PSI (*Population Stability Index*).
 4. Badanie stabilności rozkładów zmiennych.

Testy i wdrożenie modelu

Poprawność zdefiniowania modelu (błędy formalne) oraz jego funkcjonowania (jakość oceny) w silniku scoringowym została sprawdzona na wszystkich rekordach zbioru uczącego oraz na danych przetworzonych przez wewnętrzny system SKOK, archiwizujący wnioski kredytowe, w marcu 2008 r.

Wybór punktu odcięcia

Model do e-pożyczki miał, zgodnie z przyjętym założeniem, dzielić klientów na dwie grupy, odrzucając tych o pewnym ustalonym niskim poziomie punktacji. Dlatego należało ustalić taki poziom punktacji, aby przy akceptowalnym poziomie ryzyka (mierzonego wartością prawdopodobieństwa niespłacania kredytu przez klienta), przyjąć akceptowalny poziom wniosków odrzuconych (wyrażony jako procent wszystkich wniosków przetworzonych), z których nie każdy musi okazać się zły.

Zestaw Skoringowy nie dawał jednoznacznej odpowiedzi, jaki punkt odcięcia (wartość *cut off*) można przyjąć, aby pozostać przy dopuszczalnym poziomie ryzyka. W związku z tym wykonano analizę istniejących grup ryzyka w SKOK i przyjęto, że model powinien odrzucić klientów z grupy o najwyższym ryzyku. Za wyborem prawidłowego punktu odcięcia przemawiała specyfika produktu, a tym samym obecność wielu czynników zwiększających poziom ryzyka, charakterystycznych dla tego produktu, nie obserwowanych natomiast w innych produktach, m.in.:

- ◆ brak jakichkolwiek zabezpieczeń niezależnie od kwoty pożyczki (maksymalna kwota to 20 000,00 zł), wieku klienta, punktacji scoringowej (brak grup ryzyka),
- ◆ odejście od niektórych przyjętych w SKOK Stefczyka zasad udzielania kredytów, tzn.: nie wymaga się drugiego dokumentu tożsamości, uregulowanej służby wojskowej, zgody współmałżonka,
- ◆ pożyczka wypłacana zawsze na oświadczenie o zatrudnieniu i zarobkach – nie wymaga się zaświadczenia,



- ◆ analiza wniosków i weryfikacja oświadczeń przez nowo zatrudnione osoby bez doświadczenia,
- ◆ kanał dystrybucji produktu (sprzedaż elektroniczna - brak bezpośredniego kontaktu z klientem) i przyjmowanie oświadczeń (zamiast zaświadczeń) zwiększają ryzyko wyłudzeń.

Odrzucenie przez model wniosku klienta (odmowa udzielenia e-pożyczki) oznacza w praktyce, że klient nie może się ubiegać o ten konkretny produkt, ale może uzyskać inną pożyczkę w oddziale SKOK, na zasadach określanych przez model, funkcjonujący dla innych grup produktów.

Podsumowanie

Uzyskana dobra jakość modelu e-pożyczki, zbudowanego przy wykorzystaniu *Zestawu Skoringowego*, załączonego do programu *STATISTICA*, pozwoliła na szybkie wdrożenie produktu i jego prawidłowe funkcjonowanie do dzisiaj. Aktualnie monitorowany wskaźnik przeterminowania dla tego produktu, znajduje się na odpowiednim poziomie, co świadczy między innymi o prawidłowości działania karty i poprawności weryfikacji klientów.