KAROL PRZANOWSKI

CREDIT SCORING W ERZE BIG-DATA

Techniki modelowania z wykorzystaniem generatora losowych danych portfela Consumer Finance

OFICYNA WYDAWNICZA
SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE

CREDIT SCORING W ERZE BIG-DATA

Techniki modelowania z wykorzystaniem generatora losowych danych portfela Consumer Finance

KAROL PRZANOWSKI

CREDIT SCORING W ERZE BIG-DATA

Techniki modelowania z wykorzystaniem generatora losowych danych portfela Consumer Finance

OFICYNA WYDAWNICZA
SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE
WARSZAWA 2014

Recenzent
Krzysztof Jajuga

Redaktor
Julia Konkołowicz-Pniewska

Wydanie I

ISBN 978-83-7378-922-7

Szkoła Główna Handlowa w Warszawie – Oficyna Wydawnicza
02-554 Warszawa, al. Niepodległości 162
tel. 22 564 94 77, 22 564 94 86, fax 22 564 86 86
www.wydawnictwo.sgh.waw.pl
e-mail: wydawnictwo@sgh.waw.pl

Projekt i wykonanie okładki
Małgorzata Przestrzelska

Skład i łamanie
Karol Przanowski

Druk i oprawa
QUICK-DRUK s.c.
tel. 42 639 52 92
e-mail: quick@druk.pdi.pl

Zamówienie 84/VII/14

Mojej ukochanej żonie Małgosi
oraz dwóm synom Jasiowi i Franciszkowi,
dzięki którym dźwiganie ciężaru dnia codziennego
stało się możliwe i jest nawet radosne,
w podziękowaniu za umożliwianie mi poświęcania się pasji
Credit Scoring aż do późnych godzin nocnych

Przedmowa

W serii WHITE PAPERS publikowanej przez SAS Institute, Cary USA, *Manage the Analytical Life Cycle for Continuous Innovation. From Data to Decision*, SAS 2013, No. 106179, s.1, czytamy: „Modele analityczne leżą u podstaw najważniejszych decyzji w biznesie – wyszukiwania nowych możliwości, minimalizowania niepewności i zarządzania ryzykiem. Wobec tego przy podejmowaniu decyzji w czasie rzeczywistym i systemach operacyjnych powinny być wykorzystywane dziesiątki, jeśli nie setki, modeli predykcyjnych. Modele te powinny być traktowane jako aktywa o wysokiej wartości – którymi w istocie są. Muszą być tworzone z wykorzystaniem potężnych i pewnych procesów i zarządzane w taki sposób, aby w okresie swojej użyteczności wykazywały się jak najwyższą wydajnością. Zespoły analityczne i IT potrzebują powtarzalnych i skutecznych procesów oraz niezawodnej architektury do tworzenia i rozwijania predykcyjnych modeli analitycznych wykorzystywanych w szeroko definiowanym biznesie".

Złożoność procesu zarządzania cyklem analitycznym obejmuje następujące etapy: określenie zagadnienia, przygotowanie danych, eksplorację danych, przygotowanie modelu, walidację i przygotowanie dokumentacji modelu, wdrożenie modelu, a w końcu monitorowanie i ocenę jego jakości. Jakość, szybkość i skuteczność modelowania w erze Big Data opisywanej przez 5V *(Volume, Velocity, Variety, Veracity, Value)* jest można rzec „chlebem powszechnym" ale i „dużym wyzwaniem". Zwieńczeniem całego procesu modelowania są dobrej jakości modele predykcyjne.

Powinnością nauczycieli akademickich zajmujących się szkoleniem analityków w zakresie tego, co nazywamy Advanced Analytics and Data Science jest uczenie młodego pokolenia analityki w sposób kompleksowy, to jest w kategorii tego, co określamy *procesem zarządzania cyklem analitycznym*. W takim procesie myślenie i modelowanie statystyczne łączy się z myśleniem procesami biznesowymi, wykorzystaniem nowoczesnych technologii i Bussines Intelligence.

Książka dr. Karola Przanowskiego ***Credit Scoring w erze Big Data*** rozpoczyna nową serię publikacji adresowanych do środowisk akademickiego i biznesowego związanych z naszą coroczną konfe-

rencją **Adavanced Analytics and Data Science** organizowaną przez Szkołę Główną Handlową (Kolegium Analiz Ekonomicznych – Instytut Statystyki i Demografii – Zakład Analizy Historii Zdarzeń i Analiz Wielopoziomowych) we współpracy z SAS Institute Polska.

Tytuł publikacji, zdaniem Autora kontrowersyjny, zestawia dwa pojęcia: Credit Scoring i Big Data, które we wstępie do publikacji są szczegółowo omówione i wyjaśnione. Kolejne rozdziały to opis autorskiej konstrukcji generatora losowych danych portfela Consumer Finance i liczne przykłady modelowania skoringowego przedstawionego wyjątkowo dobrze zarówno od strony statystycznej jak i istoty oraz filozofii modelowania skoringowego. Warto podkreślić, że w procesie przygotowania publikacji wykorzystano doświadczenia i materiały z zajęć dydaktycznych prowadzonych przez Autora w ramach semestralnego przedmiotu *Credit Scoring i makroprogramowanie w SAS* dla studentów studiów magisterskich w Szkole Głównej Handlowej w Warszawie. Dla ułatwienia Czytelnikowi, Autor publikacji we wstępie daje wskazówki: *w jakiej kolejności czytać rozdziały książki*.

Zatem, serdecznie zapraszam do lektury tej i kolejnych naszych publikacji.

Ewa Frątczak

Spis treści

Wstęp

Uzasadnienie podjętego tematu

Kontrowersyjny tytuł książki stawia dwa pojęcia – dziś powszechnie znane: Credit Scoring i Big Data – trochę jakby na przeciwnych biegunach. Można by odnieść wrażenie, że Credit Scoring albo powoli usuwa się w cień ze względu na Big Data albo też nabiera nowego znaczenia w świetle nowej ery podejścia do danych i analiz. Owszem, nowa era danych nadchodzi i powoduje poważne rewolucje w myśleniu, ale Credit Scoring nadal pozostanie, ewentualnie stanie się jednym z ciekawych i dobrze rozpracowanych przykładów dla Big Data.

Czym jest Credit Scoring? Pierwotnie Credit Scoring związany był z procesem akceptacji wniosków kredytowych w bankach (Thonabauer i Nosslinger, 2004), gdzie używano prostych eksperckich kart skoringowych do wyznaczania oceny punktowej wniosku. Sposób naliczania punktów musiał być łatwy i umożliwiać, nawet mniej wykwalifikowanym analitykom, obiektywne zbadanie zdolności do wywiązania się ze zobowiązania kredytowego (Thomas *et al.*, 2002). Z nastaniem epoki komputerów oceny punktowe stały się zaawansowanymi modelami predykcyjnymi, na początku opartymi głównie na modelu regresji logistycznej. Dziś śmiało można pojęcie rozszerzyć o wiele innych metod modeli predykcyjnych, włączając w to techniki Data Mining: sieci neuronowe, drzewa decyzyjne, lasy losowe, czy też wiele innych technik ciągle się rozwijających i powodujących silną presję wyścigu w poszukiwaniu najlepszych, by wygrywać konkursy i lansować swego rodzaju modę na jedną z nich. Nie trzeba też Credit Scoringu utożsamiać tylko z bankowym procesem akceptacji. Stosuje się go także dziś w wielu innych procesach, w których klient podpisujący umowę, najczęściej zobowiązujący się do regularnych zobowiązań finansowych (takich jak abonament telefoniczny, TV itp.), musi być wstępnie oceniony w celu przygotowania najlepszych warunków umowy, by instytucja świadcząca dane usługi nie naraziła się na zbyt duże straty. Starając się najlepiej

odpowiedzieć na postawione pytanie: czym jest Credit Scoring, trzeba określić sposób przedstawienia najważniejszych jego aspektów. Wydaje się, że najlepszy sposób to studium przypadków, to wprowadzenie czytelnika w wiele istotnych, ważnych i praktycznych problemów. Tak postawione zadanie staje się głównym powodem napisania książki.

Czym jest Big Data? Niełatwo jest podać poprawną definicję tego pojęcia. Stało się dziś ono bardzo modne i większość autorów koncentruje uwagę głównie na własnościach danych, określając je jako duże i wyjątkowo zmieniające się w czasie. Mówi się o słynnych 3V (Gartner-Report, 2001; Ohlhorst, 2013; Soubra, 2012; Berman, 2013): ang. volume, czyli dużej wielkości danych, ang. velocity, czyli szybko zmieniające się oraz ang. variety, różnorodne, ze stałą strukturą i niestrukturalne, jak np. filmy, albo treści SMS. Dodaje się dziś także ang. veracity – prawdziwość czy też ang. value – wartość, aby podkreślić bogactwo cennej wiedzy ukrytej w Big Data. Jest to bardzo zastanawiające, dlaczego w definicji mówi się tylko o danych. Już Credit Scoring, pojęcie określone w latach pięćdziesiątych, stara się obejmować coś więcej niż dane. Przywołuje się tu modelowanie statystyczne oraz często podkreśla wagę wdrożenia modeli w systemach informatycznych. Podobnie ma się rzecz z Big Data. Tu szczególnie trzeba podkreślić rolę systemów zbierających i składujących dane. Co więcej przy dużych, inaczej masywnych[1] danych problemy ich przetwarzania, czy obróbki nabierają nowej jakości, stąd, mówiąc o Big Data, zawsze trzeba od razu myśleć o dużych systemach IT, bez których nie jest możliwe wydobycie cennej wiedzy o kliencie i procesie. Poprawnie zatem powinno się definiować Big Data jako układ składający się z: danych opisanych własnościami 3V (5V), metod składowania i przetwarzania danych, technik zaawansowanej analizy danych oraz wreszcie całego środowiska sprzętu informatycznego. Jest to zatem połączenie nowoczesnej technologii i teorii analitycznych, które pomagają optymalizować masowe procesy związane z dużą liczbą klientów, czy użytkowników.

[1] Trwa spór o polskie określenie Big Data, czasem pojawia się tłumaczenie: masywne dane.

Należy mocno podkreślić, że pojęcie Big Data wyrosło ze specyficznego podejścia do danych. Otóż pojawiła się istotna różnica pomiędzy pierwotną przyczyną gromadzenia danych a ich późniejszym użyciem. Owa różnica jest dziś szczególnie uważnie badana przez prawników, gdyż coraz częściej wykorzystanie danych jest nadużywane. Wprowadza się nowe pojęcie *profilowania* i bardzo prawdopodobne staje się rozszerzenie treści umów o dodatkowe klauzury zgody klientów na wykorzystanie ich danych do analiz profilowania.

Drugi aspekt wspomnianej różnicy, czyli oddzielenie potrzeby gromadzenia od używania danych, rodzi jeden z największych problemów w analizie danych, a mianowicie z założenia nie pozwala utrzymywać danych w pożądanej jakości. Powoduje to nieskończone dodatkowe problemy poprawiania jakości danych. Dziś trudno jest przewidzieć, do czego to doprowadzi.

Czy można zatem odróżnić omawiane pojęcia? Czy istnieje coś wyjątkowego, co specyficzne jest tylko dla jednego z nich? Credit Scoring pierwotnie wspomagał procesy akceptacji w bankach. Big Data pojawiło się na początku głównie w firmach rozwijających e-Usługi, takich jak Google, Amazon czy Facebook. W Polsce – w takich firmach, jak Onet czy NaszaKlasa. Ten rodzaj biznesu z założenia musiał uporać się z dużymi ilościami danych oraz z ich szybko zmieniającą się naturą.

Oba pojęcia ogólnie odnoszą się do tego samego problemu. Istota sprowadza się do lepszego zarządzania procesami, produktami i relacjami z klientami na podstawie lepszych analiz, lepszych danych. Można tu przytoczyć wiele innych podobnych nazw metod używanych dziś w biznesie, które różnią się jedynie miejscem ich powstania, gałęzią przemysłu, gdzie po raz pierwszy je zastosowano. Mamy zatem: Systemy Wspomagania Decyzji (Kwiatkowska, 2007) (ang. Decision Support System, DSS), Systemy Informowania Kierownictwa (ang. Executive Information Systems, EIS), narzędzia inteligencji biznesowej (ang. Business Intelligence, BI) – do dziś raczej nietłumaczone i używane w języku oryginalnym. W innym środowisku spotkamy się z: marketingiem zdarzeniowym (ang. Event-Driven Marketing czy Event Based Marketing) lub podejmowaniem decyzji na podstawie danych (ang. Data-Driven Decision Making), zarządzaniem relacją z klientem (Payne, 2005) (ang. Customer Rela-

tionship Management, CRM) lub zarządzaniem przedsiębiorstwem w czasie rzeczywistym (Goldenberg, 2008) (ang. Real-Time Enterprise). Choć istnieje wielość pojęć, to rewolucja Big Data zaczyna to porządkować. Ma się wrażenie, że Big Data powoli przysłania inne wcześniej używane pojęcia i być może ma to głęboki sens, byleby nie zapomnieć o podstawach.

Nawet jeśli dziś mówi się o nowej erze w kontekście Big Data, to i tak analizy skoringowe są doskonałym tego przykładem, w szczególności stosowanym przy bardzo prostym modelu biznesowym. Istnieje duże prawdopodobieństwo, że w przyszłości także dane posiadające własności 5V będą w pełni używane w Credit Scoringu. Nie będzie to raczej zmieniać metod modelowania, a jedynie rozszerzy zakres danych i najprawdopodobniej spowoduje zwiększenie jakości dyskryminacji.

Ze względu na prostotę modeli skoringowych (głównie kart skoringowych) doskonale nadają się one dla początkujących, którzy chcą rozumieć, czym jest analiza danych i jej zastosowania w biznesie, aby wyrobić sobie ważne umiejętności i nie zgubić istoty, co może się niestety zdarzyć przy bardziej skomplikowanych modelach biznesowych, strukturach danych i technikach modelowych, takich jak lasy losowe czy sieci neuronowe. Prostota daje nieocenione doświadczenie, którego później nie da się wyrobić. Można to porównać do doświadczeń z nauki komputerów. Osoby pamiętające procesory, takie jak Z80, przyznają, że nauka Asemblera, języka programowania najniższego rzędu związanego z poleceniami procesora, była łatwa i dawała wyobrażenie złożoności pracy komputera. Asembler dla obecnych procesorów jest już tak trudny, że mało kto w ogóle go używa. Podobnie rzecz ma się w przypadku statystyki. Jeśli dobrze i starannie pozna się podstawy, w szczególności wszelkie parametry testu t-studenta, włączając moc testu i minimalną wielkość próby, to zupełnie inaczej patrzy się na całą statystykę i bardziej złożone modele, takie jak uogólnione modele liniowe, modele proporcjonalnych hazardów czy modele mieszane. Szuka się wtedy w zaawansowanej teorii analogii do w pełni rozpracowanej dla testu t-studenta, rozumie się zatem lepiej, a czasem nawet żałuje, że owe tematy nie do końca zostały podjęte w obecnych teoriach. Takich przykładów można przytoczyć więcej. Jeszcze jeden bardzo ważny, pochodzi z dziedzi-

ny, którą dziś nieco się zapomina, a mianowicie z metod numerycznych. Otóż coraz mniej programistów i analityków zdaje sobie sprawę, jak wykonywane są działania liczbowe przez komputer. Wielu spodziewa się, że komputer zawsze liczy z największą dokładnością i że błędy obliczeń są wynikiem złego wprowadzania danych lub ich jakości, nie zaś samych obliczeń. Na proste pytanie zadane studentom, które działania arytmetyczne generują największe błędy obliczeniowe, najczęściej nie ma odpowiedzi. Wszystkie przedstawione rozumowania prowadzą zatem do prostego wniosku: trzeba dobrze opanować podstawy. Credit Scoring jest doskonałą nauką podstaw. Tu właśnie wykształciły się wszystkie pożądane elementy modelowania predykcyjnego, takie jak: proste modele biznesowe, rozumienie populacji, dobór próby, testowanie na różnych próbach, walidacja modeli, analiza wpływu wniosków odrzuconych, ocena modeli, kalibracja do wartości prawdopodobieństwa, wyznaczenie punktów odcięcia, testowanie strategii, implementacja w systemie decyzyjnym oraz testowanie po wdrożeniu. Cały cykl życia modelu został właśnie tu poprawnie zdefiniowany i należy się tylko uczyć od Credit Scoringu i wcielać go w innych dziedzinach zastosowań.

Z drugiej jednak strony, badając materiał teoretyczny Credit Scoringu zebrany w powszechnie znanych książkach i metodologiach ściśle chronionych przez grupy kapitałowe, można dość szybko odkryć, że wiele metod i rekomendacji powstało na podstawie pewnych doświadczeń ekspertów i nie są one udowodnione naukowo. Credit Scoring jawi się jako zestaw dobrych praktyk nieosadzonych głęboko w nauce. Jest to jedna z poważniejszych przyczyn wyboru podjętego tematu książki.

Jednym z ważnych i aktualnych problemów przy Big Data, to poprawne określenie, kim jest naukowiec od danych, czy inżynier danych, ang. Data Scientist (Kincaid, 2013). Jedną z odpowiedzi może być: to ten, który dobrze opanował podstawy analizy danych i szybko będzie w stanie uzupełnić brakującą wiedzę, kiedy spotka się z prawdziwymi problemami w życiu biznesowym. Inną odpowiedzią może być umiejętne opanowanie kilku dziedzin z odpowiednimi wagami: statystyki, by operować właściwym zestawem narzędzi zaawansowanej analizy; programowania, by samodzielnie pisać algorytmy i tworzyć zaawansowane analizy i raporty. Trzeba także znać

się na biznesie, by statystykę i programowanie umieć stosować przynajmniej w jakiejś jednej dziedzinie. Owa umiejętność związana jest z rozumieniem modeli biznesowych, czyli gdzie się traci, inwestuje i gdzie zarabia pieniądze oraz jak zgrać wszystkie wymienione procesy, by sumarycznie przynosiły zyski. Ostatnią umiejętnością jest komunikacja. Tej cechy jest nadal stanowczo za mało w dzisiejszym biznesie i dlatego na naszych oczach biznes oddziela się od informatyki (działów IT), często nie mogąc się porozumieć. Pomiędzy te dwie grupy wchodzi inżynier danych i jeśli potrafi umiejętnie przekonać obie strony do wspólnej pracy, przedstawić właściwe argumenty, często oparte na prostych, przemawiających do wyobraźni analizach, to sprawia, że firma zaczyna przekształcać się powoli z przedsiębiorstwa opartego na wiedzy eksperckiej w firmę szybko reagującą na zmianę oraz podejmującą decyzje na podstawie danych. Wtedy okazuje się, że dane zaczynają stanowić jedno z najważniejszych źródeł podejmowania decyzji i każdy z departamentów zaczyna rozumieć swoją misję.

W znanej książce o Big Data (Mayer-Schonberger i Cukier, 2013) sformułowane są dwie myśli bardzo ważne dla nowej rewolucji. Pierwsza odwołuje się do potęgi gromadzonych danych i poucza: pozwólmy mówić danym. Jest to istotnie ważny krok, gdyż wyraźnie podkreśla znaczenie danych i ich ogromną, do tej pory, niewykorzystaną moc. Prowadzi ona bezpośrednio to rozszerzenia horyzontów i rozpoczynania budowy modeli od przygotowania danych na bazie znacznie szerszych zakresów niż zwykło się to robić dotychczas. Hasła 5V prowadzą zatem do przygotowania zmiennych objaśniających, które mają przewidzieć badane zjawisko, na bazie wszelkiej istniejącej i dostępnej informacji w gromadzonych bazach, nawet jeśli z pozoru nie istnieje jakakolwiek przesłanka, czy logika przyczynowo-skutkowa. Druga myśl przytoczonej książki budzi niepokój, mianowicie stwierdza: nieważne dlaczego, ważne, że coś działa (ang. not why, but what). Jeśli tylko udaje się zarobić większe pieniądze, jeśli tylko uzyskuje się lepsze narzędzia optymalizacyjne, to nie musimy ich rozumieć, wystarczy, że działają. Takie rozumowanie jest bardzo niebezpieczne. Możemy sobie wyobrazić, że w przyszłości zapomni się o logice i przyczynowo-skutkowej weryfikacji zależności, a tym samym doprowadzi narzędzia analityczne

do roli automatów bez ingerencji analityka. Automaty będą używane w systemach informatycznych i być może instalowane przez pracowników technicznych nieznających się na analizie danych. W szybkim czasie może to doprowadzić do inwigilacji społeczeństwa lub manipulacji, a zawód inżyniera danych będzie najbardziej poszukiwany, gdyż automaty będą się psuły, algorytmy nie wytrzymają próby czasu i trzeba będzie je zmieniać. Byleby tylko firmy sprzedające owe automaty miały tego świadomość.

Fakt, że będziemy mieli coraz więcej zmiennych objaśniających, nie musi być niebezpieczny, byleby w jakimś etapie budowy modelu predykcyjnego dokonać przeglądu odkrytych reguł i wyeliminować pozorne. Właśnie dlatego Credit Scoring ponownie staje się dobrym przykładem, gdyż są tu automaty, które pomagają przyspieszyć etapy złożonych obliczeń bez absorbowania analityka, a także takie etapy, gdzie praca analityka jest jedyna w swoim rodzaju i nie może być zastąpiona przez komputer.

Początki Credit Scoring sięgają lat 50. XX w., kiedy firma konsultingowa o nazwie Fair Isaac & Company stworzyła pierwszy komercyjny system skoringowy (Poon, 2007). Pierwsze ważne argumenty optymalizacji koncentrowały się wokół haseł: szybciej, taniej i obiektywniej (Mester, 1997), ale taniej głównie dzięki eliminacji ręcznej pracy w ocenianiu wniosków kredytowych. Dziś przytoczone hasła są niepodważalne i oczywiste, natomiast nadal zbyt rzadko wykazuje się potęgę optymalizacyjną modeli skoringowych w kontekście przymnażania zysku, kapitału, co zostało pokazane w podrozdziale 2.1.

W książce głównie koncertujemy się na statystycznych modelach oceny punktowej, zwanych także kartami skoringowymi, z ang. credit scorecard lub ogólniej Credit Scoring: (Thomas *et al.*, 2002; Anderson, 2007; Matuszyk, 2008). Najczęściej modele te tworzone są na bazie regresji logistycznej. Ich konstrukcja jest dość prosta oraz łatwa w interpretacji i dlatego na stałe modele te zagościły w optymalizacji wielu procesów instytucji finansowych. Znalazły one szczególne zastosowanie w bankowości (Huang, 2007) do optymalizacji procesów akceptacji produktów kredytowych i modeli PD (ang. probability of default) stosowanych w rekomendacjach Basel II

i III do liczenia wymogów kapitałowych RWA (ang. Risk Weighted Assets) (BIS–BASEL, 2005).

Celem napisania książki jest głównie próba stworzenia uniwersalnego repozytorium danych Credit Scoring i wykazanie jego przydatności do rozwijania technik modelowych, by stworzyć narzędzia do przeprowadzania dowodów naukowych. Aby rozwijać badania nad Credit Scoring, trzeba po pierwsze mieć dobre dane – to jest punkt startowy. Podjęte są tu następujące wyzwania:

- Czy możliwe są badania Credit Scoring bez konieczności posiadania rzeczywistych danych?
- Czy możliwe są sposoby dowodzenia wyższości jednej techniki modelowej nad drugą w oderwaniu od konkretnej reprezentacji danych?
- Czy można stworzyć ogólne repozytorium danych i na jego bazie prowadzić różnego rodzaju badania?
- Czy można w zarządzaniu procesem akceptacji kredytowej minimalizować wpływ wniosków odrzuconych? Czy można, pomimo braku danych o tych wnioskach, poprawnie estymować ryzyko kredytowe?
- Czy można stworzyć, na bazie repozytorium, metody i teorię zarządzania strategiami w procesie akceptacji kredytowej?

Podstawową tezą stawianą i udowadnianą w całej pracy jest stwierdzenie, że dane symulacyjne, choć nie mogą zastąpić rzeczywistych, to są bardzo użyteczne w rozumieniu złożoności procesów w instytucjach finansowych i dają szanse stworzenia ogólnej teorii do porównywania technik modelowych. Wydaje się pozornie, że dane symulacyjne upraszczają rzeczywiste procesy. W głębszym sensie pozwalają jednak rozważyć więcej przypadków, gdyż ich konstrukcja związana jest z odpowiednią listą parametrów i założeń. Każda ich modyfikacja daje kolejny nowy układ danych, a rozważenie wszystkich możliwych kombinacji daje szersze spektrum niż dane rzeczywiste zaobserwowane tylko w kilku instytucjach finansowych.

Poza głównym wątkiem wykazywania przydatności danych symulacyjnych w książce zaprezentowane są wszystkie najważniejsze

problemy Credit Scoring oraz etapy budowy modelu bardzo szczegółowo ujęte właśnie dzięki danym symulacyjnym, gdyż można publikować ich dowolne raporty liczbowe bez narażenia się na ujawnienie danych wrażliwych.

Związek z procesami biznesowymi z jednej strony czyni Credit Scoring dziedziną popularną i znaną, z drugiej – utrudnia jej pełny rozwój w oderwaniu od wpływu dużych korporacji i firm konsultingowych. Przepisy chroniące dane praktycznie uniemożliwiają pełne i rzetelne studia konkretnych układów danych.

Modele Credit Scoring są szczególnym przypadkiem statystycznych modeli predykcyjnych służących do prognozowania zjawisk na podstawie dotychczasowej zaobserwowanej historii danych. Najlepszym sprawdzianem ich użyteczności i poprawności jest zatem testowanie prognozy z rzeczywistymi wynikami. Niestety często, aby przeprowadzić tego typu testy, potrzeba czasu, nawet kilku lat. W skrajnych przypadkach, jeśli chce się obserwować pełny cykl życia nawet zwykłych kredytów, takich jak kredyt ratalny, potrzeba przynajmniej pięciu, a może i dziesięciu lat, jeśli chce się uwzględnić także wszystkie etapy procesów windykacyjnych, włączając prace komorników po wypowiedzeniu umowy.

Obserwacja cyklu koniunkturalnego, choć jesteśmy już po kolejnym dużym kryzysie (Benmelech i Dlugosz, 2010; Konopczak *et al.*, 2010), nadal nie wydaje się tak prosta. Jak podają raporty NBP, obecny czas odnotowuje wyjątkowo niskie wartości ryzyka kredytów konsumenckich. Nikt jednak nie jest w stanie zagwarantować, że kryzys nie powróci. Konsekwencje rekomendacji T i rozwinięcia się parabanków ciągle nie są do końca zbadane. Pojawia się ciekawy problem niereprezentatywności danych rynku kredytowego w bazach Biura Informacji Kredytowej (BIK) i warto temu poświęcić obszerniejsze badania. Obecny kryzys ekonomiczny skłania także wielu badaczy ku poszukiwaniu lepszych modeli predykcyjnych, bardziej stabilnych w czasie (Mays, 2009).

Rozważmy teraz sytuację najczęściej pojawiającą się w rozpoczynaniu pracy naukowej z dziedziny statystyki stosowanej. Z reguły problem zaczyna się i kończy na banalnym pytaniu: skąd wziąć rzeczywiste dane? Niektóre dyscypliny, w szczególności zastosowanie statystyki w medycynie, nie mają takiego problemu, przynajmniej

nie jest on tak istotny. Dane medyczne są dość powszechnie dostępne. Rzecz ma się zupełnie inaczej w stosunku do danych bankowych. Z reguły trzeba występować z formalnymi podaniami do członków zarządów banków i nie zawsze odpowiedzi są pozytywne, albo też pozwolenie otrzymuje się na bardzo zafałszowane dane często pozbawione interpretacji. Nawet jeśli uda się już takie dane pozyskać, to ich wielkości często nie spełniają pożądanych oczekiwań. W najnowszej publikacji prezentowanej na konferencji „Credit Scoring and Credit Control XIII" w Edynburgu (Lessmanna *et al.*, 2013) zebrano prawie cały dorobek badań nad Credit Scoring z ostatnich dziesięciu laty. Wymienia się tu dziesiątki danych, na których budowano różne modele, stosując bardzo wiele technik, włączając w to także Data Mining. Niestety tylko kilka wymienionych danych zawierało więcej niż kilkadziesiąt tysięcy wierszy, a tylko jeden testowany zbiór miał więcej niż sto charakterystyk. Dane te jednak nie są powszechnie dostępne. Autorzy opisali także wszystkie publicznie znane dane do badań Credit Scoring. Wymieniono tylko siedem pozycji, gdzie tylko jeden zbiór miał 150 tysiecy wierszy, a inny – 28 charakterystyk. Słusznie zatem (Kennedy *et al.*, 2011) piszą o potrzebie tworzenia danych symulacyjnych.

Przeprowadźmy teraz proste rozumowanie teoretyczne. Przypuśćmy, że chcemy udowodnić poprawnie naukowo, że modele kart skoringowych oparte na metodzie WoE (Siddiqi, 2005) (ang. Weight of Evidence) są najlepsze wśród obecnie stosowanych. Nawet jeśli (Lessmanna *et al.*, 2013) argumentują, że już inne metody, takie jak: ang. bootstrap sampling with a greedy hill-climbing (HCES-Bag), lasy losowe, czy sieci neuronowe, wypierają regresję logistyczną, to i tak pozostaje ogólne pytanie: jak przeprowadzić poprawne dowodzenie wyższości jednej techniki modelowej nad drugą? W cytowanej pracy modele porównano na kilku zbiorach. Jest to i tak dość oryginalne podejście, gdyż większość prac naukowych z dziedziny statystyki stosowanej z reguły opisuje tylko jeden przypadek danych rzeczywistych i na jego bazie próbuje wyciągać ogólne wnioski. Można stwierdzić, że większość prezentacji z wielu lat konferencji „Credit Scoring and Credit Control" właśnie w ten sposób powstawało, w szczególności (Malik i Thomas, 2009; Huang i Scott, 2007).

Spróbujmy nasz dowód wyższości WoE nad innymi technikami sprowadzić do czysto matematycznego języka. Można by wtedy sformułować problem w następujący sposób: $Q(n) > W(n)$ dla każdego $n \in N$, czyli że nierówność jest prawdziwa dla wszystkich liczb naturalnych. Dowód teoretyczny można by przeprowadzić, stosując zasadę indukcji matematycznej. Wracając do języka statystyki, należało by przy zastosowaniu statystyki teoretycznej przeprowadzić dowód wyższości WoE na wszystkich możliwych zbiorach. Tego typu zadanie jest jednak niemożliwe; można by tu przytoczyć tezy Cantora, że nie istnieje zbiór wszystkich zbiorów. Nawet jeśli pominie się tezy z teorii mnogości, to na gruncie ekonomii czy statystyki matematycznej nie wydaje się możliwe jakiekolwiek podejście do tego problemu. Można jedynie przeprowadzać dowody własności pewnych funkcji, czy rozkładów, zakładając podane z góry rozkłady wejściowe. Zagadnienie sprowadza się zatem do badań nad przykładami danych rzeczywistych, czyli do uprawiania statystyki stosowanej. Niestety tego typu czynność można by nazwać udowadnianiem prawdziwości nadmienionej nierówności tylko dla przykładowej wartości liczby naturalnej, chociażby dla liczby jeden.

Jeśli zatem niemożliwe staje się rozważenie wszystkich zbiorów, a kilku przykładowych jest zbyt trywialne, to powstaje potrzeba stworzenia czegoś pośredniego. Mianowicie możliwość rozważenia zbiorów najbardziej typowych, spotykanych w rzeczywistości oraz ich wielu uogólnień. Jak należy to rozumieć? Jeśli z możliwych dostępnych danych pobierze się wzorce rozkładów i zależności pomiędzy cechami, to obserwując te wzorce, można stworzyć więcej kombinacji niż pojawiły się w rzeczywistości. Jeśli tylko wzorce będą poprawnie zbadane, będziemy mieli możliwie najbliższe rzeczywistości układy danych, a jednocześnie także ogólniejsze z racji ich różnych kombinacji. O tego typu danych będzie można pisać w sposób bardzo precyzyjny, podając ich parametry, własności rozkładów itp. Będzie zatem możliwe przeprowadzanie dowodów nie tylko na zasadzie, która technika modelowa jest lepsza, ale także przy jakich parametrach. Być może na razie przedstawione rozumowanie wydaje się być utopijne, ale nie jest łatwe znalezienie innej drogi. Jeśli chcemy być bardziej wiarygodni w metodach porównawczych, musimy testować na większej liczbie zbiorów, musimy nauczyć się tworzyć

lepsze mierniki, lepsze sposoby mierzenia różnic pomiędzy sposobami modelowania. Sam powód poszukiwania poprawnych kryteriów wydaje się celem samym w sobie.

Przypuśćmy, że posiadamy już stworzone dane symulacyjne. Znamy zatem dokładny przepis ich tworzenia. Nawet jeśli używana jest tu symulacja Monte Carlo, wprowadzająca element losowy, to i tak jest on w pełni deterministyczny. Każdorazowe uruchomienie generatorów losowych może zwracać dokładnie ten sam ciąg losowych liczb. Możemy zatem stwierdzić, że znamy każdą liczbę i każdy rozkład takich danych. Budowanie modeli statystycznych na tego typu danych wydawać się może absurdalne, gdyż modele te odkryją, czy też wykażą dokładnie te reguły, które były podstawą do ich tworzenia. Jednak nawet tak spreparowane dane zaczynają mieć własności, które zaskakują, nie były przewidywane, a jednak się pojawiły.

Złożoność procesu budowy takich danych powoduje, że nawet sam autor nie panuje nad wynikiem. Dobranie wszystkich współczynników jest nie lada wyzwaniem. Tworzy się zatem układ, który staje się nie do końca znany i trzeba się go uczyć, badać, tak jakby po raz pierwszy się pojawił, nic o nim wcześniej nie wiedząc. Można by tu zacytować prof. Grzegorza Kołodko: „rzeczy dzieją się i wiele rzeczy dzieje się na raz". Owe wiele rzeczy na raz powoduje, że dane przestają być zrozumiałe i posiadają swego rodzaju tajemnicę, którą trzeba odkryć. Trzeba odkryć własności, których nie planowaliśmy, a które przy okazji powstały, bo wiele wielkości zależy od siebie.

Wszystkie przytoczone argumenty skłaniają do głębszych badań nad danymi symulacyjnymi i powodują także zmianę swego rodzaju paradygmatu w statystyce stosowanej. Nie musimy zaczynać pracy naukowej od pytania skąd wziąć rzeczywiste dane. Być może należy problem w ogóle inaczej sformułować. Mianowicie: jakie dane potrzebne są do poprawnego działania danego procesu. Jak pokazać, że posiadanie takich a nie innych danych jest wystarczające do wyznaczenia trafnej prognozy? Mylne bardzo jest oczekiwanie, że dane rzeczywiste informują nas istotnie więcej o procesie niż dane symulacyjne. Obserwowane zjawiska nie ujawniają zmiennych ukrytych. Obserwowalne nie oznacza możliwe do wyjaśnienia. Tak czy inaczej potrzeba jest poszukiwania zmiennych ukrytych.

Potrzebę tworzenia danych symulacyjnych można sformułować jeszcze inaczej. W typowym zarządzaniu procesem akceptacji kredytowej standardowo wykonuje się raporty takie, jak: przyczyny odrzutów, raporty vintage, badania profili klientów, raporty flow-rate itp. Są to właśnie obserwowane rozkłady. Jakie zmienne ukryte, jakie procesy ukryte tworzą takie wyniki? Jak złożone muszą być zależności, aby uzyskać zbliżone do rzeczywistych wyniki? Zadanie nie jest trywialne. Każdy analityk budujący modele kart skoringowych po pewnym czasie doświadczenia jest w stanie przytoczyć bardzo wiele reguł zależności w danych. Nawet jeśli zmienia bank i pracuje w zupełnie nowym środowisku, ponownie odkrywa zbliżone reguły. Różnią się one co do bezwzględnych wartości, ale nie w ogólnym fakcie. Zawsze emeryci spłacają lepiej kredyty niż inni. Z reguły klienci z większym wynagrodzeniem spłacają lepiej, aczkolwiek już przy dużych pensjach jest inaczej, bardzo bogaci potrafią mieć nieregularne płatności. Jak teraz to pogodzić: z jednej strony emeryt zarabia mało, czyli nie powinien dobrze spłacać, ale z drugiej obserwuje się, że spłaca dobrze. Czy da się stworzyć takie dane, aby spełniały obie własności? Nie jest to takie oczywiste i samo poszukiwanie modeli, które temu sprostają staje się ważnym elementem pracy badawczej przybliżającej nas do rozumienia bankrutów.

Kim jest bankrut? Całe zadanie tworzenia danych Credit Scoring jest próbą odpowiedzi właśnie na to pytanie. Czy profil bankruta jest tylko jeden? Czy zawsze tak samo się zachowuje w czasie? Czy zawsze tak samo reaguje na zdarzenia? Nigdy nie dowiemy się, jeśli nie zaczniemy budować symulatorów.

Ostatni argument pochodzi z innej dziedziny. W badaniach sieci telekomunikacyjnych, dokładnie w planowaniu ruchu w sieci już od wielu lat znane są symulatory. Istnieją wyspecjalizowane oprogramowania takie jak OPNET [2]. Badania sieci testuje się łatwo i szybko. Sieć projektuje się, składając ją z zaprogramowanych węzłów, gdzie każdy z nich może generować inny, z góry zadany rozkład ruchu. Wydaje się, że badanie procesu akceptacji kredytowej też powinno zaowocować gotowymi układami, które przy z góry zadanych parametrach wygenerują ruch klientów składających wnioski kredytowe,

[2] OPNET Technologies Inc. (`http://www.opnet.com`).

spłacających lub nie i wreszcie obliczą wszelkie typowe miary, takie jak: opłacalność procesu, jaki musi być optymalny punkt odcięcia, jak sterować parametrami cenowymi itp.

Opis zawartości rozdziałów

Książka rozpoczyna się od rozdziału 1, w którym przedstawiony jest szczegółowy opis tworzenia symulacyjnych danych podobnych do portfela Consumer Finance. W podrozdziale 1.1 przedstawiono uproszczony opis całego algorytmu. W każdym z podrozdziałów opisane są kolejne etapy tworzenia danych. Finalnie powstają dane związane z dwoma produktami: kredytem ratalnym i gotówkowym oraz zmieniające się w czasie dane klientów. Wszystkie te dane są wzajemnie powiązane, stąd konieczność opisu generatora w wielu podrozdziałach, w których szczegółowo pokazuje się zależność przyczynowo-skutkową pomiędzy danymi historycznymi a nowo powstałymi, uwzględniając oczywiście generatory liczb losowych. Jest to istotny rozdział książki, gdyż bez dobrych danych nie byłoby możliwe przedstawienie wielu ciekawych studiów przypadków opisanych w następnych rozdziałach.

W rozdziale 2, omawiane są dwa przykłady analiz, które wymagają uproszczonego generatora danych ograniczonego tylko do jednego produktu. Pierwszy przykład opisany w podrozdziale 2.1 przedstawia metodę wyznaczania zysku ekstra dla banku z procesu akceptacji kredytowej, dzięki zwiększeniu mocy predykcyjnej modelu. Jest to pierwsze tak śmiałe badanie wiążące zmianę mocy predykcyjnej modelu skoringowego z finalnymi zyskami banku. Dotychczas wielu autorów formułowało ogólne wnioski, że modele odgrywają zasadniczą rolę w zmniejszaniu straty, ale nie było to nigdy pokazane do końca na konkretnych liczbach. Podrozdział ten stanowi zatem cenną wiedzę, która może pomóc w wielu strategicznych decyzjach związanych z szacowaniem potencjalnych zysków i kosztów budowy nowego modelu. Kolejny podrozdział – 2.2 stanowi ważną naukę dla każdego analityka budującego modele. Dotychczas nie została poprawnie i naukowo zdefiniowana metoda porównawcza technik skoringowych. Z reguły ograniczano się do porównywania kliku modeli na jednym przykładowym zbiorze danych, tymczasem w podroz-

dziale tym można znaleźć gotowy przepis na bardziej wiarygodną metodę, która jest w stanie znacznie subtelniej porównywać techniki modelowe, a także uczynić to porównanie bardziej uniwersalnym wnioskiem, niezależnym od danych, na których je wykonano. Przedstawiony schemat postępowania wyraźnie wykazuje wyższość kilku metod budowy kart skoringowych nad innymi oraz jednocześnie jest szansą dla badań nad Credit Scoringiem, by stworzyć solidne naukowe podstawy do porównań.

W rozdziale 3, omówiony jest szczegółowo przypadek zarządzania portfelem Consumer Finance dla modelu biznesowego dwóch produktów: kredytu ratalnego jako akwizycji i kredytu gotówkowego związanego ze sprzedażą krzyżową. Poruszony jest tu ważny temat szukania optymalnych strategii akceptacji obu produktów, gdyż są one ze sobą mocno powiązane. Nie można traktować obu tych produktów oddzielnie. Kredyt ratalny jako akwizycja jest oferowany klientom, by pozyskać ich dla banku i związane jest to najczęściej z niskimi marżami dla banku. Klient musi odnieść wrażenie, że mu się opłaca korzystać z usług naszego banku. Dopiero na kredycie gotówkowym bank jest w stanie zarobić na tyle dużo, by pokryć stratę z kredytu ratalnego. Nie można zatem optymalizować samego procesu kredytu ratalnego, bo liczy się finalny zysk na obu produktach, a na samym ratalnym możemy zarabiać, np. tylko akceptując 10% populacji wniosków kredytowych, co jest niemożliwe ze względu na konkurencję innych banków. Niestety tylko bank z dużym procentem akceptacji ma szanse pozyskiwać dużą liczbę dobrych klientów. Przy okazji badania optymalnej strategii po raz pierwszy w książce uwydatniony jest problem wpływu wniosków odrzuconych (ang. Reject Inference), który dzięki danym symulacyjnym może być pokazany szczegółowo w liczbach. Cały rozdział, włączając w to także ciekawe rozumowanie podrozdziału 3.3.2, jest ważnym przykładem udowadniającym jego istotność, pozwalającym uświadomić sobie jego skutki oraz przyczynę powstawania. Problem ten do dzisiejszego dnia nie jest do końca rozwiązany, choć istnieje już spora lista publikacji, to ma się wrażenie, że nie ma sensownego pomysłu, jak dalej tę dziedzinę rozwijać. Być może dane symulacyjne i prezentowany rozdział staną się właśnie nowym i dobrym pomysłem rozpędzającym naukę w kierunku lepszych badań nad Reject Inference.

W kolejnym, rozdziale 4, przeprowadza się czytelnika przez wszystkie etapy budowy modelu akceptacyjnego, czyli używanego do akceptacji wniosków kredytowych procesowanych w systemie decyzyjnym. Najczęściej w podręcznikach z Credit Scoring, lub ogólnie poruszających tematykę modeli predykcyjnych, czy Data Mining, omawiane są różne techniki, ale rzadko prezentowane są wyniki pośrednie oparte na konkretnych danych. Dzięki danym symulacyjnym istnieje jedyna okazja pokazania każdego etapu w sposób pełny i zawierający konkretne raporty i liczby. Co więcej, przedstawiona jest tu pełna metoda radzenia sobie z problemem wpływu wniosków odrzuconych. Zawsze nowy model budowany jest na danych, które zostały podzielone wcześniej przez stary model na dwie populacje: wniosków zaakceptowanych, gdzie możemy obliczyć statystykę spłacalności kredytów oraz na wnioski odrzucone, gdzie takiej możliwości obliczenia statystyki nie mamy. Finalny model karty skoringowej musi mieć własności rozróżniania klientów, pod kątem spłacania kredytów, na całej populacji przychodzącej, włączając w to także populację wniosków odrzuconych przez stary model. Zadanie to nie jest proste. Przedstawiona jest tu metoda jak i jej pełna krytyka, dzięki której z dystansem będzie się podchodzić do metod Reject Inference, by nie traktować ich jako złotych zasad, które zawsze przyniosą pożądane efekty. Co więcej, dzięki owej krytyce czytelnik może wyrobić sobie swoją własną intuicję i być może nawet samodzielnie zdefiniować swoją metodę dopasowaną do kontekstu danego projektu budowy modelu. Dodatkowo w podrozdziale 4.5 poruszony jest bardzo ważny temat segmentacji (tj. podziału populacji na segmenty) i budowy dedykowanych modeli dla segmentów. Ogólna idea sprowadza się do prostego przesłania, że nie da się zbudować modelu dobrego dla wszystkich wniosków, ale jak się ich zbiór podzieli na właściwe segmenty, to dla każdego z nich można dobrać bardziej subtelne zmienne i finalnie zbudować lepsze modele. Wniosek ten jest ogólnie znany, problem tkwi jedynie w tym, że nie do końca wiadomo, na ile i jakie segmenty dzielić populację i czy na pewno da się zbudować na każdym nich lepsze modele. Przedstawione w rozdziale rozumowanie i różnego rodzaju porównania pomiędzy modelami i zmiennymi dają czytelnikowi możliwość wyrobienia so-

bie intuicji w tworzeniu segmentów, by wiedzieć, jakimi kryteriami powinno się kierować, aby mieć gwarancję lepszych modeli.

W ostatnim rozdziale 5, załączone są informacje, które pomogą lepiej zrozumieć treści omówione we wcześniejszych rozdziałach. Przedstawione są tu wszelkie zmienne (z tabeli analitycznej) używane do budowy modeli (podrozdział 5.1). Następnie załączone są wszystkie uproszczone dokumentacje modeli (podrozdział 5.2) używanych w omawianych wcześniej strategiach akceptacji, czyli: karty skoringowe, podstawowe statystki modeli i ich funkcje kalibracji do właściwego prawdopodobieństwa modelowanego zdarzenia. Na końcu (podrozdział 5.3) umieszczono opis dobranych parametrów generatora danych. Jest tu wiele szczegółowych, omal technicznych informacji, umożliwiających pełne zrozumienie przygotowanych danych jak i samodzielne ich wykonanie. Większość z parametrów jest ustalona ekspercko i wiąże się z dużym doświadczeniem autora w pracy z danymi portfela Consumer Finance. Nie jest możliwe uzasadnienie ich wartości przez specyficzne rozumowanie, czy też jakieś proste logiczne zasady. Dobór tych parametrów jest, z jednej strony, atutem pracy, bo dzięki niemu dane są ciekawe, można na ich bazie pokazać szereg złożonych problemów, ale z drugiej – słabością, bo nie można wykazać ich zgodności z rzeczywistością. Można jedynie mieć nadzieję, że przyszłe badania pozwolą coraz lepiej dobierać parametry na podstawie konkretnych danych rzeczywistych. Niestety wiąże się to z odwiecznym problemem dostępu do danych, już wcześniej opisanym, co powoduje że, musimy zadowolić się takimi danymi, jakie udaje nam się samodzielnie stworzyć metodami symulacyjnymi.

W jakiej kolejności czytać rozdziały książki

Zebrany materiał jest przedmiotem regularnych zajęć semestralnych o nazwie *Credit Scoring i makroprogramowanie w SAS* prowadzonych dla studentów studium magisterskiego w Szkole Głównej Handlowej w Warszawie. Dodatkowo, w większym lub mniejszym zakresie, wykładany jest także w ramach podyplomowych studiów *Akademia analityka – analizy statystyczne i Data Mining w biznesie*. W związku z tym książkę można traktować jako podręcznik aka-

demicki dla studentów uczęszczających na zajęcia. Z drugiej strony materiał jest na tyle ciekawy i obszerny, że może zainteresować szerokie grono czytelników zainteresowanych badaniami naukowymi, czy zastosowaniami w biznesie. W zależności od zainteresowań czytelnika i jego głównej potrzeby książkę można czytać w różnej kolejności rozdziałów. Pierwotna kolejność jest dobrana, by zrozumieć cały materiał dogłębnie.

Pomimo dużej liczby rozdziałów, kilka z nich stanowi najważniejsze dokonania i metody proponowane do zastosowania w biznesie oraz do kontynuowania badań naukowych.

Z punktu widzenia biznesu i korzyści wynikających z zastosowań modeli skoringowych w optymalizacji procesów najważniejszy jest podrozdział 2.1, w którym w prosty sposób możemy przekonać się, że wykorzystanie skoringów przynosi miesięcznie milionowe zyski dla firmy.

Drugim bardzo ważnym tematem zastosowań modeli skoringowych jest najprostszy biznesowy model, omówiony w rozdziale 3, tania akwizycja i droga sprzedaż krzyżowa. W szczególności w podrozdziale 3.3.1, omówione są różne strategie akceptacyjne, w których używa się zarówno modeli ryzyka kredytowego, jak i marketingowych razem, aby optymalizować połączony proces akwizycji i sprzedaży krzyżowej oraz aby przynosić współmierne korzyści dla firmy. Bardzo ważne jest tu zwrócenie uwagi na kluczowe narzędzie portfeli Consumer Finance, jakim jest proces akceptacji kredytowej oraz że przy decyzji kredytowej należy brać pod uwagę nie tylko prognozowane parametry wnioskowanego kredytu, ale także przyszłego wynikającego ze sprzedaży krzyżowej. Na podstawie lektury przytoczonych rozdziałów czytelnik przekonany będzie, że nie powinno się optymalizować procesu akwizycji w oderwaniu od sprzedaży krzyżowej.

W kilku rozdziałach został poruszony kolejny istotny temat modelowania, związany z wpływem wniosków odrzuconych, który skutkuje błędnym estymowaniem ryzyka kredytowego i ogólnie zaburza wnioskowanie statystyczne związane z przyszłym zachowaniem klientów. Temat znany pod nazwą angielską Reject Inference jest najlepiej omówiony i zbadany właśnie w kontekście procesu akceptacji kredytowej, choć występuje powszechnie przy wielu innych.

W podrozdziale 3.3, problem pojawia się po raz pierwszy jako poważna trudność w dobieraniu punktów odcięcia strategii akceptacji, gdyż okazuje się, że to, co jest planowane, nie pokrywa się z tym, co jest procesowane w systemie decyzyjnym, czyli co obserwowane. Mocniej problem zarysowany jest w podrozdziale 3.3.2, gdzie pokazane są dwie różne estymacje ryzyka tego samego segmentu klientów, wynikające tylko z konsekwencji zastosowania różnych strategii akceptacyjnych. Wreszcie najmocniej problem ten poruszony jest w rozdziale 4, gdzie coraz mocniej przekonujemy się, że nie jest możliwe poprawne estymowanie ryzyka odrzuconych wniosków, jeśli ich nigdy nie zaakceptowaliśmy.

Analityków, inżynierów danych (ang. Data Scientist), którzy na co dzień budują modele predykcyjne, w szczególności karty skoringowe, najbardziej zainteresuje rozdział 4, gdzie przedstawione są prawie wszystkie powszechnie znane techniki budowy modelu, a także kilka dobrych praktyk, szczególnie przydatnych i nieopisanych dotychczas w literaturze.

Osoby zarówno ze środowiska biznesowego, jak i naukowego zainteresowane pytaniami: jakie modele predykcyjne są najlepsze? jakie techniki budowy gwarantują pożądane efekty? powinny przestudiować podrozdział 2.2, w którym zarysowana jest metoda porównywania technik budowy modeli. Temat ten bynajmniej nie jest taki prosty i jak na razie – mało rozwijany. Pomimo dość obszernej obecnie listy najprzeróżniejszych metod budowy modeli, to brakuje narzędzi do ich porównywania. Brakuje także argumentów, czy też myśli przewodniej, która mogłaby być podstawą do stwierdzeń, że np. w przypadku modeli ryzyka kredytowego dla Consumer Finance najlepszą praktyką jest stosowanie modeli typu WoE, czy LOG, opisanych w książce. Czytelnik po przestudiowaniu proponowanego rozdziału na pewno rozszerzy swoje horyzonty i będzie w stanie samodzielnie stworzyć swoje własne porównawcze narzędzie dedykowane do specyfiki jego zastosowań.

Elementem łączącym całą książkę i każdy poruszany temat są dane. Dużą wygodą i atutem sposobu prezentowania trudnych zagadnień skoringowych w książce jest właśnie oparcie się na studiach przypadków. Bez praktycznych przykładów, bez konkretnych liczb byłoby znacznie trudniej przyswoić sobie złożoność algoryt-

mów i problemów, z jakimi boryka się typowy inżynier danych, podejmując kolejne próby: wykorzystania zaawansowanej analizy danych w optymalizacji procesów biznesowych. Dlatego też na koniec proponuje się temat, który rozpoczyna książkę, właśnie dlatego, że bez danych i zrozumienia założeń konstrukcji danych, czyli możliwości ich poprawnej interpretacji i modelowania, nie da się wniknąć głęboko w poruszane tematy książki i nie zostawią one mocnego śladu, który byłby w stanie zmienić poglądy czytelnika, czy jego zachowanie. A przecież jednym z istotnych celów napisania i oddania w ręce czytelników tej książki było i jest: przekonanie szerokiego grona dzisiejszego biznesu i świata nauki, że tematy modeli predykcyjnych są nadal bardzo otwarte, że trzeba się znać na zagadnieniu, by umieć poprawnie wnioskować i wreszcie, że potrzeba ekspertów, którzy muszą zdobywać nowe doświadczenia, uczestnicząc w dużej liczbie projektów, by potem z pokorą świadomie wyznawać, że potrafią budować modele i istotnie przyczyniają się do pomnażania kapitału w przedsiębiorstwie. W tym wszystkim trzeba jeszcze pamiętać, że rewolucja Big Data, której nie unikniemy, musi być kontrolowana, by nie zgubić istoty, pokory w stosunku do danych i świadomości, że za każdym modelem, za każdym genialnym automatem prognozującym przyszłe zachowanie klienta, zawsze stoi jakiś autor, który musi nieustająco się rozwijać i który swoje efekty zawdzięcza ciężkiej i mozolnej górniczej pracy, wydobywania z danych cennej i zyskownej wiedzy biznesowej. Zatem prędzej, czy później trzeba lekturę książki rozpocząć od rozdziału 1.

1. Ogólna konstrukcja generatora losowych danych portfela Consumer Finance

1.1. Ogólny opis algorytmu

Zanim w kolejnych rozdziałach zostanie przedstawiony szczegółowy opis tworzenia generatora danych, spróbujmy go opisać w dość prosty sposób. Dane tworzone są miesiąc po miesiącu. W każdym miesięcznym etapie tworzenia danych modyfikowane są informacje o posiadanych rachunkach klientów oraz cechy samych klientów. Historia danych każdego rachunku składa się z kilku zmiennych aktualizowanych miesięcznie: liczby rat spłaconych, liczby rat opóźnionych i statusu rachunku. Każdy nowy miesiąc powinien zatem być dodawany poprzez określenie tych trzech nowych wartości zmiennych dla każdego rachunku. Na początku obliczany jest skoring, który każdemu rachunkowi przypisuje pewną wartość oceny punktowej na bazie dotychczasowej historii kredytowej i zagregowanych danych o kliencie. Dodatkowo wykorzystuje się macierz przejść pomiędzy stanami opóźnienia (liczbami opóźnionych rat). Bazując na ocenach punktowych można określić, które rachunki w następnym miesiącu spłacą kredyt, a które wpadną w większe zadłużenie. Mechanizm jest zatem związany z łańcuchem Markowa i skoringiem. Zmiany cech klienta dokonywane są także przez odpowiednie macierze przejść, które powodują, że klientowi powiększa się lub zmniejsza wynagrodzenie, powiększa się lub zmniejsza liczba dzieci itp.

1.2. Podstawowe założenia

Zastosowania Credit Scoring w procesie akceptacji kredytowej umożliwiają osiąganie istotnych korzyści finansowych. Modele bazujące na historii potrafią dobrze prognozować. Można śmiało założyć, że spłacanie kolejnego kredytu przez danego klienta jest wypadkową jego wcześniejszej historii kredytowej oraz jego aktualnej sytuacji

materialnej, zawodowej i rodzinnej, które podaje na wniosku kredytowym. Nie można jednak traktować każdego historycznego rachunku kredytowego z taką samą wagą, inaczej każdy klient w dłuższym lub krótszym horyzoncie czasowym wpadałby w opóźnienia i nie spłacał kredytów. Muszą zatem istnieć priorytety, którymi kieruje się klient przy spłacaniu kredytów. Jest powszechnie znane, że klient będzie starannie przestrzegał terminowości spłat przy kredycie hipotecznym, a niekoniecznie przy gotówkowym, czy ratalnym na zakup żelazka. Automatycznie w jego świadomości ujawniają się przykre konsekwencje utraty mieszkania znacznie boleśniejsze od straty przysłowiowego żelazka. Priorytety zatem w dużej mierze związane są z samymi procesami kredytowymi i sposobami zabezpieczeń kredytów. Jest tu także miejsce na nieracjonalne upodobania i przywiązania klienta do marki, do zaufanej pani w okienku i wielu innych subtelności, których nie da się jednoznacznie uwzględnić w modelowaniu. Istnienie priorytetów jest jednocześnie jedynym słusznym rozwiązaniem, które w przeciwnym wypadku kończyło by się hasłem: co było pierwsze: jajko, czy kura? Spłacanie kredytu A nie może zależeć od spłacania kredytu B i w tym samym czasie kredytu B od kredytu A. Wszystko od wszystkiego zależeć nie może.

Pojawia się jeszcze inny problem natury czysto algorytmicznej. Przypuśćmy, że klient miał dwa kredyty: pierwszy, a po jego spłaceniu – drugi. Przypuśćmy jednak, że przyszły testowany proces kredytowy na historii tego klienta odrzuci pierwszy z jego wniosków kredytowych, gdyż klient miał zbyt duże prawdopodobieństwo niespłacenia. Bank zatem nie posiada informacji o historii pierwszego kredytu tego klienta. Drugi wniosek kredytowy zostanie zaakceptowany. Czy jego spłacanie ma zależeć od historii pierwszego kredytu? Jeśli damy odpowiedź przeczącą, to nie potrafimy stworzyć danych symulacyjnych, gdyż nie jesteśmy w stanie przewidzieć akceptacji przyszłych testowanych procesów. Należy zatem sformułować kolejne bardzo ważne założenie: klient zawsze gdzieś kredyt weźmie. Jeśli nie uda mu się w jego ulubionym i cenionym banku, to pójdzie do innego, jeśli tam także zostanie odrzucony, to pójdzie do parabanku, a jeśli i tam mu się nie uda, to pożyczy od znajomych lub rodziny. Można tu podążać z myślą klasyków ekonomii, że klient konsumuje niezależnie od jego wynagrodzenia. Jego potrzeby konsumpcyjne,

a zatem także kredytowe, są wynikiem czegoś więcej, co związane jest z jego aspiracjami, poglądami i długofalowymi planami.

Wypiszmy zatem podstawowe założenia generatora danych, ogólnego modelu danych kredytów konsumenckich (Consumer Finance).

- Klient może otrzymać dwa rodzaje kredytów: ratalny – na zakup dóbr konsumpcyjnych i gotówkowy na dowolny cel.
- Kredyty ratalne rządzą się swoimi prawami, ich spłacanie nie jest związane z historią kredytową kredytów gotówkowych. Jest to obserwowany w bankach fakt, który najprawdopodobniej wynika z różnicy profili ogółu klientów korzystających z kredytów ratalnych, którzy czasem godzą się na kredyt ze względu na wygodę finansową, np. raty z zerowym oprocentowaniem, choć wcale ich sytuacja finansowa do tego nie zmusza. Mogliby zakupić dany towar bez wiązania się z bankiem. Kredyt gotówkowy, wybierany przez pewien podzbiór populacji kredytów ratalnych, jest czasem koniecznością i zatem jego spłacalność bardziej jest wrażliwa na sytuację finansową klienta.
- Ryzyko kredytów ratalnych jest znacząco mniejsze od gotówkowych.
- Spłacalność kredytów gotówkowych zależy od historii obu rodzajów kredytów: ratalnego i gotówkowego.
- Jeśli klient posiada wiele aktywnych kredytów, to najgorzej będzie spłacał kredyt ostatnio zaciągnięty. Od momentu wzięcia kolejnego kredytu klient staje się bardziej przeciążony zobowiązaniami i będzie mu trudniej spłacać kredyty. Z przyzwyczajenia zatem spłaca wcześniej zaciągnięte, traktując je jako bardziej priorytetowe. Można dyskutować nad słusznością tego założenia, niemniej trzeba jakoś zróżnicować spłacalność wielu kredytów. Nie jest prawdą, że klient spłaca wszystkie kredyty tak samo w tym samym czasie.

- Kredyt gotówkowy pojawia się w danym miesiącu tylko wtedy, kiedy w tym czasie klient posiadał aktywne rachunki, czyli niezamknięte. Związane jest to z procesem sprzedaży krzyżowej (ang. cross-sell), gdzie kredyt ratalny traktuje się jako akwizycję (koszt pozyskania klienta), a gotówkowy jako okazję do zarobku banku, który może organizować kampanie tylko dla swoich znanych klientów.

- Każdy kredyt posiada datę wymagalności (ang. due date) każdego 15. dnia miesiąca.

- Miesięczne zobowiązanie, czyli rata, może być albo spłacone w całości, albo wcale. Odnotowuje się tylko dwa zdarzenia: spłacenia lub niespłacenia w danym miesiącu.

- Spłacenie może jednak być związane z wpłaceniem kilku rat kredytowych.

- Identyfikowane i mierzone są tylko liczby spłaconych i niespłaconych rat.

- Wszystkie rozkłady charakterystyk klientów są wyznaczane na bazie ustalonych i precyzyjnie dobranych rozkładów losowych.

- Jeśli klient doprowadzi do sytuacji 7 niespłaconych rat (180 dni opóźnień), to rachunek kredytowy jest zamykany ze statusem B (ang. bad status), wszystkie dalsze etapy windykacyjne są pominięte.

- Jeśli klient spłaci wszystkie raty, to rachunek jest zamykany ze statusem C (ang. closed).

- Spłacenie lub niespłacenie jest zdeterminowane przez trzy czynniki: ocenę punktową liczoną na bazie wielu charakterystyk rachunku kredytowego i klienta, macierzy migracji i makroekonomicznej zmiennej modyfikującej macierz migracji.

1.3. Schemat algorytmu

Podstawowe idee algorytmu zostały opublikowane przez (Przanowski, 2013). Prezentowany w książce opis jest wersją rozszerzoną o poziom klienta, wiele kredytów dla każdego klienta i o dwa rodzaje produktów: ratalny i gotówkowy. Wszystkie nazwy zmiennych oraz niektóre raporty są prezentowane w języku angielskim, gdyż dane mogą być przydatne w środowisku naukowym międzynarodowym, używane zatem nazwy powinny być sformułowane tylko raz w jednym języku.

1.3.1. Główne parametry

Cały proces przebiega od daty startu T_s do daty końcowej T_e.

Macierz migracji M_{ij} (ang. transition matrix) jest zdefiniowana jako procent ilościowy przejścia klientów w danym miesiącu ze stanu i niespłaconych rat do j.

Na macierz wpływa zmienna makroekonomiczna $E(m)$, gdzie m jest liczbą miesięcy od T_s. Zmienna ta powinna być w granicach $0{,}01 < E(m) < 0{,}9$, ponieważ modyfikuje ona współczynniki macierzy, powodując zwiększanie się bądź zmniejszanie ryzyka portfela.

Podstawowymi strukturami danych całego procesu są tabele: dane produkcji, gdzie gromadzone są dane kolejno wnioskowanych kredytów oraz dane transakcyjne, gdzie pojawiają się miesięczne informacje o spłatach kolejnych zobowiązań.

1.3.2. Dane produkcji

Zbiór danych produkcji zawiera listę charakterystyk klienta i rachunku.

Charakterystyki klienta (dane aplikacyjne):

- Birthday – T_{Birth} – Data urodzenia klienta z rozkładem D_{Birth}.
- Income – x^a_{Income} – Wynagrodzenie klienta w czasie składania wniosku z rozkładem D_{Income}.
- Spending – $x^a_{Spending}$ – Wydatki klienta z rozkładem $D_{Spending}$.

- Job code – x^a_{job} – Kod zawodu z rozkładem D_{job}.
- Marital status – $x^a_{marital}$ – Status małżeński z rozkładem $D_{marital}$.
- Home status – x^a_{home} – Status mieszkaniowy z rozkładem D_{home}.
- City type – x^a_{city} – Wielkość miasta z rozkładem D_{city}.
- Car status – x^a_{car} – Posiadanie samochodu z rozkładem D_{car}.
- Gender – x^a_{gender} – Płeć z rozkładem D_{gender}.
- Number of children – $x^a_{children}$ – Liczba dzieci z rozkładem $D_{children}$.
- CID – Identyfikator klienta.

Parametry kredytu (dane zobowiązania):

- Installment amount – x^l_{Inst} – Kwota raty kapitałowej z rozkładem D_{Inst}.
- Number of installments – $x^l_{N_{inst}}$ – Liczba rat kapitałowych – $D_{N_{inst}}$.
- Loan amount – $x^l_{Amount} = x^l_{Inst} \cdot x^l_{N_{inst}}$ – Kwota kredytu.
- Date of application (year, month) – T_{app} – Data wniosku.
- Branch – x^l_{branch} – Kategoria towaru z rozkładem D_{branch}.
- AID – Identyfikator rachunku.

Liczba wniosków przypadająca w każdym miesiącu, określona jest rozkładem $D_{Applications}$.

1.3.3. Dane transakcyjne

Historia spłat kredytowych gromadzona jest w strukturze (dane transakcyjne):

- AID, CID – identyfikatory rachunku i klienta.

- Date of application (year, month) – T_{app} – Data wniosku.
- Current month – T_{cur} – Aktualna data, stan kredytu na tę datę.
- Due installments – $x^t_{n_{due}}$ – Liczba opóźnionych rat.
- Paid installments – $x^t_{n_{paid}}$ – Liczba spłaconych rat.
- Status – x^t_{status} – Active (A) – Status aktywny (w trakcie spłacania), Closed (C) – spłacony i Bad (B) – gdy $x^t_{n_{due}} = 7$.
- Pay days – x^t_{days} – liczba dni przed lub po dacie wymagalnej, liczba z przedziału $[-15, 15]$ jeśli jest brakiem danych, to oznacza, że nie było spłaty.

1.3.4. Wstawianie miesięcznych danych produkcji do danych transakcji

Każdy miesiąc produkcji rozpoczyna nowe historie spłat kredytów. Wszystkie nowe kredyty w danym miesiącu z danych produkcji są wstawiane do danych transakcji w następujący sposób:

$$T_{cur} = T_{app}, \quad x^t_{n_{due}} = 0, \quad x^t_{n_{paid}} = 0, \quad x^t_{status} = A, \quad x^t_{days} = 0.$$

1.3.5. Tabela analityczna ABT (Analytical Base Table)

Jak już było wspomniane, wśród podstawowych założeń generatora danych jest przekonanie, że przyszłe zachowanie klienta, jego spłacalność, jest zależna od jego dotychczasowej historii kredytowej. W celu wyznaczenia kolejnych miesięcy danych należy na dany stan, czyli miesiąc, obliczyć charakterystyki klienta agregujące informacje o aktualnym rachunku klienta, o jego poprzednich i innych rachunkach, a także ująć jeszcze ewolucję w czasie jego sald, zadłużeń i innych pochodnych wielkości. Istnieje wiele sposobów tworzenia takich charakterystyk. Układ danych z wieloma charakterystykami oraz z funkcją celu (ang. target variable) nazywa się najczęściej tabelą analityczną, (ang. Analytical Base Table – ABT). Pojęcie zostało zdefiniowane w rozwiązaniu Systemu SAS o nazwie: SAS Credit Scoring Solution[3].

[3] SAS Institute Inc. (`http://www.sas.com`).

Poniżej wypisane zostaną najbardziej typowe zmienne wchodzące w skład ABT. Pozostałe użyte w procesie generowania danych i budowy modeli opisane są w podrozdziale 5.1 (od strony 137, tabele 35–43), ich liczba to 204. Wiele zespołów analitycznych potrafi zbudować tabele analityczne gromadzące kilka tysięcy zmiennych. W takiej rzeczywistości zaczynają odgrywać istotną rolę poprawne algorytmy selekcji zmiennych do finalnych modeli.

Bazowe zmienne, na których oblicza się zmiany w czasie są zdefiniowane w następujący sposób:

$$\begin{aligned}
x^{act}_{days} &= x^{t}_{days} + 15, \\
x^{act}_{n_{paid}} &= x^{t}_{n_{paid}}, \\
x^{act}_{n_{due}} &= x^{t}_{n_{due}}, \\
x^{act}_{age} &= \mathrm{years}(T_{Birth}, T_{cur}), \\
x^{act}_{capacity} &= (x^{l}_{Inst} + x^{a}_{Spending})/x^{a}_{Income}, \\
x^{act}_{loaninc} &= x^{l}_{Amount}/x^{a}_{Income}, \\
x^{act}_{seniority} &= T_{cur} - T_{app} + 1,
\end{aligned}$$

gdzie funkcja $\mathrm{years}()$ oblicza całkowitą liczbę lat pomiędzy dwiema datami.

Rozważmy dwa szeregi czasowe zmiennych związanych z dniami płatności i zaległymi ratami:

$$\begin{aligned}
x^{act}_{days}(m) &= x^{act}_{days}(T_{cur} - m), \\
x^{act}_{n_{due}}(m) &= x^{act}_{n_{due}}(T_{cur} - m),
\end{aligned}$$

gdie $m = 0, 1, ..., 11$ oznacza względny czas określający liczbę miesięcy przed T_{cur}.

Na podstawie tak zdefiniowanych szeregów czasowych można obliczyć agregaty określające zmiany w czasie. Jeśli wszystkie elementy szeregu dla ostatnich t-miesięcy są niepuste to definiujemy:

$$\begin{aligned}
x^{beh}_{days}(t) &= (\textstyle\sum_{m=0}^{t-1} x^{act}_{days}(m))/t, \\
x^{beh}_{n_{due}}(t) &= (\textstyle\sum_{m=0}^{t-1} x^{act}_{n_{due}}(m))/t,
\end{aligned}$$

gdzie $t = 3, 6, 9$ i 12. Pokazano w tym wypadku przykładowe metody liczące średnie wartości liczby dni od daty wymagalnej i liczby

opóźnionych rat w ciągu ostatnich 3, 6, 9, czy 12 miesięcy. Tego typu zmienne oznaczane są przedrostkiem AGR3, AGR6, AGR9 i AGR12 oraz środkowym członem nazwy MEAN, czyli np. AGR3_MEAN. Można też obliczyć statystyki agregujące typu MAX i MIN. Dodatkowo można też obliczyć w przypadku, gdy tylko kilka wartości jest niepustych. Wystarczy wtedy, by minimalnie jedna wartość z szeregu 12 punktów czasowych była niepusta, a już wartość agregatu się policzy. Takie zmienne oznacza się przedrostkiem AGS (patrz tabela 37, strona 139).

1.3.6. Korekta macierzy migracji

Czynnik makroekonomiczny $E(m)$ wpływa na macierz migracji, lekko ją modyfikując, głównie współczynniki powyżej przekątnej i uzyskując dla pewnych miesięcy większe ryzyko portfela, a dla innych mniejsze:

$$M_{ij}^{adj} = \begin{cases} M_{ij}(1 - E(m)) & \text{dla} \quad j \leqslant i, \\ M_{ij} & \text{dla} \quad j > i + 1, \\ M_{ij} + \sum_{k=0}^{i} E(m) M_{ik} & \text{dla} \quad j = i + 1. \end{cases}$$

Warunek $j \leqslant i$ oznacza tu współczynniki poniżej i razem z przekątną, czyli odpowiedzialne za przejście do stanów o mniejszym zadłużeniu lub takim samym. Innymi słowy wartość $E(m)$ powiększa szansę na przejście w stany większego zadłużenia, czyli na większe ryzyko.

1.3.7. Krok iteracyjny

Krok ten wykonywany jest do wyznaczenia stanów płatności wszystkich rachunków przechodząc z miesiąca T_{cur} do $T_{cur} + 1$. Wszystkie nowe rachunki dodawane są do transakcji według schematu opisanego w podrozdziale 1.3.4. Część z pozostałych rachunków w nowym miesiącu kończy się, ze względu na spłacenie wszystkich rat lub wejście w stan siedmiu opóźnionych:

$$x_{status}^{t} = \begin{cases} C, & \text{gdy} \quad x_{n_{paid}}^{act} = x_{N_{inst}}^{l}, \\ B, & \text{gdy} \quad x_{n_{due}}^{act} = 7. \end{cases}$$

Te rachunki kredytowe nie będą już miały dalszej historii w kolejnych miesiącach.

Dla pozostałej części aktywnych rachunków, nienowych i niekończących się, zostaje wyznaczone zdarzenie płatności lub jego braku na podstawie dwóch modeli regresyjnych:

$$\begin{aligned} Score_{Main} = &\ \textstyle\sum_\alpha \beta^a_\alpha x^a_\alpha + \sum_\gamma \beta^l_\gamma x^l_\gamma + \sum_\delta \beta^{act}_\delta x^{act}_\delta \\ + &\ \textstyle\sum_\eta \sum_t \beta^{beh}_\eta(t) x^{beh}_\eta(t) + \beta_r \varepsilon + \beta_0, \qquad (1.1) \\ Score_{Cycle} = &\ \textstyle\sum_\alpha \phi^a_\alpha x^a_\alpha + \sum_\gamma \phi^l_\gamma x^l_\gamma + \sum_\delta \phi^{act}_\delta x^{act}_\delta \\ + &\ \textstyle\sum_\eta \sum_t \phi^{beh}_\eta(t) x^{beh}_\eta(t) + \phi_r \epsilon + \phi_0, \qquad (1.2) \end{aligned}$$

Gdzie $t = 3, 6, 9, 12$, α, γ, η, δ przebiegają wszystkie indeksy możliwych zmiennych z ABT (patrz podrozdziały 1.3.5 i 5.1) oraz ε i ϵ są liczbami losowymi zestandaryzowanego rozkładu normalnego N. Innymi słowy obliczone skoringi są kombinacjami wszystkich zmiennych opisujących dany rachunek na dany miesiąc, włączając informacje o zmianach w czasie do danego miesiąca oraz uwzględniając zaburzenie losowe związane ze zmiennymi ε i ϵ.

Należy dodatkowo nadmienić, że przed obliczeniem wzorów regresyjnych wygodnie jest wszystkie zmienne wejściowe zestandaryzować. W takim wypadku współczynniki można interpretować, ustawiając je w porządku malejącym, które bardziej, a które mniej ważą w stawaniu się bankrutem. Standaryzacja daje też dodatkowy efekt, mianowicie wszyscy klienci zostają zeskalowani do poziomu średniej w danym miesiącu, w danym etapie cyklu koniunkturalnego. Oznacza to, że w kryzysie najlepsi klienci też będą mieli większe problemy ze spłacalnością kredytów.

Wprowadzamy nową zmodyfikowaną macierz migracji:

$$M^{act}_{ij} = \begin{cases} M^{adj}_{ij}, & \text{gdy} \quad Score_{Cycle} \leqslant \text{Cutoff}, \\ M_{ij}, & \text{gdy} \quad Score_{Cycle} > \text{Cutoff}, \end{cases}$$

gdzie Cutoff jest dodatkowym parametrem takim, jak wszystkie współczynniki β, czy ϕ. Nowa macierz jest zatem dla segmentu poniżej punktu odcięcia modyfikowana przez zmienną makroekonomiczną $E(m)$, a powyżej – stała w czasie. Tak zdefiniowana macierz pozwala tworzyć segmenty portfela z ryzykiem bardziej i mniej zależnym od zmiennej makroekonomicznej.

Dla ustalonego miesiąca T_{cur} i ustalonego stanu aktywnych rachunków $x^{act}_{n_{due}} = i$ możemy je podzielić na grupy, bazując na ocenie $Score_{Main}$, aby każda grupa posiadała udział taki, jak odpowiednie współczynniki macierzy M^{act}_{ij}, mianowicie: grupa pierwsza $g = 0$ z najwyższymi ocenami z udziałem równym M^{act}_{i0}, druga $g = 1$ z udziałem M^{act}_{i1}, ... i wreszcie ostatnia $g = 7$ z udziałem M^{act}_{i7}.

Jeśli zatem rachunek w miesiącu T_{cur} jest w zadłużeniu i ($x^{act}_{n_{due}} = i$) oraz jest przypisany do grupy g, to dokona płatności w miesiącu $T_{cur} + 1$, gdy $g \leqslant i$, w przeciwnym przypadku nie wykona płatności i wpadnie w większe zadłużenie.

Dla braku płatności dane transakcji zostaną uzupełnione o następny miesiąc według wzorów:

$$x^t_{n_{paid}} = x^{act}_{n_{paid}},$$

$$x^t_{n_{due}} = g,$$

$$x^t_{days} = \text{Missing}.$$

Dla dokonanej płatności do danych transakcji zostaną wstawione:

$$x^t_{n_{paid}} = \min(x^{act}_{n_{paid}} + x^{act}_{n_{due}} - g + 1, x^l_{N_{inst}}),$$

$$x^t_{n_{due}} = g,$$

gdzie x^t_{days} zostanie wygenerowane z rozkładu D_{days}. Liczba wpłaconych rat $x^t_{n_{paid}}$ wyznaczana jest w prosty sposób na bazie aktualnego stanu zadłużenia, czyli $x^{act}_{n_{due}} = i$ oraz nowego g. Jeśli rachunek pozostaje w tym samym stanie, czyli $i = g$, to musi być wpłacona jedna rata, jeśli wraca do mniejszego zadłużenia, to musi być wpłaconych odpowiednio tyle rat, o ile stan zadłużenia rachunku zmniejsza się plus jedna.

Opisany krok iteracyjny wykonywany jest dla wszystkich miesięcy od $T_s + 1$ do T_e.

1.3.8. Poziom klienta, zmiany cech aplikacyjnych

W praktyce bankowej charakterystyki obliczane w tabelach analitycznych ABT są ze sobą powiązane: skorelowane i zależne. Dane

behawioralne oparte na opóźnieniach z reguły przez samą ich konstrukcję są istotnie skorelowane.

O wiele poważniejszy i trudniejszy jest problem zależności zmiennych typowo aplikacyjnych związanych z klientem takich, jak: wiek, wynagrodzenie, status małżeński, status mieszkaniowy, kod zawodu, posiadanie samochodu, liczba dzieci, wydatki, czy typ miasta. Oczywistością jest, że nie może być emeryt w wieku osiemnastu lat. Zależności te nie są jednak trywialne. Ułatwieniem może być tu istnienie słowników danych, gdzie wiele zmiennych grupuje się w dość szerokie kategorie. Kod zawodu z reguły reprezentuje kilka grup. Wynika to z potrzeby budowy modeli i analiz. Na bardzo mało licznych kategoriach wszelkie statystyki będą zbyt nieodporne i niepoprawnie estymowane, by przeprowadzać jakiekolwiek wnioskowanie.

Obecne opracowanie koncentruje się na opisaniu metody, a współczynniki modeli zostały dobrane ekspercko, bez konfrontacji z publikacjami demograficznymi, czy materiałami GUS. Byłoby bardzo pożądane, gdyby udało się zebrać stosowne opracowania mogące jednoznacznie zbliżyć się do bardziej realnych zależności pomiędzy zmiennymi i uzyskać bardziej rzeczywiste rozkłady. Jest to okazja, by zainteresować i zgromadzić środowisko naukowe, które sprostałoby zadaniu, zebrałoby dotychczasowy dorobek i stworzyłoby kompletny opis modelu. Tego typu praca byłaby bardzo cennym materiałem do przeprowadzania poprawnych scenariuszy warunków skrajnych (ang. stress testing) budowanych modeli w bankach (BIS, 2009).

Problem po części sprowadza się do stworzenia modelu gospodarstwa domowego ewoluującego w czasie. Początkiem szeregu czasowego jest w tym przypadku wiek osiemnaście lat, gdyż młodsi nie mogą być klientami banku. Potencjalny klient startuje z reguły od ustalonych grup zawodowych, jest w stanie wolnym, raczej mieszka z rodzicami, nie posiada dzieci. Tego typu założenia można dość szybko określić. W kolejnych miesiącach lub latach życia klient zmienia zawody, zmieniają mu się wynagrodzenia. Wszystkie te zmiany stanów można opisać ogólnie przez macierze przejść i oceny punktowe, podobnie jak to zostało szczegółowo opisane w podrozdziale 1.3.7, gdzie pokazano, jak klient z zadłużenia i przechodzi do zadłużenia j. Dodatkowo należy pamiętać, że wynagrodzenia po-

siadają swoje właściwe rozkłady losowe zależne także od ocen punktowych. Specjalnie nie podaje się tu szczegółowej listy zmiennych, w ogólnym przypadku wynagrodzenie jest zależne od wszystkich zmiennych z ABT. Można się jedynie domyślać, że najważniejsze są: kod zawodu, wiek, typ miasta i pewnie płeć. W tym rozdziale nie podaje się ogólnych wzorów, gdyż byłyby jeszcze bardziej złożone od przedstawionych w podrozdziale 1.3.7 a i tak nie pomogły by w zrozumieniu tematu. W tym wypadku najtrudniejsze jest określenie poprawnych parametrów modeli, współczynników macierzy przejść i współczynników we wzorach regresyjnych. Jak na razie zostały one dobrane dość swobodnie, na bazie doświadczenia autora, aby doprowadzić cały proces do końca. Na zasadzie: zobaczmy na ile wygenerowane dane będą użyteczne.

Nie wszystkie zmiany stanów klienta są widoczne lub obserwowalne przez banki. Najczęściej bank dowiaduje się o zmianie, gdy klient aplikuje o kolejny kredyt. Pomiędzy datami aplikacji nie obserwowane są zmiany klienta. Czasami przy niektórych typach produktów, jak np. hipotekach, banki wprowadzają monitoringi portfeli i proszą klientów o wypełnianie stosownych oświadczeń, gdzie uaktualnia się zmiany wynagrodzenia i kodu zawodu. Z reguły jednak tylko nowa aplikacja jest źródłem nowych danych. Jest to jeden z ciekawych i istotnych problemów banków kredytów konsumenckich, gdzie nie ma się pewności co do nowego wynagrodzenia klienta, a jednocześnie chce się wysłać ofertę kredytu z gwarancją przyszłej akceptacji przez system decyzyjny.

1.3.9. Dodatkowe algorytmy dla kredytu gotówkowego

Kredyt gotówkowy jest traktowany jako działanie cross-sell, czyli jest efektem akcji marketingowych kierowanych do istniejących klientów banku. Oznacza to, że kredyt ten może się pojawić w danych produkcji w miesiącu $T_{cur} + 1$ tylko dla klientów, którzy posiadają aktywne rachunki ($x^{act}_{status} = A$) w miesiącu T_{cur}.

Dodatkowo należy dodawać kredyty gotówkowe do produkcji według oceny punktowej do tego celu zdefiniowanej, która określa szansę na pozytywną reakcję w prowadzonej kampanii. Tylko pewien z góry ustalony procent klientów, zwany współczynnikiem

konwersji, z uzyskanymi najlepszymi punktami korzysta z kredytu gotówkowego.

1.3.10. Definicje zdarzeń: default i response

Zdarzenie niewywiązania się ze zobowiązania (ang. default) jest kluczowym pojęciem modeli Credit Scoring oraz rekomendacji Basel II i III.

Każdy rachunek jest badany od miesiąca jego aplikowania i uruchomienia T_{app}, zwanego punktem obserwacji, w ciągu kolejnych 3, 6, 9 i 12 miesięcy. W tym okresie, zwanym okresem obserwacji, oblicza się statystykę, maksymalne opóźnienie:

$$\text{MAX} = \text{MAX}_{m=0}^{o-1}(x_{n_{due}}^{act}(T_{app} + m)),$$

gdzie $o = 3, 6, 9, 12$. Innymi słowy pytamy o maksymalną liczbę opóźnionych rat w ciągu pierwszych 3, 6, 9 i 12 miesięcy od uruchomienia kredytu. Na podstawie tej statystyki definiujemy trzy typowe kategorie.

Dobry (Good): Gdy MAX $\leqslant 1$ lub dodatkowo w czasie okresu obserwacji rachunek spłacił całe swoje zobowiązanie: $x_{status}^{t} = C$.

Zły (Bad): Gdy MAX > 3 lub w czasie okresu obserwacji rachunek wpadł w maksymalne zadłużenie: $x_{status}^{t} = B$. Wyjątkowo dla $o = 3$, gdy MAX > 2.

Nieokreślony (Indeterminate): dla pozostałych przypadków.

Istnienie stanu nieokreślonego może podlegać dyskusji; jest to kwestia pewnej konwencji. Istotą jest próba koncentracji na bardzo złych przypadkach, a nie tylko trochę złych. Jeśli klient ma małe opóźnienie, to może jeszcze z niego wyjść, ale przy dużym staje się istotnym problemem banku.

Podstawową statystyką mierzącą ryzyko jest udział procentowy (ilościowy) kategorii złych (ang. bad rate, default rate). Istnienie kategorii nieokreślonego powoduje zmniejszenie ryzyka.

W przypadku modeli do badania prawdopodobieństwa pozytywnej reakcji na kampanie wprowadza się zdarzenie określane w języku angielskim cross-sell lub response. Punktem obserwacji jest tu każdy kolejny miesiąc całego aktywnego portfela, czyli miesiąc T_{cur}.

Definiujemy dwie kategorie.

Dobry (Good): Gdy w czasie okresu obserwacji klient wziął kredyt gotówkowy.

Zły (Bad): Gdy w czasie okresu obserwacji klient nie wziął żadnego kredytu gotówkowego.

Statystykę udziału kategorii dobrego nazywa się współczynnikiem konwersji (ang. response rate). W całej pracy przyjmuje się założenie, że zdarzenie response bada się albo w okresie obserwacji jednego miesiąca albo sześciu.

2. Model uproszczony, kredyty ratalne

Generator danych losowych Consumer Finance w uproszczonej wersji został obszernie opisany w (Przanowski, 2013). Jest to model tylko jednego produktu: kredytu ratalnego. Każdy klient posiada tylko jeden kredyt. Zmienne są zatem budowane tylko na podstawie aktualnej historii kredytowej jednego kredytu. Wszystkie pozostałe szczegóły odpowiadają ogólnej konstrukcji opisanej w rozdziale 1.

Zbiór zawiera 2 694 377 wierszy (obserwacji) i 56 kolumn (zmiennych). Każdy wiersz reprezentuje wniosek kredytowy, a kolumny to wszelkie zmienne opisujące ten wniosek łącznie ze zmiennymi ABT.

2.1. Opłacalność procesu, wpływ mocy predykcyjnej na zysk

Rozdział ten jest rozszerzoną wersją publikacji (Przanowski, 2014).

Modele Credit Scoring są powszechnie stosowane w optymalizacji procesów bankowych. Nikt już tego dziś nie kwestionuje, ale mało jest opracowań wykazujących ich przydatność, konkretne kwoty zysku, czy oszczędności. Być może jest to spowodowane chęcią utrzymania tajemnicy przedsiębiorstwa, gdyż tym samym ujawniałoby się w prosty sposób dość krytyczne dla funkcjonowania banku informacje. Wygodne jest zatem wykorzystanie danych losowych, gdyż w tym wypadku nie chroni nas tajemnica. Z drugiej strony nie jest istotne wykazanie przydatności co do złotego, gdyż bardzo wiele składowych kosztów jest związanych ze specyfiką funkcjonowania danej firmy i nie da się ich uogólnić.

Dane uproszczone kredytu ratalnego zostały specjalnie przerobione, aby podnieść wartość globalnego ryzyka aż do 47% oraz zostały zbudowane karty skoringowe z różną mocą predykcyjną, wyrażoną w statystyce Giniego (Siddiqi, 2005).

Niestety bez posiadania szczegółowych kosztów przedsiębiorstwa nie da się przedstawić całego arkusza zysków i strat (ang. P&L). Ale wystarczającą informacją jest policzenie straty oczekiwanej, prowizji i przychodów odsetkowych. Wszystkie inne koszty będą tylko

odejmować się od zysku, nie wpłyną one zatem na wartości przyrostów.

Wprowadźmy oznaczenia: APR – roczne oprocentowanie kredytu, $r = \frac{\mathrm{APR}}{12}$ (można też oprocentowanie to traktować jako marżę dla banku, czyli oprocentowanie dla klienta pomniejszone o koszt kapitału, który bank musi ponieść, udzielając kredytu), p – prowizja za udzielenie kredytu, płatna przy uruchomieniu kredytu, $x^l_{Amount}(i) = A_i$ – kwota kredytu i $x^l_{N_{inst}}(i) = N_i$ – liczba rat, gdzie i jest numerem kredytu. Zgodnie z obecnymi regulacjami Basel II stratę oczekiwaną można wyrazić jako sumę iloczynów trzech członów: prawdopodobieństwa zdarzenia, default, niewywiązania się z zobowiązania (PD), procentu straty zobowiązania dla zdarzenia default (LGD) i kwoty zobowiązania w czasie zdarzenia default (EAD). Bez większych obliczeń można przyjąć, w ujęciu ostrożnościowym, że: $\mathrm{LGD} = 50\%$ a EAD jest kwotą kredytu. Dla historycznych danych nie musimy bazować na prognozie, ale możemy przyjąć, że PD jest oparte na wartości zmiennej $\mathrm{default}_{12}$. Jak nastąpiło zdarzenie default, to $\mathrm{PD} = 100\%$ a jak nie, to $\mathrm{PD} = 0\%$, wtedy wyznacza się obserwowaną stratę L. Kwotę I przychodów odsetkowych łącznie z prowizją oblicza się na podstawie procentu składanego. Mamy zatem dla każdego i-tego kredytu:

$$L_i = \begin{cases} 50\% A_i, & \text{gdy} \quad \mathrm{default}_{12} = \textbf{Zły}, \\ 0, & \text{gdy} \quad \mathrm{default}_{12} \neq \textbf{Zły}. \end{cases}$$

$$I_i = \begin{cases} A_i p, & \text{gdy} \quad \mathrm{default}_{12} = \textbf{Zły}, \\ A_i(N_i r \frac{(1+r)^{N_i}}{(1+r)^{N_i}-1} + (p-1)), & \text{gdy} \quad \mathrm{default}_{12} \neq \textbf{Zły}. \end{cases}$$

Czyli sumaryczny zysk (profit) P całego portfela obliczamy następującym wzorem:

$$P = \sum_i I_i - L_i. \tag{2.1}$$

Dla każdego modelu skoringowego z różnymi mocami predykcyjnymi możemy posortować wszystkie wnioski po wartości oceny punktowej, od najmniej do najbardziej ryzykownego. Ustalając

punkt ocięcia, wyznaczamy sumaryczną wartość zysku na zaakceptowanej części portfela i procent akceptacji. Otrzymujemy w ten sposób krzywe Profit prezentowane na rysunku 1. Niektóre z nich, np. dla Giniego z wartością 20%, nigdy nie przyniosą zysku dla banku, niezależnie od procentu akceptacji zawsze tracimy. Przy tak niskiej mocy predykcyjnej procesu akceptacji (jego modelu skoringowego lub wszystkich reguł decyzyjnych) nie daje się prowadzić biznesu. Co więcej, przy akceptacji wszystkich wniosków całkowity zysk jest ujemny i wynosi około $-44{,}5$ miliona PLN (mPLN). Najlepsze trzy krzywe pokazano dokładniej na rysunku 2. Modele z mocą większą od około 60% potrafią zidentyfikować opłacalne segmenty, przy czym im lepszy model, tym więcej możemy zaakceptować i więcej zarobić. Dla modelu o mocy 89%, dość dużej, by pojawiła się w praktyce, można zaakceptować prawie 44% wniosków i zyskać $10{,}5$ mPLN. Dla tego modelu na rysunku 3 pokazano dodatkowo składowe zysku, czyli przychód i stratę narastająco. Przy pełnej akceptacji strata sięga aż $72{,}2$ mPLN. Krzywa straty narasta wykładniczo przy wzroście procentu akceptacji, natomiast przychody rosną prawie liniowo. Brak idealnej linii jest efektem różnych kwot kredytów. Co więcej, przychody rosną bardzo podobnie dla każdego modelu, niezależnie od jego mocy predykcyjnej. Zupełnie inaczej ma się sprawa z krzywymi strat, patrz rysunek 4. Im większa moc modelu tym krzywa straty jest bardziej zakrzywiona, tym dłuższy odcinek od zerowej akceptacji jest spłaszczony, czyli niepowodujący dużej straty. Stopień zakrzywienia krzywej straty jest bardzo prostą interpretacją statystyki Giniego. Dla modelu o zerowej mocy, strata będzie narastać liniowo. Dla stu procent będzie to łamana: do wartości dopełnienia ryzyka globalnego $1 - 47\% = 53\%$ będzie linią zerową a potem gwałtownie liniowo będzie rosnąć do całkowitej straty $72{,}2$ mPLN.

Warto sobie zdawać sprawę z przedstawionych kwot, gdyż one właśnie w prosty sposób udowadniają, jak ważną funkcję pełnią modele skoringowe w pomnażaniu kapitału przedsiębiorstwa.

Można także obliczyć proste wskaźniki poprawy zysku, straty i procentu akceptacji, przy założeniu zwiększenia mocy predykcyjnej modelu o 5%. W tabeli 1 przedstawiono zebrane wskaźniki. Wystarczy zwiększyć moc modelu o 5%, a miesięcznie bank zarobi

o prawie $1{,}5$ miliona PLN więcej, zwiększając przy tym procent akceptacji o $3{,}5\%$. Można też, nie zmieniając procentu akceptacji (ang. acceptance rate, AR) na poziomie 20%, polepszać model i oszczędzać na stracie. W tym wypadku prawie 900 tysięcy PLN (kPLN) zaoszczędzimy miesięcznie. Dla akceptacji 40% oszczędność będzie aż $1{,}5$ miliona PLN miesięcznie.

Zaprezentowane kwoty zysku, czy oszczędności uzasadniają istnienie zespołów analitycznych w bankach oraz zapraszają wszelkich analityków do ciągłego rozwoju i doskonalenia zawodowego. Pobudzają, by nieustająco testować i sprawdzać, czy nie da się zbudować lepszych modeli.

Niestety wszelkie prezentowane liczby są poprawne przy założeniu sensownej estymacji straty, czy też prawdopodobieństwa zdarzenia default. Jeśli, wprowadzając nowy model, zwiększamy procent akceptacji, to możemy liczyć się z niedoszacowaną stratą wynikającą z wpływu wniosków odrzuconych. Dokładniejszą analizę ich wpływu omawia się w podrozdziale 3.3.2.

2.2. Porównywanie technik budowy modeli

Rozdział ten jest modyfikacją i rozszerzoną wersją publikacji (Przanowski i Mamczarz, 2012).

Jak już zostało wcześniej wspomniane, obecna literatura dotycząca Credit Scoring przedstawia różne techniki budowy modeli predykcyjnych na stosunkowo dość mało licznych zbiorach. Dotyczy to zarówno liczby obserwacji, jak i zmiennych.

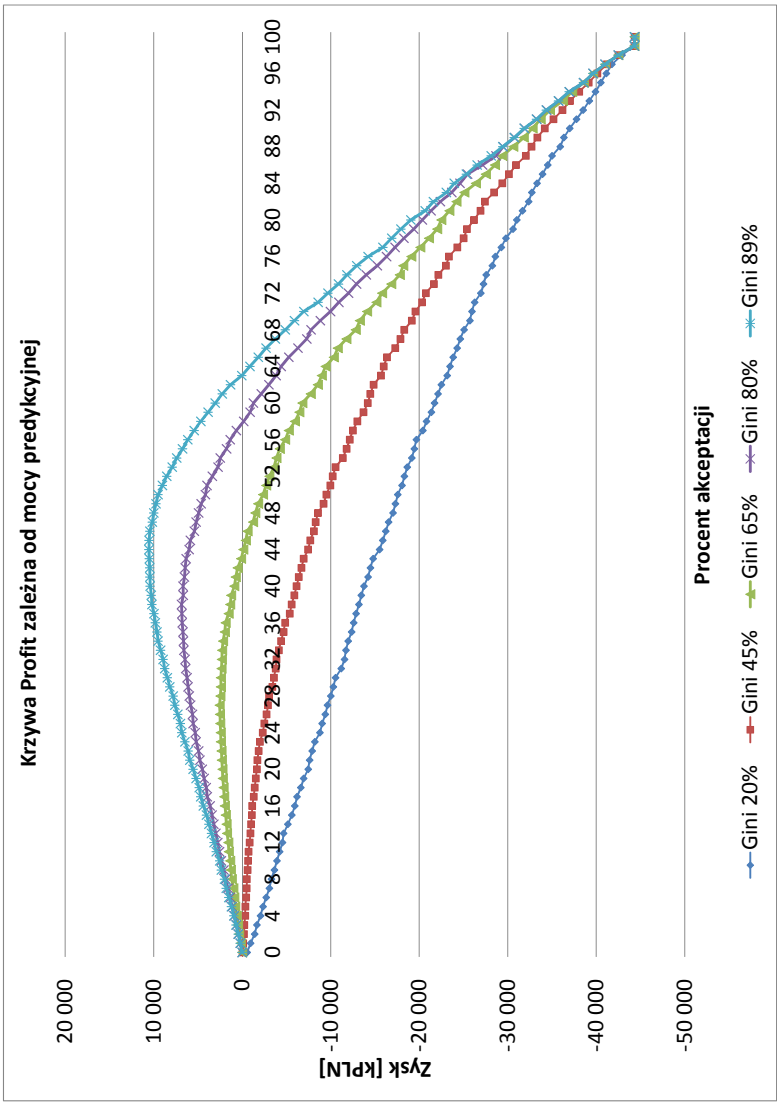
Obecne czasy tworzą coraz to większe wyzwania co do wielkości danych. Temat Big Data na stałe zagościł już na konferencjach i musimy umieć się przygotowywać do zupełnie nowego paradygmatu. Pierwszą poważną zmianą będzie pogodzenie się z automatycznymi procesami budowy modeli. Nie chodzi tu o budowanie modeli w całkowicie automatyczny sposób, przez przysłowiowe kliknięcie myszką, ale o umiejętne wykorzystanie automatów, by ułatwiały analitykowi te etapy, które powinny i dały możliwość koncentracji na tych, których nie da się zautomatyzować. Jeśli zaczyna się od kilku tysięcy zmiennych, nie można oglądać wykresów i raportów dla każdej z nich. Muszą powstać wygodne narzędzia wspomagające decyzje

Tabela 1: Przyrosty wskaźników finansowych zależne od zmiany mocy predykcyjnej modelu

Wskaźnik	Wartość
Liczba wniosków w miesiącu	50 000
Średnia kwota kredytu	5 000 PLN
Średni czas kredytowania	36 miesięcy
Roczne oprocentowanie kredytów	12%
Prowizja za udzielenie kredytu	6%
Globalne ryzyko portfela	47%
Zmiana mocy predykcyjnej	5%
Zmiana procentu akceptacji	3,5%
Zmiana zysku	1 492 kPLN
Zmiana straty oczekiwanej ($\mathrm{AR} = 20\%$)	872 kPLN
Zmiana straty oczekiwanej ($\mathrm{AR} = 40\%$)	1 529 kPLN

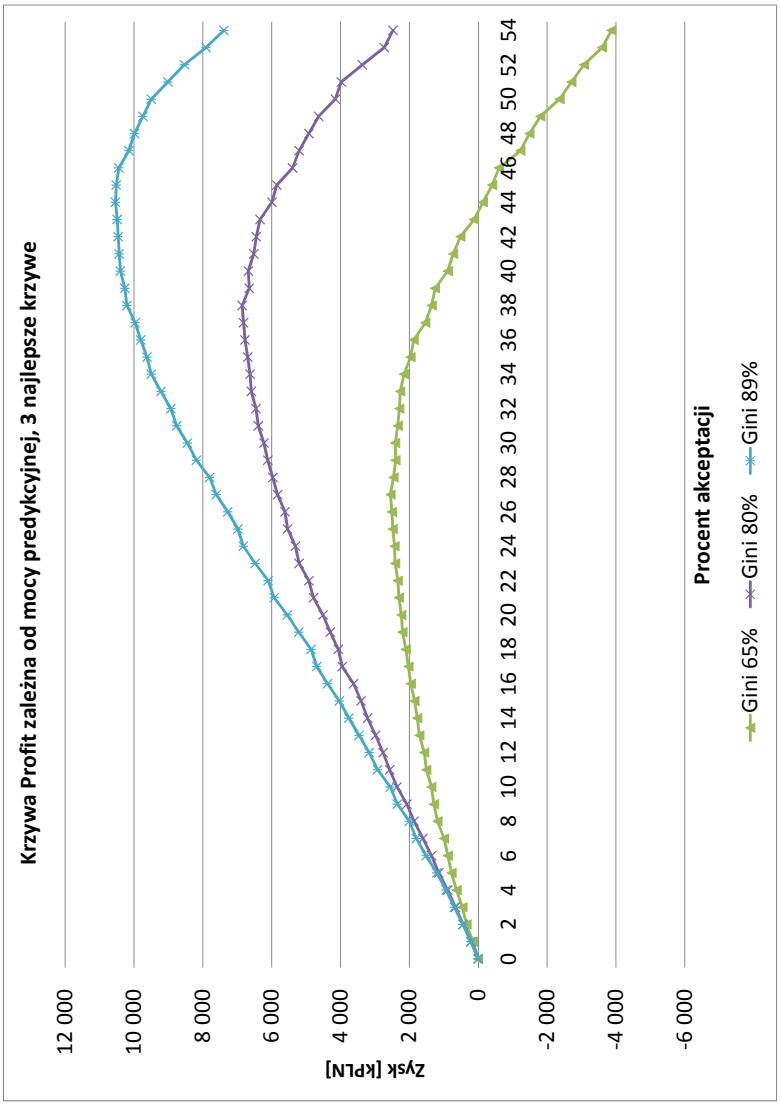
Źródło: (Przanowski, 2014).

Rysunek 1: Krzywe Profit



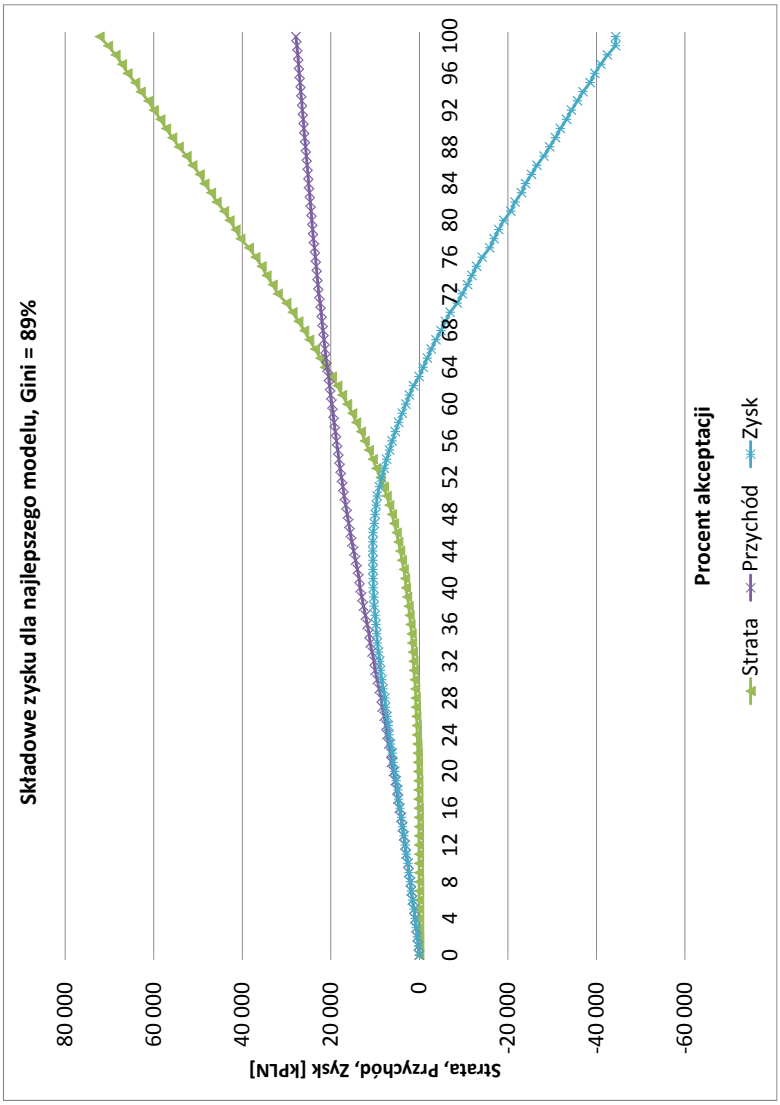
Źródło: (Przanowski, 2014).

Rysunek 2: Najlepsze krzywe Profit



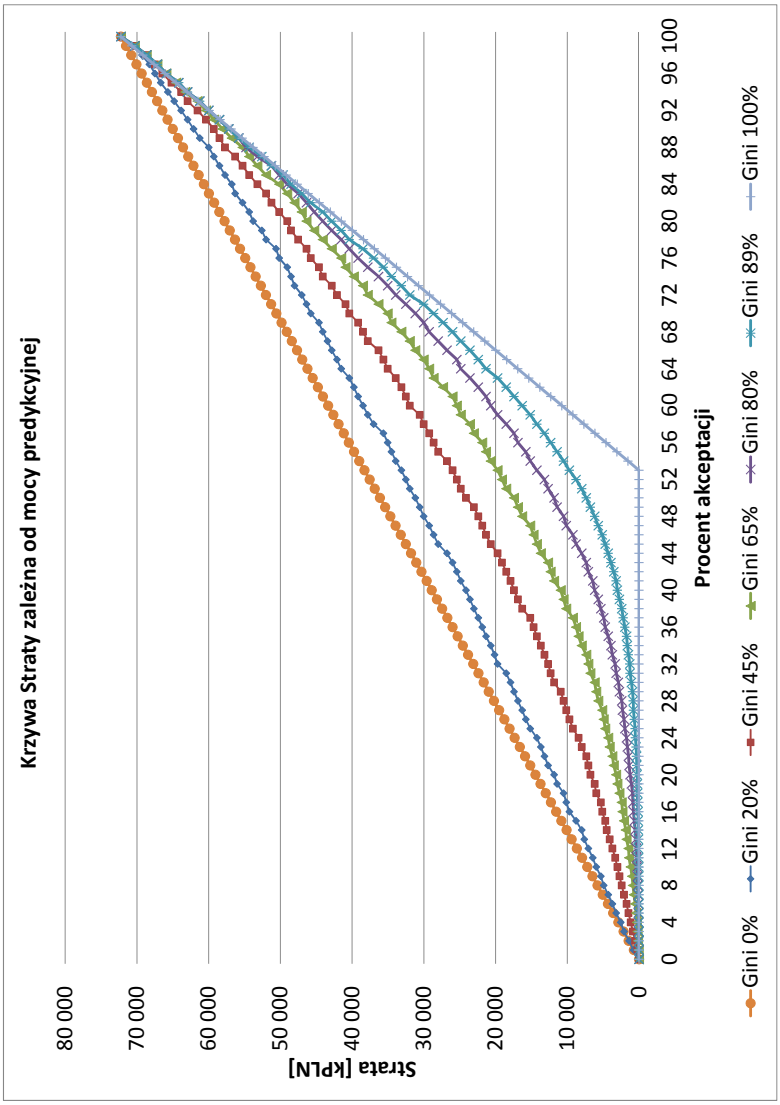
Źródło: (Przanowski, 2014).

Rysunek 3: Składowe zysku dla najlepszego modelu



Źródło: opracowanie własne.

Rysunek 4: Krzywe Straty



Źródło: opracowanie własne.

i doprowadzające do mniejszej liczby kandydatów, by później bardziej indywidualnie analizować każdą z ich. Muszą zatem pojawić się nowe statystyki, mierniki, czy też kryteria, którymi owe automaty będą się kierowały. Druga poważna zmiana w myśleniu polega na przejściu z czci i zaufania do wiedzy eksperckiej na korzyść wiedzy czerpanej z danych. Nie chodzi tu o negowanie poprzednich doświadczeń, czy wiedzy zdobywanej latami, ale o świadomość, że coraz to szersze zakresy dostępnych danych spowodują brak możliwości posiadania doświadczenia, gdyż w skrajnych przypadkach będzie się zawsze budowało modele na bazie nowych danych, których jeszcze nikt wcześniej nie widział. Dodatkowo musimy zauważyć, że rynek zmienia się coraz bardziej dynamicznie i tylko aktualne dane pozwolą rozumieć pojawiające się tendencje.

2.2.1. Dane wykorzystane do analiz

W celu przeprowadzenia dowodu uniwersalności danych losowych do badań zostały wykorzystane dwa zestawy danych rzeczywistych pochodzące z odległych dziedzin: bankowej i medycznej.

Rzeczywiste dane bankowe. Pochodzą z jednego z Polskich banków sektora Consumer Finance. Zawierają 50 000 wierszy i 134 kolumny. Nazwy kolumn są utajnione, a zmienna celu reprezentuje informację o klientach, którzy weszli w opóźnienia więcej niż 60 dni w ciągu 6 miesięcy od daty obserwacji.

Rzeczywiste dane medyczne. Opisują przypadki zachorowania na raka w Stanach Zjednoczonych (Delen *et al.*, 2005). Zawierają 1 343 646 wierszy i 40 kolumn. Zmienna celu opisuje przeżycie lub zgon pacjenta z powodu nowotworu danego typu w ciągu pięciu lat od diagnozy.

2.2.2. Ogólny proces budowy modelu karty skoringowej

Dla każdego zestawu danych zostało zbudowanych kilkaset modeli predykcyjnych. Wszystkie obliczenia dokonano w Systemie SAS, wykorzystując moduły Base SAS, SAS/STAT i SAS/GRAPH.

- Próby losowe. Stworzone zastały dwa zbiory: treningowy i walidacyjny przesunięte w czasie; zbiór walidacyjny zawierał obserwacje z późniejszego okresu niż zbiór treningowy. Taka metoda zwana z ang. time sampling pozwala badać stabilność modeli w czasie.

- Tworzenie kategorii – kategoryzacja zmiennych lub grupowanie (ang. binning). Na podstawie miary entropii każda zmienna ciągła jest kategoryzowana do zmiennej porządkowej. Jest to typowa metoda stosowana przy drzewach decyzyjnych. Zmienne nominalne lub porządkowe także się grupuje, tworząc finalnie liczniejsze kategorie, bardziej reprezentatywne w populacji.

- Wstępna selekcja zmiennych (preselekcja) – odrzucenie zmiennych o niskiej predykcyjności. W tym etapie na podstawie prostych jednowymiarowych kryteriów odrzuca się zmienne nienadające się do finalnego modelu, które mają albo znacząco słabą wartość predykcji, albo są bardzo niestabilne w czasie.

- Wielowymiarowa selekcja zmiennych – generator modeli. Dla zmiennych ciągłych w procedurze Logistic istnieje metoda selekcji oparta na heurystyce branch and bound (Furnival i Wilson, 1974). Jest to bardzo wygodna metoda, gdyż pozwala stworzyć wiele modeli, w tym wypadku 700, po 100 najlepszych z modeli o 6 zmiennych, 7 zmiennych, ..., 12 zmiennych.

- Ocena modeli. Nie ma jednego kryterium do oceny modelu. Stosuje się zatem kilka kryteriów głównie związanych z mocą predykcyjną (AR inaczej Gini), stabilnością $\mathrm{AR}_{\mathrm{diff}}$ (inaczej delta Gini – różnica względna predykcji pomiędzy zbiorami: treningowym i walidacyjnym), miary współliniowości: $\mathrm{MAX}_{\mathrm{VIF}}$ – maksymalny współczynnik inflacji wariancji[4] (ang. Variance Inflation Factor) (Koronacki i Mielniczuk, 2001), $\mathrm{MAX}_{\mathrm{Pearson}}$ – maksymalny współczynnik korelacji Pearsona na parach zmiennych i $\mathrm{MAX}_{\mathrm{ConIndex}}$ – maksymalny indeks

[4] Czasem tłumaczony jako współczynnik podbicia wariancji.

warunkujący (ang. Condition Index) (Welfe, 2003) oraz miara istotności: $\mathrm{MAX}_{\mathrm{ProbChiSquare}}$ – maksymalna p-value dla zmiennych. W pracy (Lessmanna *et al.*, 2013) wymienia się jeszcze kilka innych. Jest to temat nadal bardzo aktualny, wielu autorów kwestionuje opieranie się na statystyce Giniego (Scallan, 2011).

- Wdrożenie modelu. Jak zostanie pokazane, etap ten jest także bardzo ważny i pomimo dobrych wskaźników budowanego modelu jego zastosowanie może przynieść inne wyniki niż się spodziewamy (podrozdział 3.3.1). Umiejętne wdrożenie modelu to wyzwanie dla badań naukowych.

- Monitoring i testowanie. Ten temat nie będzie poruszany bezpośrednio w obecnej pracy, ale w praktyce zawsze jest nieodzownym etapem cyklu życia modelu. Trzeba wiedzieć, jak testować poprawność działania modelu, jak upewnić się, czy pojawiła się już sposobność zbudowania lepszego i jak oba modele porównać. Wreszcie trzeba wiedzieć, jakimi kryteriami kierować się przy decyzji o wymianie na nowy.

2.2.3. Różne kodowania i selekcje zmiennych

Model skoringowy, choć oparty na tym samym zestawie zmiennych, może być estymowany w regresji logistycznej na różne sposoby, zależnie od metody kodowania zmiennych.

Pierwszy sposób, oznaczany jako REG, to model bez jakiejkolwiek transformacji zmiennych. W tym wypadku potrzebna jest metoda uzupełniania braków danych. Wybrano najprostszą: uzupełnianie przez średnią. Należy pamiętać, że sama metoda uzupełniania braków jest już tematem samym w sobie i może znacząco zmienić wyniki przedstawione w pracy.

Drugi sposób – LOG oparty jest na transformacji logit: każdej kategorii zmiennej przypisana jest jej wartość logit. Jest to prawie identyczna metoda do WoE stosowanej w SAS Credit Scoring Solution (Siddiqi, 2005), patrz także podrozdział 4.2.4.

Trzeci sposób – GRP związany jest z kodowaniem binarnym referencyjnym zwanym z ang. reference lub dummy; patrz tabela 2.

W tym wypadku poziomem referencyjnym jest grupa o najwyższym numerze, czyli o najmniejszym ryzyku. W ogólnej teorii wybór poziomu referencyjnego nie jest tematem trywialnym. Najczęściej przyjmuje się dominującą kategorię (Frątczak, 2012), aczkolwiek sposób kodowania może znacząco wpłynąć na współliniowość w modelu nawet przy najliczniejszych kategoriach referencyjnych. Na kodowanych zmiennych nie można przeprowadzić selekcji metodą branch and bound, gdyż liczba zmiennych rozrasta się do tak dużej, że czas obliczeń staje się prawie nieskończenie długi. Zmienne sztuczne lub inaczej zerojedynkowe (ang. dummy) znacząco zmieniają postać modelu. Każda zmienna z n - kategoriami będzie miała $(n-1)$ zmiennych sztucznych. Dodatkowo firma Score Plus (Scallan, 2011) słusznie proponuje selekcję tych zmiennych wykonywać na innym kodowaniu kumulatywnym, zwanym z ang. ordinal lub nested; patrz tabele 3, 4 i 5. Dla ostatniej metody wszystkie bety (współczynniki z modelu) z jedną zmienną mają ten sam znak, co pozwala dodatkowo badać współliniowość. Jeśli znak bety się odwraca, znaczy, że zmienne zbyt mocno zależą od siebie.

W przypadku GRP nie jest możliwa selekcja zmiennych taka, jak dla REG i LOG. Zostały zatem zebrane wszystkie modele wyprodukowane dla REG i LOG, a potem ocenione metodą GRP. Ze względu na brak możliwości zdefiniowania indywidualnej metody selekcji dla GRP stworzono kilka algorytmów korekcji liczby zmiennych w GRP. Wiele kategorii w zetknięciu się z innymi w modelu staje się nieistotnymi, potrzeba zatem metody ich detekcji. Wybrano dwie: eliminacji wstecznej –ang. backward i krokowej – ang. stepwise. Finalny model można estymować zarówno przy kodowaniu dummy jak i nested, co powoduje, że otrzymuje się 12 metod korekcji modeli GRP; patrz tabela 6.

Wszystkie modele z wyjątkiem REG są modelami karty skoringowej, patrz tabela 7. Modele REG zostały tu rozważone jedynie jako dodatkowa skala porównawcza.

Tabela 2: Kodowanie referencyjne – ang. dummy

Numer grupy	Zmienna 1	Zmienna 2	Zmienna 3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Źródło: opracowanie własne.

Tabela 3: Kodowanie kumulatywne malejące – ang. decending nested

Numer grupy	Zmienna 1	Zmienna 2	Zmienna 3
1	0	0	0
2	1	0	0
3	1	1	0
4	1	1	1

Źródło: opracowanie własne.

Tabela 4: Kodowanie kumulatywne rosnące – ang. ascending nested

Numer grupy	Zmienna 1	Zmienna 2	Zmienna 3
1	1	1	1
2	0	1	1
3	0	0	1
4	0	0	0

Źródło: opracowanie własne.

Tabela 5: Kodowanie kumulatywne monotoniczne – ang. monotonic nested

Numer grupy	Zmienna 1	Zmienna 2	Zmienna 3
1	1	1	1
2	1	1	0
3	1	0	0
4	0	0	0

Źródło: opracowanie własne.

Tabela 6: Metody korekcji modeli GRP

Nawa metody	Estymacja	Selekcja	Kodowanie
NBA	nested	backward	ascending nested
NBD	nested	backward	descending nested
NBM	nested	backward	monotonic nested
NSA	nested	stepwise	ascending nested
NSD	nested	stepwise	descending nested
NSM	nested	stepwise	monotonic nested
DBA	dummy	backward	ascending nested
DBD	dummy	backward	descending nested
DBM	dummy	backward	monotonic nested
DSA	dummy	stepwise	ascending nested
DSD	dummy	stepwise	descending nested
DSM	dummy	stepwise	monotonic nested

Źródło: opracowanie własne.

Tabela 7: Przykładowa karta skoringowa

Zmienna	Warunek (kategoria)	Ocena cząstkowa
Wiek	$\leqslant 20$	10
	$\leqslant 35$	20
	$\leqslant 60$	40
Wynagrodzenie	$\leqslant 1500$	15
	$\leqslant 3500$	26
	$\leqslant 6000$	49

Źródło: opracowanie własne.

2.2.4. Etapy obliczeń, zebranie wyników

Dla każdego typu danych wylosowano próby do zbiorów: treningowego i walidacyjnego oraz przeprowadzono wstępną selekcję zmiennych; patrz tabela 8.

Następnie policzono 700 modeli dla REG i LOG oddzielnie. Potem 1400 modeli dla GRP, które w kolejnym kroku korygowano 12 metodami. W sumie zbudowano i oceniono około 19 600 modeli dla każdego typu danych oddzielnie (czyli w sumie 58 800 modeli). Istnienie tak dużej liczby ocenionych modeli daje możliwość badania rozkładów kryteriów oceny, a tym samym pozwala porównywać metody budowy modeli.

Można inaczej sformułować istotę metody porównawczej proponowanej w książce. Chodzi o znalezienie schematu generowania dużej liczby modeli danego typu lub danej techniki, czyli o stworzenie chmur modeli – każda chmura dla innego typu. Wtedy, dysponując różnego rodzaju kryteriami ocen modeli można precyzyjnie porównywać i badać rozkłady tych ocen.

Wszystkie obliczenia wykonano na Laptopie Core Duo 1,67GHz. Obliczenia trwały 2 miesiące.

Tabela 8: Liczebności w próbach oraz liczby zmiennych po wstępnej selekcji

Typ danych	Treningowy	Walidacyjny	Liczba zmiennych
Bankowe	27 325	12 435	60
Medyczne	29 893	17 056	23
Losowe	66 998	38 199	33

Źródło: opracowanie własne.

2.2.5. Interpretacja zebranych wyników

W celu uzyskania tej samej skali dla wszystkich metod, technik modelowych, z każdego zestawu modeli wybrano 700 najlepszych modeli względem kryterium $\mathrm{AR}_{\mathrm{Valid}}$, czyli mocy predykcyjnej Giniego na zbiorze walidacyjnym. Na rysunkach 5, 6 i 7 przedstawiono jednowymiarowe rozkłady wartości podstawowych kryteriów oceny modeli. Wyraźnie daje się zauważyć, że zróżnicowanie predykcyjności ma miejsce dla modeli REG, LOG i GRP. Wszystkie modele GRP i różne ich korekcje mają podobne wartości. Podobnie ma się rzecz z kryterium stabilności $\mathrm{AR}_{\mathrm{diff}}$. W przypadku współliniowości występują dość duże rozbieżności. Wyraźnie widać, że korekcje GRP zdecydowanie polepszają $\mathrm{MAX}_{\mathrm{VIF}}$, a dla modeli LOG wartości skupiają się w obrębie akceptowalnych; literatura podaje granice pomiędzy 2 i 10 dla VIF (Belsley, 1991) w przypadku regresji liniowej, a dla regresji logistycznej najlepiej przyjąć górną granicę 3.

Znacznie lepszą metodą jest porównywanie modeli na podstawie wielowymiarowego kryterium odległości od ideału. W tym wypadku wszystkie kryteria jednocześnie biorą udział w ocenianiu modelu. W następnym kroku wprowadza się wagi, aby dane kryterium uczynić najmocniejszym; patrz rysunki 8, 9 i 10. Im niżej są punkty, tym lepsze modele, tym bliższe ideałowi. Tak wykonane rysunki pozwalają lepiej porównać modele. Widać wyraźnie, że modele REG odbiegają od ideału dla każdego zestawu danych. Modele GRP mają duży rozrzut i też nie powinny być nazywane „dobrymi". Modele LOG mają pożądane rozkłady, choć nie zawsze są bardzo nisko. Nie-

które korekcje GRP wykazują najlepsze własności, szczególnie modele estymowane przy technice kodowania kumulatywnej (nested).

2.2.6. Finalne porównanie technik LOG i NBM

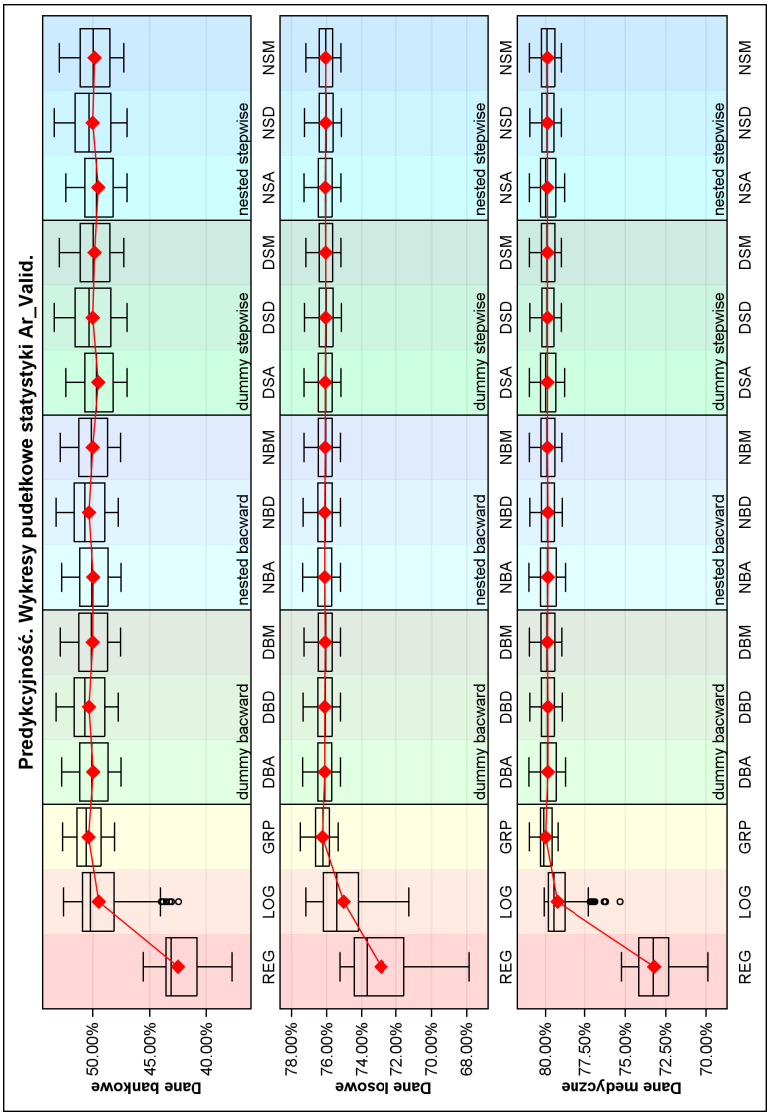
Na podstawie wielu wykresów trójwymiarowych, które trudno jest prezentować w pracy, zostały wybrane dwa najważniejsze kryteria, by najlepiej uwydatnić różnice pomiędzy dwiema technikami modelowymi.

Moc predykcyjna $\mathrm{AR}_{\mathrm{valid}}$ oraz stabilność $\mathrm{AR}_{\mathrm{diff}}$ okazały się wystarczające do wykazania różnic. Na rysunkach 11, 12 i 13 przedstawiono dwuwymiarowe wykresy rozproszenia. Modele LOG, reprezentowane jako gwiazdki na wykresach, wykazują lepsze własności stabilności niż NBM, ale mają trochę mniejsze moce predykcyjne. Pomimo mniejszych wartości predykcyjności dla LOG lepiej jest budować modele stabilniejsze i prostsze. Jest to podejście bardziej konserwatywne, łatwiejsze w budowie i implementacji. Modele mniej stabilne w czasie, choć z lekko większą mocą predykcyjną, trzeba będzie częściej monitorować (ze względu na większe prawdopodobieństwo przetrenowania). Jeśli zatem nie mamy za dużo czasu, a to zdarza się częściej niż się spodziewamy, wtedy lepiej używać jednej sprawdzonej metody: LOG.

2.2.7. Podsumowanie

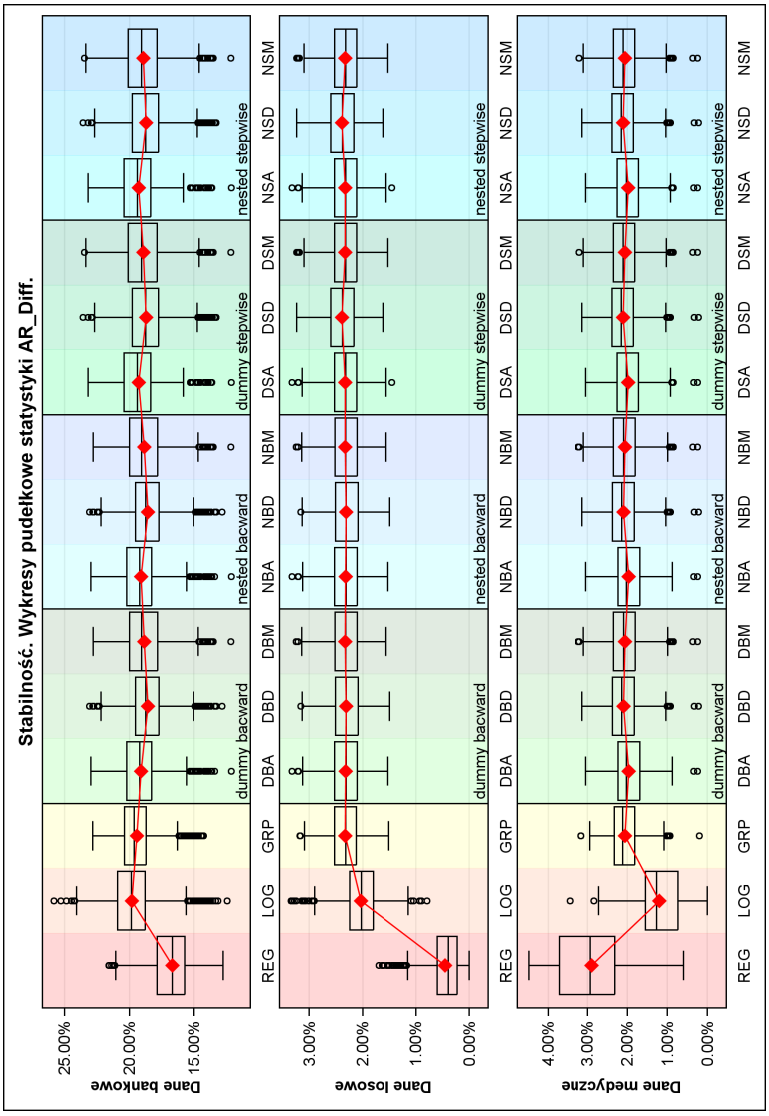
Mimo trzech różnych typów danych wyniki porównujące techniki modelowe są zbieżne. Innymi słowy można stwierdzić, że metoda porównawcza przyjęta w pracy nie jest związana z konkretnym zestawem danych. Daje to zatem możliwość testowania tylko na jednym wybranym zestawie, czyli np. na losowym generatorze danych. Ponadto zaprezentowaną metodę porównywania technik można stosować na dowolnie wybranym zestawie danych. Osoba przystępująca do modelowania, nawet jeśli spodziewa się już wyniku i przekonana jest do jakiejś techniki modelowej, może upewnić się i przeliczyć cały proces, testując na konkretnym zestawie danych. Jedynym minusem jest czas obliczeń. To kryterium skłania nas do wykonywania wielu testów na danych losowych, gdyż zawsze są one dostępne i mogą być udostępniane wszystkim zainteresowanym. Możemy je

Rysunek 5: Rozkłady jednowymiarowe. Predykcyjność



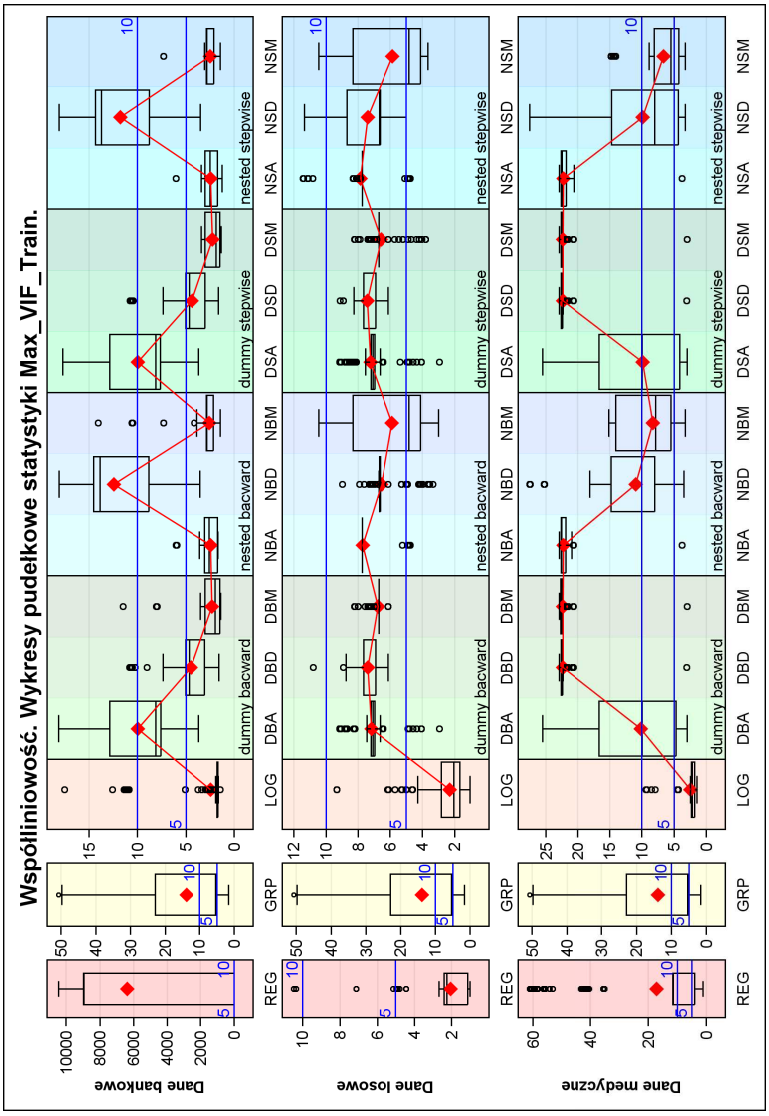
Źródło: (Przanowski i Mamczarz, 2012).

Rysunek 6: Rozkłady jednowymiarowe. Stabilność



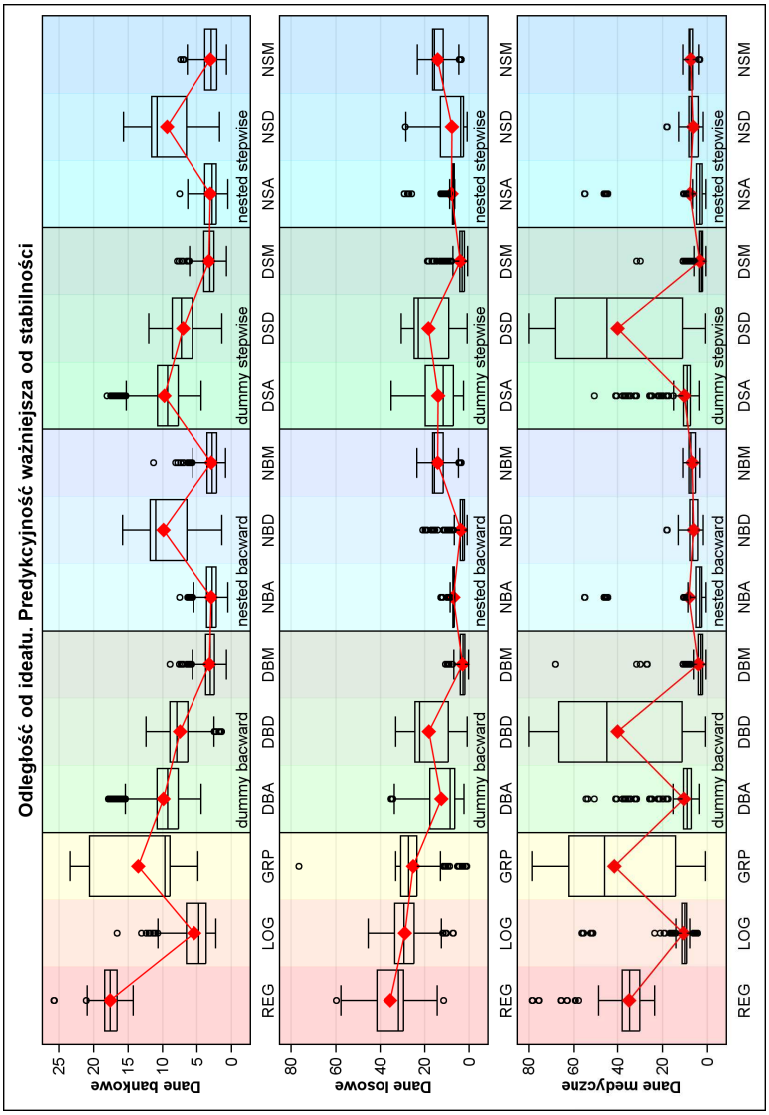
Źródło: (Przanowski i Mamczarz, 2012).

Rysunek 7: Rozkłady jednowymiarowe. Współliniowość



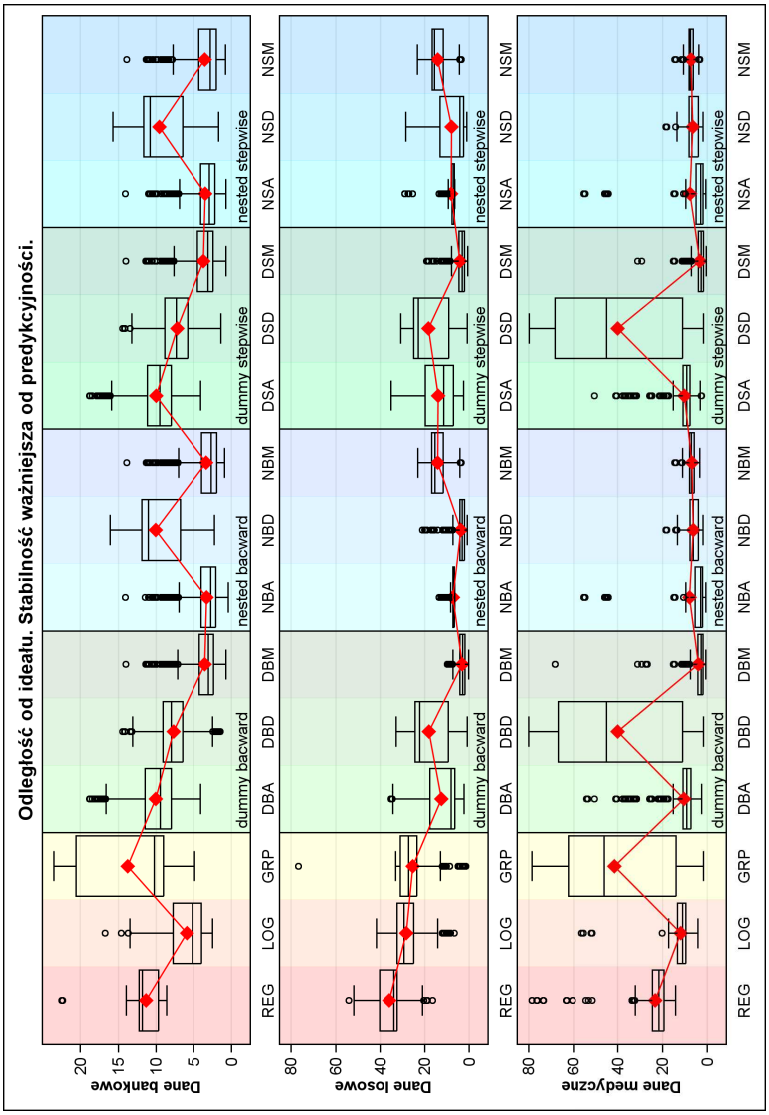
Źródło: (Przanowski i Mamczarz, 2012).

Rysunek 8: Ujęcie wielowymiarowe. Predykcyjność ważniejsza od stabilności



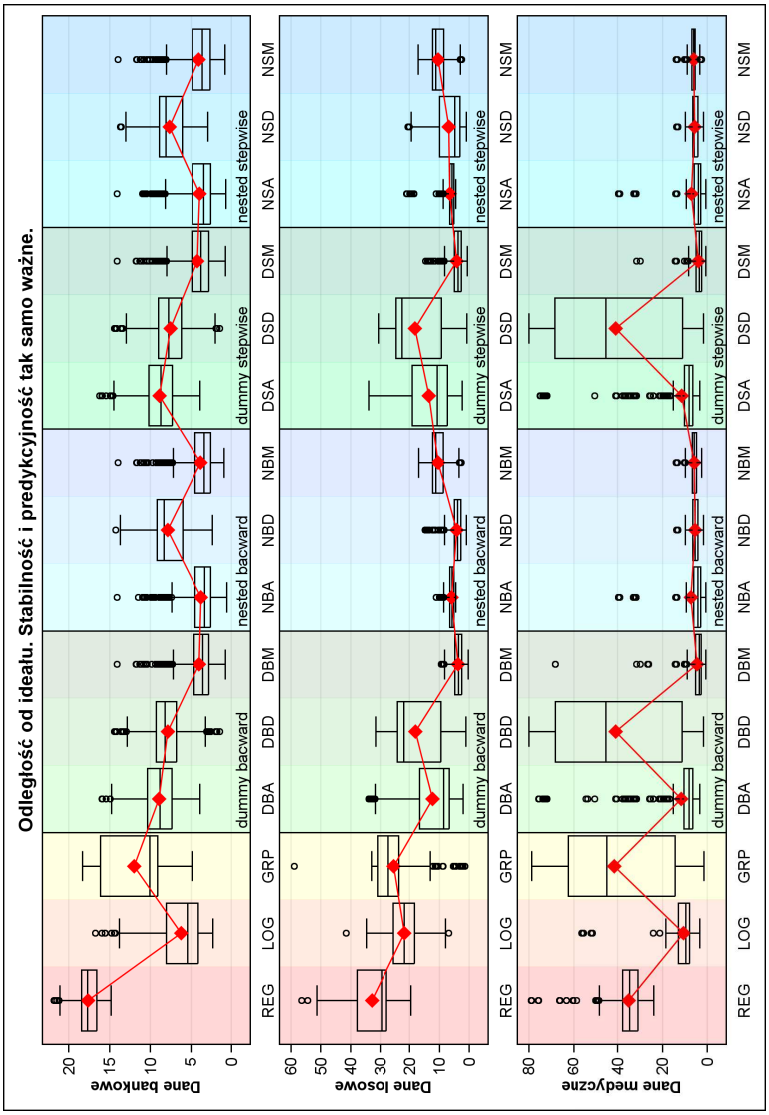
Źródło: (Przanowski i Mamczarz, 2012).

Rysunek 9: Ujęcie wielowymiarowe. Stabilność ważniejsza od predykcyjności



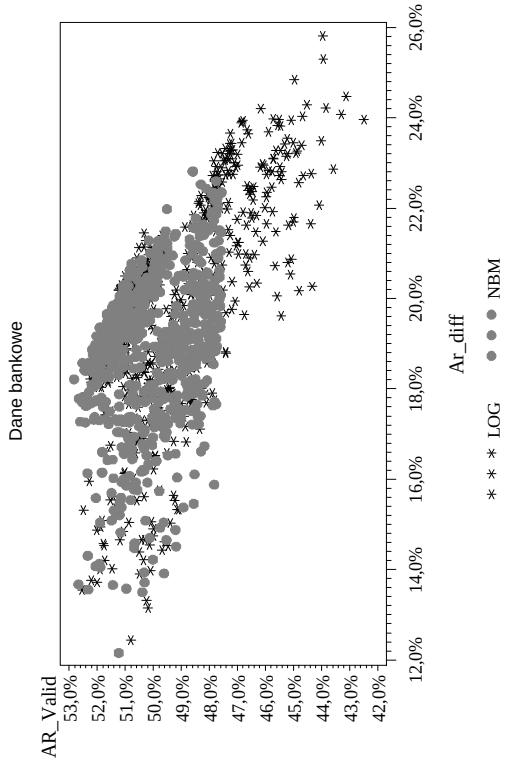
Źródło: (Przanowski i Mamczarz, 2012).

Rysunek 10: Ujęcie wielowymiarowe. Stabilność i predykcyjność tak samo ważne



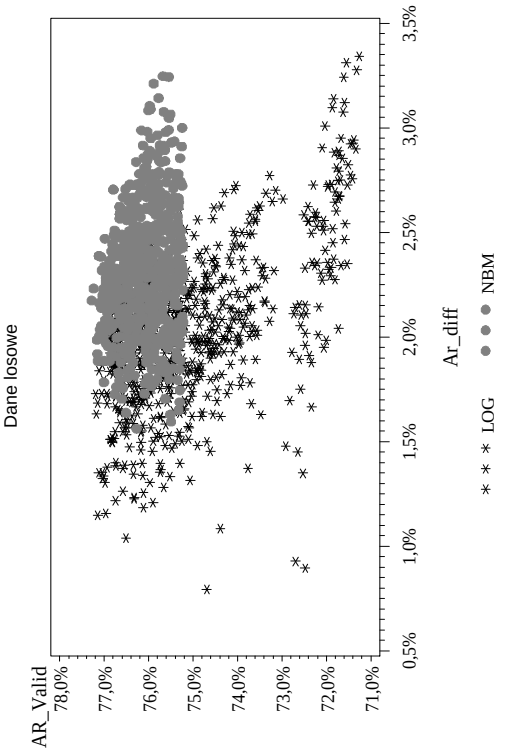
Źródło: (Przanowski i Mamczarz, 2012).

Rysunek 11: Wykres rozrzutu, porównanie metody LOG z NBM. Dane bankowe



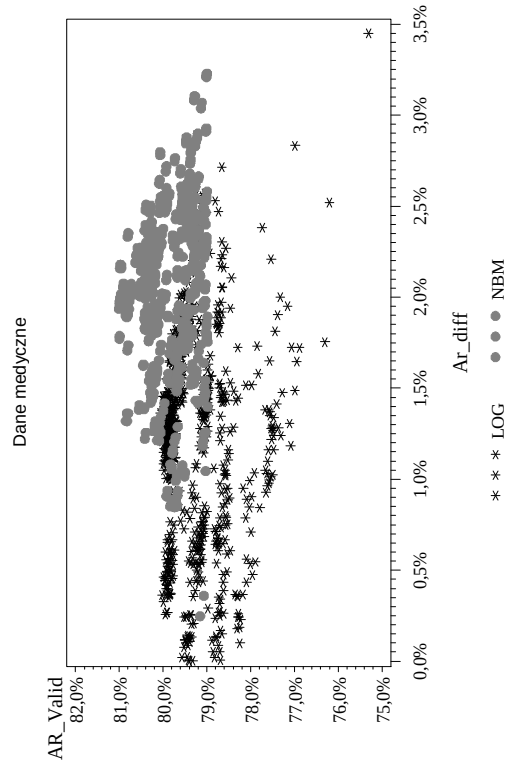
Źródło: opracowanie własne.

Rysunek 12: Wykres rozrzutu, porównanie metody LOG z NBM. Dane losowe



Źródło: opracowanie własne.

Rysunek 13: Wykres rozrzutu, porównanie metody LOG z NBM. Dane medyczne



Źródło: opracowanie własne.

dowolnie modyfikować i ulepszać, a tym samym formułowane wnioski będą się stawać coraz poprawniejsze.

Można zatem śmiało odwołać się do postawionego celu pracy. Choć zestawy danych różnią się od siebie, to istnieje możliwość stworzenia repozytorium danych Credit Scoring na bazie generatora losowego i jest możliwe testowanie na nim wielu różnych technik modelowych. Okaże się wtedy, że nie posiadamy jeszcze poprawnie definiowanych kryteriów, że trzeba dopiero stworzyć całą teorię porównywania modeli. Wreszcie trzeba stworzyć narzędzia, które będzie można uruchamiać, kiedy będzie się posiadało dostęp do rzeczywistych danych, by przetestować finalne techniki na konkretnym rzeczywistym przypadku.

3. Model biznesowy: akwizycja i sprzedaż krzyżowa

Rozdział ten stanowi rozszerzoną wersję publikacji (Przanowski, 2014).

Większość firm finansowych, mając nadzieję, że kryzys już się skończył, rozpoczęło walkę o klienta na dużą skalę. Czas kryzysu (2008–2009) był okresem, gdzie banki bardzo ostrożnie zarządzały procesem akceptacji i znacząco zmniejszały populację akceptowaną. Dziś coraz mocniej zdajemy sobie sprawę, że małe ryzyko niekoniecznie przynosi przychody, trzeba szukać złotego środka, być bardziej otwartym na ryzykownego klienta, byle nie za bardzo. Najczęściej klienci o małym ryzyku nie są aktywni kredytowo, bo właśnie dlatego ich ryzyko jest małe. Trzeba zatem zabiegać o klienta, aby chciał wziąć nowy kredyt. A ci, o których się nie zabiega i sami proszą o kredyt, często są zbyt ryzykowni i przynoszą duże straty. Pojawia się zatem potrzeba tworzenia modeli biznesowych, które zawsze mają podobny mechanizm: trzeba klienta zachęcić czymś wygodnym, atrakcyjnym i dość tanim albo zupełnie bezpłatnym, a jak się z nami zżyje, to zaproponować produkty drogie, na których będzie się zarabiać. Ogólnie tego typu modele mają strukturę: tania akwizycja, droga sprzedaż krzyżowa (ang. cross-sell). Dziś na rynku spotykamy się z dość licznymi przykładami tego typu modeli, przoduje w tym firma Google, która oferuje szeroki wachlarz produktów internetowych całkowicie za darmo, traktując to właśnie jako akwizycję. Także branża FMCG, czy AGD pełna jest przykładów: tania drukarka, drogie tusze; tani serwis do kawy, drogie kapsułki itp.

Jednym z typowych i dość już starych modeli biznesowych w sektorze bankowym jest akwizycja w postaci taniego kredytu ratalnego i sprzedaż krzyżowa kredytów gotówkowych wysoko oprocentowanych. Klient, biorący na raty lodówkę, czy telewizor plazmowy, jest bardzo zadowolony z niskich rat. Wielu z nich nigdy nie skorzysta z kredytu gotówkowego, ale pewna część przyzwyczai się do kredytów i zacznie generować przychody dla banku. Choć model ten jest powszechnie znany, bynajmniej nie jest łatwo uczynić go opłacal-

nym. Jest to najlepszy przykład zastosowania Credit Scoring i wykazania jego użyteczności w osiąganiu niemal milionowych korzyści finansowych.

Omówione metody uzyskania przykładowych danych symulacyjnych służą właśnie pogłębieniu analiz i optymalizacji modelu biznesowego, który można prosto nazwać: kredyt ratalny, potem gotówkowy. Jeśli chcemy wykazać przydatność danych losowych w badaniach nad Credit Scoring i metodach budowy kart skoringowych, to nie można zapomnieć o metodach doboru punktów odcięcia (ang. cut-off). Nie można oddzielić metod budowy modeli od ich implementacji. Całość tworzą dopiero oba tematy. Model trzeba użyć w procesach i potem jeszcze sprawdzić jego działanie. Nie zawsze wyniki potwierdzają pierwotne oczekiwania i założenia. Analiza różnic pomiędzy oczekiwanymi, prognozowanymi parametrami a osiągniętymi w rzeczywistości jest nieodzownym elementem całego procesu budowy modeli.

3.1. Parametry modelu

W podrozdziale 5.3 przedstawione są konkretne algorytmy i parametry tworzące przykładowy zestaw danych losowych. Ze względu na mnogość szczegółów, eksperckich reguł, zmiennych i zależności oraz fragmenty kodu SAS 4GL zdecydowano się umieścić je na końcu książki. Zainteresowany czytelnik może dokładnie przestudiować cały proces ich tworzenia. Dla większości osób jednak konkretna realizacja danych nie jest istotą, a jedynie przykładem danych, dzięki którym możliwe jest przedstawienie wielu aspektów Credit Scoringu w przystępny i osadzony w szczegółach sposób. Dużym atutem materiału prezentowanego w książce jest właśnie oparcie się na studiach przypadków. Wszystkie prezentowane wyniki są rezultatem konkretnych obliczeń na symulacyjnych danych, przy użyciu narzędzi programistycznych Systemu SAS. Wystarczy te same kody SAS 4GL zastosować na innym, być może pochodzącym z rzeczywistych procesów, zestawie danych i można otrzymać już raporty i narzędzia optymalizujące prawdziwe procesy w jakiejkolwiek firmie.

3.2. Wyniki symulacji, podstawowe raporty

Wszystkie obliczenia dokonano na laptopie Dell Latitude. Czas tworzenia danych kredytu ratalnego wynosi 3 godziny i 53 minuty. Zbiór Produkcja zawiera 56 335 wierszy i 20 kolumn. Zbiór Transakcje – 1 040 807 wierszy i 8 kolumn. Wartości rocznych liczb wniosków i ryzyka przedstawiono na rysunku 14.

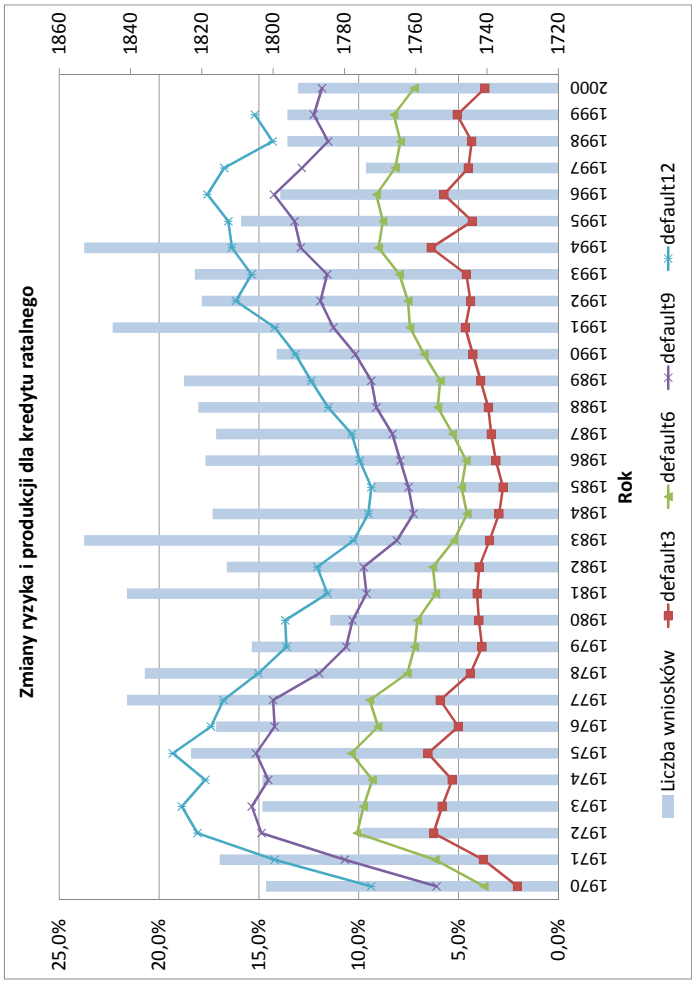
Dla kredytu gotówkowego czas obliczeń wynosi 11 godzin i 57 minut. Zbiór Produkcja_cross zawiera 60 222 wiersze i 19 kolumn, a Transakcje_cross 1 023 716 wierszy i 8 kolumn. Raport z wartości ryzyka przedstawiono na rysunku 15.

Miesięczne wielkości portfeli obu produktów razem oraz współczynnik konwersji (ang. response rate), z okresem obserwacji jeden miesiąc, przedstawiono na rysunku 16.

3.3. Implementacja modeli, system decyzyjny

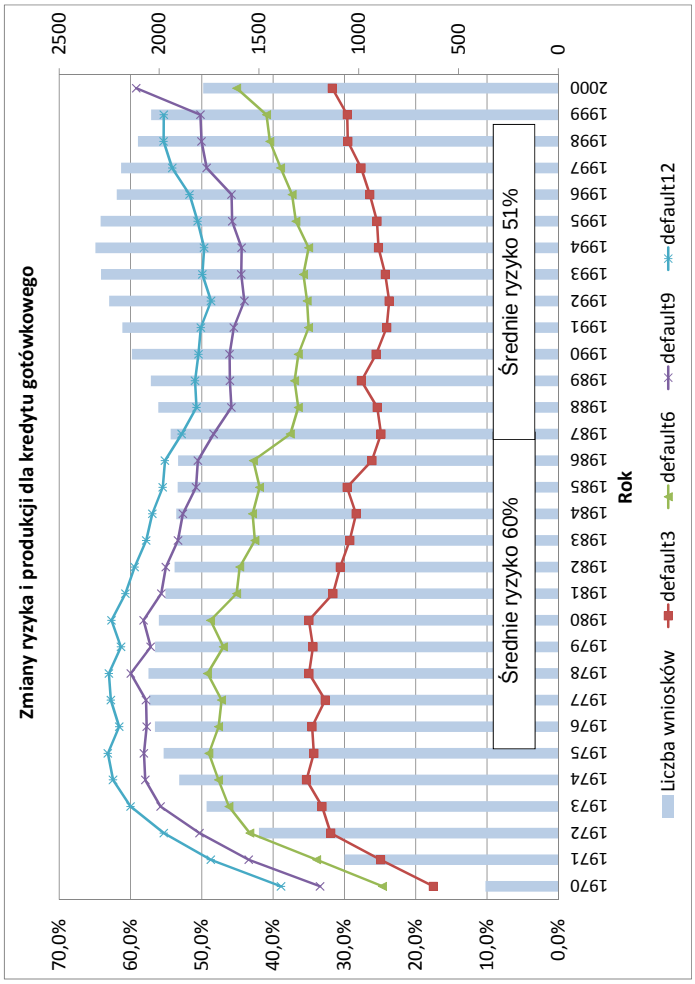
Wszystkie wygenerowane kredyty ratalne i gotówkowe traktowane są jako potencjalny portfel banku. Jak szczegółowo opisano w podrozdziale 1.2, wszystkie kredyty są tak czy inaczej brane przez klientów. Ze względu na priorytety, którymi kieruje się klient, być może także ze względu na politykę banku, nie wszystkie kredyty nawet ten sam klient spłaca tak samo. Bank może obniżyć koszty strat kredytowych, ograniczając posiadanie kredytów w jego portfelu. Umiejętny wybór kredytów jest zadaniem dla modeli skoringowych, które zaimplementowane są w systemie decyzyjnym (ang. decision engine lub scoring engine). Decyzja akceptacji powoduje, że dany kredyt brany jest do portfela banku z całą jego przyszłą historią, już z góry znaną. Zakłada się, że klient spłaca zawsze tak samo, cała jego historia jest znana, pełna i niezmienna, zmienia się tylko kredytodawca. System decyzyjny oczywiście nie zna przyszłej historii klienta, zna tylko jego historię do momentu nowego wniosku kredytowego. Co więcej, niekoniecznie jest to pełna historia kredytowa, gdyż znane są w systemie tylko te kredyty, które należały do portfela banku. Bank zatem przez swoje wcześniejsze decyzje albo posiada większą wiedzę o kliencie, albo mniejszą. Pojawia się tu bardzo ciekawy problem, czy lepiej zaakceptować klienta z jego ryzykownym kredytem

Rysunek 14: Kredyt ratalny



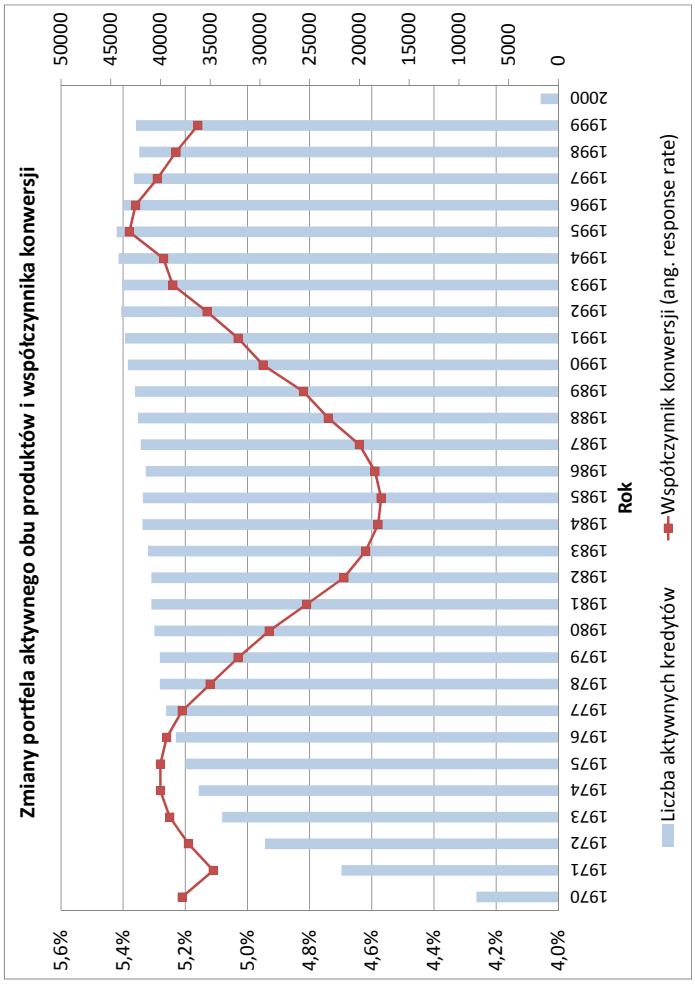
Źródło: opracowanie własne.

Rysunek 15: Kredyt gotówkowy



Źródło: opracowanie własne.

Rysunek 16: Portfele miesięczne



Źródło: opracowanie własne.

i pogodzić się z możliwością niespłacenia i powolnym odzyskiwaniem długu przez procesy naszego banku, czy też odrzucić ten kredyt i nie wiedzieć o jego historii. Jednym z rozwiązań jest posiadanie centralnych baz danych kredytowych, co właśnie w Polsce jest rozwijane i utrzymywane przez Biuro Informacji Kredytowej (BIK). Im większą wiedzę posiadamy o klientach, tym łatwiej podejmuje się właściwe decyzje i mniejszy jest błąd wynikający z istnienia wniosków odrzuconych.

Oczywiście prezentowany model zachowania klienta może być kwestionowany. Być może w rzeczywistości klient z pierwszym zdarzeniem default wpada w lawinę kolejnych i już nigdy z tej pętli nie wychodzi, finalnie stając się bankrutem zarejestrowanym we wszystkich międzybankowych bazach. Ale obserwowanym faktem jest podwójna sytuacja: część klientów spłaca kredyty dobrze w jednym banku, jednocześnie w drugim mając opóźnienia, co możemy obserwować właśnie w danych BIK. Nawet jeśli opisany model zachowania nie jest doskonały, to warto przekonać się i sprawdzić, jakie wnioski można na bazie takiego modelu sformułować.

Wszystkie wnioski z całej dostępnej historii, od 1970 do 2000 roku, system decyzyjny procesuje około 50 minut. Rozważmy sytuację skrajną: wszystkie wnioski są akceptowane. Można ten przypadek uważać za początkowy okres funkcjonowania banku, gdzie bank, dysponując kapitałem początkowym, akceptuje początkowe straty, aby nauczyć się w przyszłości optymalizować procesy. Niestety obecnie banki się zabezpieczają i zawsze mają jakieś reguły i modele skoringowe, często kupowane od firm konsultingowych. Niemniej rozważmy taki przypadek, który być może lepiej jest nazwać rajskim, bo tylko w raju nawet bankruci będą mieli te same szanse.

Na podstawie pewnej historii kredytowej możliwe jest zbudowanie modeli skoringowych. Przypuśćmy, że korzystamy z historii pomiędzy latami 1975 i 1987. Jest to już okres, kiedy procesy się w miarę ustabilizowały. Ale ryzyko w tym czasie jest większe od drugiego w latach 1988–1998, w którym korzystamy już z modeli, mając nadzieję otrzymania istotnych zysków. Zostały zbudowane cztery przykładowe modele akceptacyjne (obliczane na czas wniosku kredytowego). Trzy modele prognozujące zdarzenie $\mathrm{default}_{12}$:

model ryzyka dla kredytu ratalnego (oznaczenie PD Ins), model ryzyka dla kredytu gotówkowego (PD Css), model ryzyka dla kredytu gotówkowego w momencie aplikowania o kredyt ratalny (Cross PD Css) oraz dodatkowo model skłonności skorzystania z kredytu gotówkowego w momencie aplikowania o kredyt ratalny (PR Css, zdarzenie response w okresie sześciu miesięcy obserwacji). Opisy szczegółowe tych modeli prezentowane są w rozdziale 5.2. Każdy model finalnie oblicza prawdopodobieństwo modelowanego zdarzenia.

Podstawowym zadaniem jest sprawienie, aby proces był opłacalny i zmaksymalizował zysk. W tym procesie przyjmujemy następujące parametry: roczne oprocentowanie kredytu ratalnego $= 1\%$, oprocentowanie kredytu gotówkowego $= 18\%$, zerowe prowizje obu kredytów. Średnie wartości LGD przyjmujemy: 45% dla ratalnego, a 55% dla gotówkowego. Dodatkowo wszystkie kredyty gotówkowe zostały wygenerowane ze stałą kwotą kredytu $= 5\ 000$ PLN oraz okresem kredytowania $= 24$ miesiące. Wskaźniki finansowe procesu dla okresu modelowego 1975–1987 przedstawiono w tabeli 9 (metoda wyliczania oparta na wzorze 2.1, strona 48). W tabeli 10 przedstawiono moce modeli predykcyjnych. Są one nawet za duże, jak na możliwości rzeczywistych procesów, szczególnie model response (PR Css), gdyż w rzeczywistości uzyskuje się modele z mocą mniejszą niż 80%. Z drugiej strony obserwujemy tu modele budowane na całej dostępnej historii kredytowej, takich liczb nie da się obserwować w rzeczywistości, są one ukryte. Tylko na danych losowych możemy próbować zobaczyć ich stan początkowy, bez wpływu wniosków odrzuconych.

Średnie ryzyko procesu z tego okresu wynosi $37{,}19\%$, a średnie prawdopodobieństwo (PD) $34{,}51\%$ jest lekko niedoszacowane. Całkowity zysk około -40 mPLN. Nie lada wyzwaniem staje się doprowadzenie tego procesu do opłacalności. Problem jest na tyle poważny, że obecne opracowanie daje tylko pierwsze propozycje, wskazując istotną rolę danych losowych w dalszych badaniach naukowych. Obecnie rozwijają się coraz bardziej zagadnienia opłacalności, nazywane ogólnie z ang. Customer LifeTime Value, CLTV, lub CLV (Ogden, 2009; DeBonis *et al.*, 2002). W przypadku nasze-

Tabela 9: Wskaźniki finansowe procesu dla strategii akceptacji wszystkich kredytów (okres 1975–1987)

Wskaźnik	Ratalny	Gotówkowy	Razem
Zysk	–7 824 395	–31 627 311	–39 451 706
Przychód	969 743	10 260 689	11 230 432
Strata	8 794 138	41 888 000	50 682 138

Źródło: opracowanie własne.

Tabela 10: Moce predykcyjne modeli skoringowych (1975–1987)

Model	Gini (%)
Cross PD Css	74,01
PD Css	74,21
PD Ins	73,11
PR Css	86,37

Źródło: opracowanie własne.

go procesu uproszczoną wersją modelu CLTV jest model Cross PR Css.

Przejdźmy zatem przez etapy wyznaczania sensownych parametrów procesu. Pierwszy parametr znajdujemy, analizując krzywą Profit tylko dla kredytu gotówkowego. Przy akceptacji $18{,}97\%$ uzyskujemy największy zysk o wartości $1\ 591\ 633$ PLN. W systemie decyzyjnym dodajemy regułę, gdy $PD\_Css > 27{,}24\%$, to wniosek odrzucamy. Z punktu widzenia analiz CLTV powinno się do zagadnienia podejść bardziej dokładnie i rozważyć ciąg zaciąganych kredytów gotówkowych tego samego klienta, gdyż odrzucenie pierwszego z nich blokuje kolejne. Osoba, która nie jest klientem banku, nie otrzyma oferty, nie będzie miała zatem możliwości otrzymania kredytu. Bardzo prawdopodobne jest, że wiele kredytów gotówkowych tego samego klienta moglibyśmy akceptować na innych zasadach i sumarycznie otrzymać większy zysk. Tego typu rozumowanie przeprowadźmy tylko w sytuacji konwersji z kredytu ratalnego do gotówkowego. Rozważmy moment aplikowania o kredyt ratalny, który z góry jest kredytem raczej nieopłacalnym. Jeśli będziemy starali się go uczynić zyskownym, to będziemy akceptować bardzo mało kredytów i nie damy okazji wzięcia w przyszłości kredytu gotówkowego, który ma znacząco większe oprocentowanie.

Rozważamy tylko produkcję kredytu ratalnego, z jego globalnym zyskiem połączonych transakcji, czyli aktualnie wnioskowanego ratalnego i przyszłego kredytu gotówkowego, używając do tego celu dodatkowych modeli Cross PD Css i PR Css. Tworzymy po pięć segmentów dla modeli PD Ins i PR Css, uwzględniając już w obliczeniach pierwszą regułę na PD_Css ustaloną wcześniej dla kredytu gotówkowego. Dla każdej kombinacji segmentów obliczamy globalny zysk; patrz tabela 11. Analizując segmenty, tworzymy kolejne reguły:

Dla kredytu ratalnego, jeżeli $PD\_Ins > 8{,}19\%$, to odrzucamy oraz jeżeli $8{,}19\% \geqslant PD\_Ins > 2{,}18\%$ i ($PR\_Css < 2{,}8\%$ lub $Cross\_PD\_Css > 27{,}24\%$), to też odrzucamy.

Zwróćmy uwagę, że wprowadzona została dodatkowa reguła nie tylko oparta na mierniku ryzyka, ale także mówiąca tyle: jeśli ryzyko ratalnego jest w górnej półce, to, aby się opłacało, musimy mieć gwarancję wzięcia przyszłego kredytu gotówkowego na ustalonym

poziomie prawdopodobieństwa oraz że będzie on odpowiednio na niskim poziomie ryzyka.

Tak ustawione reguły procesu w sumie przynoszą $1\,686\,684$ PLN globalnego zysku z obu produktów razem. Gdyby nie było dodatkowej reguły związanej z przyszłym kredytem gotówkowym, to akceptowali byśmy wszystkie kredyty ratalne spełniające tylko jedną regułę: $PD\_Ins \leqslant 8{,}19\%$ i proces przyniósłby zysk $1\,212\,261$ PLN. Stracilibyśmy zatem około 470 kPLN, co stanowiłoby prawie o 30% mniejsze zyski.

Niestety przedstawione wyniki nie uwzględniają poprawnie wpływu wniosków odrzuconych. Należy zatem to sprawdzić ponownie, procesując wszystkie wnioski w systemie decyzyjnym. Tylko wtedy poznamy realny wpływ informacji ukrytej, powstałej na skutek odrzuconych wniosków. Symulacje różnych strategii prezentowane są w podrozdziale 3.3.1.

Tabela 11: Kombinacje segmentów i ich globalne zyski (1975–1987)

GR PR Css	GR PD Ins	Liczba Ins	Globalny zysk	Min (%) PR Css	Max (%) PR Css	Min (%) PD Ins	Max (%) PD Ins
4	0	1 277	372 856	4,81	96,61	0,02	2,18
4	1	581	96 096	4,81	96,61	2,25	4,61
1	0	2 452	67 087	1,07	1,07	0,32	2,18
3	0	907	46 685	2,80	4,07	0,07	2,18
3	1	734	14 813	2,80	4,07	2,25	4,61
3	2	307	12 985	2,80	4,07	4,76	7,95
4	2	361	8 039	4,81	96,25	4,76	7,95
3	3	446	–1 283	2,80	4,07	8,19	18,02
4	3	417	–5 774	4,81	95,57	8,19	18,02
1	1	3 570	–82 886	1,07	1,07	2,25	4,61
1	2	4 044	–408 644	1,07	1,07	4,76	7,95
3	4	726	–946 937	2,80	4,07	18,50	99,62
4	4	1 054	–1 108 313	4,81	96,25	18,50	99,83
1	3	3 883	–1 270 930	1,07	1,07	8,19	18,02
1	4	2 878	–4 306 859	1,07	1,07	18,50	97,00

Źródło: opracowanie własne.

3.3.1. Testowanie różnych strategii akceptacji

W celu głębszego zrozumienia problemów związanych w wdrażaniem modeli skoringowych rozważono cztery różne strategie decyzyjne (tabele 12, 13, 14 i 15). Strategia pierwsza związana jest z metodą wyznaczenia optymalnych reguł przedstawioną w podrozdziale 3.3. Zauważmy, że prognozowaliśmy zysk procesu na poziomie $1\ 686\ 684$ PLN w okresie 1975–1987. Po przeliczeniu procesu akceptacji z nowymi regułami okazało się, że prawdziwy zysk wynosi: $663\ 327$ PLN. Pomyliliśmy się aż o jeden milion złotych, jest to bardzo duży błąd. Można by oczywiście ulepszać naszą metodę, uwzględniać więcej czynników i poprawniej identyfikować wnioski odrzucone (30% udziału) lub te, które nie mogły być zrealizowane z racji nieaktywnego klienta na moment wnioskowania kredytu gotówkowego (50% udziału). Z drugiej strony możemy i tak być zadowoleni: zamiast zysku ujemnego -40 mPLN mamy około 700 tysięcy na plus. Niemniej błąd jest na tyle duży, że uświadamia potrzebę głębszych studiów i pokazuje, że sama budowa modelu z dużymi wskaźnikami predykcyjnymi nie gwarantuje sukcesu. Dopiero wdrożenie, jego wszystkie składowe kroki poprawnie wykonane, zagwarantują nam osiągnięcie spodziewanych korzyści. Aczkolwiek do końca nie da się przewidzieć skutków gwałtownej, niemal rewolucyjnej zmiany strategii. Zauważmy, że zastosowaliśmy strategię pierwszą po strategii, gdzie akceptowaliśmy wszystkich klientów. Ze 100% akceptacji przeszliśmy na 26% akceptacji kredytu ratalnego i na 16% – gotówkowego. Tak mocna zmiana całkowicie zaburzyła rozkłady naszych ocen punktowych, a tym samym i wartości prawdopodobieństw.

Zauważmy dodatkowo, że procent akceptacji dla kredytu gotówkowego $16{,}23\%$ w rzeczywistości będzie miał inną wartość. Związane jest to z niemożliwymi do zaobserwowania liczbami, które nazwiemy niewidzialnymi. Liczba klientów nieznanych jest możliwa do poznania tylko dzięki danym symulacyjnym. W rzeczywistości pomniejszy ona mianownik i procent akceptacji będzie wynosił około 33%. Problem jest jeszcze poważniejszy, a mianowicie można zadać sobie pytanie: na ile procent akceptacji kredytu ratalnego wpływa na procent akceptacji kredytów gotówkowych? Jest to już dość

trudne do wykazania na bazie danych symulacyjnych. Z reguły mamy dość dużą skłonność do zaciągania kredytów wśród klientów, którzy przy kredycie ratalnym są w okolicy punktu odcięcia, czyli są na granicy akceptowalnego ryzyka. Jeśli zatem nieznacznie zmienimy akceptowalność kredytu ratalnego, to w skrajnych przypadkach bardzo znacząco może zmienić się procent akceptacji kredytów gotówkowych. Może być jeszcze inna sytuacja. Sam procent akceptacji kredytów gotówkowych nadal będzie podobny, ale liczba akceptowanych kredytów gotówkowych może znacząco się zmniejszyć. Trzeba zawsze mieć w świadomości, że grupa „nieznany klient" odgrywa poważną rolę w procesie sprzedaży krzyżowej i dlatego w procesie akceptacji produktów akwizycyjnych powinny brać udział także modele predykcyjne prognozujące przyszłe zachowanie klienta z potencjalnymi produktami sprzedaży krzyżowej.

W nowym procesie ten sam model na tym samym okresie pokazuje prawdopodobieństwo PD o wartości $28{,}87\%$ dla wszystkich wniosków. Przypomnijmy, że model PD wykazywał wcześniej średnią równą $34{,}51\%$. Skąd pojawiła się różnica? Związana jest z brakiem informacji w danych banku o wszystkich kredytach klientów. Ponieważ nowa strategia akceptuje lepsze kredyty, bank posiada informację tylko o lepszych kredytach jego klientów, średnia wartość PD musi zatem się zmniejszyć, a co za tym idzie otrzymujemy niedoszacowaną jego wartość. Gdybyśmy teraz chcieli sytuację odwrócić, na bazie danych z procesu strategii pierwszej próbować więcej akceptować, to możemy liczyć się z dość dużym błędem niedoszacowanego ryzyka. Zauważmy, że nie tylko rozkłady parametrów PD się zmieniły, ale także moce predykcyjne modeli. Gini modelu Cross PD Css z 74% spadł do 41% na całości, a na części zaakceptowanej aż do 21%. Być może nawet analityk budujący ten model miałby dość poważne problemy z przełożonymi, gdyż pewnie przed wdrożeniem chwalił się swoimi osiągnięciami, będąc dumnym i oczekując nagrody. Zwróćmy uwagę, że w realnym świecie obserwujemy tylko liczbę 21%, a ta druga jest możliwa tylko w badaniach na danych symulacyjnych. Można by pytać się: czy model ten przestał działać? Nic bardziej mylnego, on działa, ale na części, której już nie mamy w danych; działa i odrzuca poprawne wnioski. Niestety trudno jest zmierzyć jego użyteczność po wdrożeniu. Najprawdopodobniej zbu-

dowano by nowy model i pewnie wcale nie lepszy. Obserwowane pomiary po wdrożeniu są obarczone zmianą populacji i nie zawsze da się przewidzieć tego skutki.

Mamy tu przykład bardzo poważnego zjawiska, które staje się dewizą Zarządzania Kompleksową Jakością, z ang. Total Quality Management, TQM (Blikle, 1994): świadomy menadżer podejmuje decyzje na podstawie danych oraz liczb zarówno obserwowanych, jak i ukrytych, nieobserwowanych (niewidzialnych). Brak możliwości poznania niektórych wskaźników nie powinien powodować ich pomijania przy podejmowaniu decyzji. Nawet jeśli nie znamy niektórych liczb, to i tak mogą być one kluczowe w zarządzaniu.

W metodzie opisanej w rozdziale 3.3 sprawdzane jest też, co by było, gdyby reguła przy akceptacji kredytów ratalnych była prosta, pozbawiona kryterium dla PR_Css. Prognozowano, że wtedy zysk spadnie o 470 tysięcy. W rzeczywistości zysk spadł o 550 tysięcy, czyli w przyrostach i spadkach popełniliśmy mniejszy błąd.

Wpływ wniosków odrzuconych jest na tyle nieprzewidywalny na ile dużą zmianę przeprowadza się w procesie. Być może bank powinien w pierwszym okresie swojej działalności wprowadzić prostą regułę: odrzucać klientów, którzy w historii ostatnich 12 miesięcy mieli przeterminowanie powyżej trzech zaległych rat, czyli wystąpiło na nich zdarzenie default. Strategia trzecia właśnie ten proces realizuje. Choć proces ten nie jest jeszcze opłacalny, to już znacząco zmniejszył akceptację kredytu gotówkowego – aż do 45%. Modele nadal wykazują dużą moc predykcyjną. Przeprowadzając na takich danych podobne rozumowanie i ustalając nowe punkty odcięcia, można dojść do lepszego rozwiązania: strategii czwartej. Ponieważ punktem wyjścia jest strategia trzecia, błąd szacunku zysku okazuje się być mniejszy i w efekcie strategia czwarta przynosi najlepsze korzyści finansowe. Zysk wynosi tu 732 tysiące złotych, przy 9% akceptacji kredytu gotówkowego i 26% ratalnego. Być może strategia ta nie jest dobra ze względu na zbyt małe procenty akceptacji. Zauważmy, że w okresie 1988–1998 mamy odwrócenie. Strategia pierwsza przynosi zysk $1{,}5$ mPLN a czwarta $1{,}3$ mPLN. Najlepsze zatem metody, to adaptacyjne, ciągłe testowanie nowych strategii dostosowujących się do nowych warunków, do zmian ryzyka w czasie.

Niestety w drugim okresie ryzyko jest mniejsze i strategia akceptująca więcej klientów wygrywa.

Zupełnie nietestowanym w pracy, a także bardzo ważnym, wydaje się temat dostosowania warunków cenowych kredytów do prognozowanych wartości ryzyka i przyszłych zaciąganych kredytów gotówkowych, czyli modeli CLTV. Być może strategia pierwsza wzbogacona o zróżnicowaną cenę byłaby najlepsza w każdym okresie. Nic nie stoi na przeszkodzie, aby dalej rozwijać badania strategii na bazie danych losowych i lepiej przybliżać się do zrozumienia złożoności procesów bankowych. Z prac naukowych prowadzonych razem ze studentami wynika, że polityka cenowa przynosi mniejsze korzyści od inwestycji w lepszy model predykcyjny z większą wartością statystyki Giniego. Właściwe dobieranie parametrów cenowych zachęca jednak większą liczbę kredytobiorców i zmniejsza poziom strat dla klientów bardzo ryzykownych.

Tabela 12: Strategia 1

Okres	Przychód	Strata	Zysk
1975–1987	3 407 745	2 744 418	663 327
1988–1998	3 761 299	2 246 844	1 514 455

Reguła	Opis
PD_Ins Cutoff	$PD\_Ins > 8{,}19\%$
PD_Css Cutoff	$PD\_Css > 27{,}24\%$
PD i PR	$8{,}19\% \geqslant PD\_Ins > 2{,}18\%$ i ($PR\_Css < 2{,}8\%$ lub $Cross\_PD\_Css > 27{,}24\%$)

Kredyt gotówkowy

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko (%)	Zysk
PD_Css Cutoff	8 436	32,97	42 180 000	67,99	–13 098 591
Nieznany klient	12 999	50,80	64 995 000	65,91	–19 171 357
Akceptacja	4 152	16,23	20 760 000	22,35	642 637
Razem	25 587	100,00	127 935 000	59,53	–31 627 311

Kredyt ratalny

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko (%)	Zysk
PD_Ins Cutoff	9 289	39,30	60 214 008	26,95	–7 339 423
PD i PR	8 131	34,40	31 340 808	5,37	–505 662
Akceptacja	6 217	26,30	22 698 240	2,14	20 690
Razem	23 637	100,00	114 253 056	13,00	–7 824 395

Średnie wartości parametrów (%)

Parametr	Akceptacja	Razem
PD (razem PD Ins i PD Css)	7,93	28,87
PR Css	17,15	21,76
Cross PD Css	21,71	17,73

Moc predykcyjna (Gini w %)

Model	Akceptacja	Razem
Cross PD Css	21,34	40,72
PD Css	31,66	53,28
PD Ins	41,93	68,58
PR Css	72,56	68,88

Źródło: opracowanie własne.

Tabela 13: Strategia 2

Okres	Przychód	Strata	Zysk
1975–1987	4 008 258	3 896 818	111 441
1988–1998	4 539 328	3 829 634	709 694

Reguła	Opis
PD_Ins Cutoff	$PD\_Ins > 8{,}19\%$
PD_Css Cutoff	$PD\_Css > 27{,}24\%$

Kredyt gotówkowy

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko (%)	Zysk
PD_Css Cutoff	9 297	36,33	46 485 000	67,84	–14 381 482
Nieznany klient	11 661	45,57	58 305 000	67,34	–17 822 432
Akceptacja	4 629	18,09	23 145 000	23,16	576 604
Razem	25 587	100,00	127 935 000	59,53	–31 627 311

Kredyt ratalny

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko (%)	Zysk
PD_Ins Cutoff	9 325	39,45	60 221 856	26,98	–7 359 232
Akceptacja	14 312	60,55	54 031 200	3,89	–465 163
Razem	23 637	100,00	114 253 056	13,00	–7 824 395

Średnie wartości parametrów (%)

Parametr	Akceptacja	Razem
PD (razem PD Ins i PD Css)	6,82	29,05
PR Css	12,79	22,89
Cross PD Css	17,62	18,34

Moc predykcyjna (Gini w %)

Model	Akceptacja	Razem
Cross PD Css	19,39	39,86
PD Css	31,23	55,05
PD Ins	41,73	69,04
PR Css	80,56	64,40

Źródło: opracowanie własne.

Tabela 14: Strategia 3

Okres	Przychód	Strata	Zysk
1975–1987	7 496 614	21 801 230	–14 304 616
1988–1998	7 881 992	18 510 342	–10 628 350

Reguła	Opis
Zły klient	$\mathrm{agr12\_Max\_CMaxA\_Due} > 3$

Kredyt gotówkowy

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko (%)	Zysk
Zły klient	7 114	27,80	35 570 000	79,83	–14 195 320
Nieznany klient	7 036	27,50	35 180 000	67,04	–10 673 871
Akceptacja	11 437	44,70	57 185 000	42,28	–6 758 120
Razem	25 587	100,00	127 935 000	59,53	–31 627 311

Kredyt ratalny

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko (%)	Zysk
Zły klient	483	2,04	2 047 188	27,74	–277 899
Akceptacja	23 154	97,96	112 205 868	12,69	–7 546 496
Razem	23 637	100,00	114 253 056	13,00	–7 824 395

Średnie wartości parametrów (%)

Parametr	Akceptacja	Razem
PD (razem PD Ins i PD Css)	21,81	32,70
PR Css	21,79	28,83
Cross PD Css	43,09	24,48

Moc predykcyjna (Gini w %)

Model	Akceptacja	Razem
Cross PD Css	64,83	63,59
PD Css	63,67	64,82
PD Ins	71,94	72,56
PR Css	79,96	64,72

Źródło: opracowanie własne.

Tabela 15: Strategia 4

Okres	Przychód	Strata	Zysk
1975–1987	2 010 242	1 278 361	731 882
1988–1998	2 452 716	1 134 729	1 317 986

Reguła	Opis
Zły klient	$\mathrm{agr12\_Max\_CMaxA\_Due} > 3$
PD_Ins Cutoff	$PD\_Ins > 7{,}95\%$
PD_Css Cutoff	$PD\_Css > 19{,}13\%$
PD i PR	$7{,}95\% \geqslant PD\_Ins > 2{,}8\%$ i ($PR\_Css < 2{,}8\%$ lub $Cross\_PD\_Css > 19{,}13\%$)

Kredyt gotówkowy

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko (%)	Zysk
Zły klient	2 253	8,81	11 265 000	74,26	–4 026 033
PD_Css Cutoff	5 375	21,01	26 875 000	53,66	–5 462 687
Nieznany klient	15 739	61,51	78 695 000	65,29	–22 845 756
Akceptacja	2 220	8,68	11 100 000	17,97	707 165
Razem	25 587	100,00	127 935 000	59,53	–31 627 311

Kredyt ratalny

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko (%)	Zysk
Zły klient	209	0,88	891 720	27,75	–121 550
PD_Ins Cutoff	9 253	39,15	60 130 704	26,46	–7 208 030
PD i PR	8 029	33,97	31 118 232	5,49	–519 531
Akceptacja	6 146	26,00	22 112 400	2,05	24 717
Razem	23 637	100,00	114 253 056	13,00	–7 824 395

Średnie wartości parametrów (%)

Parametr	Akceptacja	Razem
PD (razem PD Ins i PD Css)	4,24	25,17
PR Css	11,37	15,68
Cross PD Css	17,02	14,61

Moc predykcyjna (Gini w %)

Model	Akceptacja	Razem
Cross PD Css	3,23	19,19
PD Css	33,15	47,81
PD Ins	36,79	67,67
PR Css	70,59	64,89

Źródło: opracowanie własne.

3.3.2. Paradoks Customer Bank Seniority. Wpływ wniosków odrzuconych

Rozważmy jedną konkretną zmienną o nazwie ACT_CCSS_SENIORITY. Reprezentuje ona liczbę miesięcy od pierwszego wniosku kredytu gotówkowego liczoną w momencie aplikowania o kolejny kredyt gotówkowy. Zmienną tę można nazwać czasem życia klienta w banku lub z ang. Customer Bank Seniority lub Time in Booking. Zanim przeanalizowane zostaną jej własności na wstępie przeprowadźmy proste rozumowanie. W sytuacji pełnej akceptacji (zwaną rajską), gdzie akceptowane są zawsze wszystkie wnioski, klient z dłuższą historią w banku może okazać się bardziej ryzykowny. Tak jak w prostym doświadczeniu polegającym na rzucie kostką, im więcej razy rzucamy, tym większe jest prawdopodobieństwo wypadnięcia sześciu oczek. Zatem im częściej klient jest aktywny, im dłużej posiada swoje produkty bankowe, tym jego szansa na bankructwo większa. Zauważmy, że teza ta potwierdza obserwowane wyniki dla wyznaczonych kategorii tej zmiennej w przypadku strategii pełnej akceptacji. W tabeli 16 pokazano, że najbardziej ryzykowny jest klient z historią pomiędzy 25 i 57 miesięcy. Nie mamy tu monotonicznych przedziałów, może to być efektem konkretnego przypadku danych. Najmniej ryzykowny jest klient z pustą historią, czyli nowy. Sytuacja ta wydaje się absurdalna w rzeczywistym świecie. Każdy analityk bankowy bez wahania stwierdzi, że klient nowy, bez historii bankowej, jest najbardziej ryzykowny. Co więcej, zmienna czas życia klienta zawsze uważana jest za predyktywną, dyskryminuje z mocą około $20-30\%$ wartości statystyki Giniego. Zawsze klient z dłuższą historią jest bardziej wiarygodny i ma mniejsze ryzyko. Można tu pozwolić sobie na drobny żart: najprostszym i w miarę skutecznym modelem skoringowym jest układ dwóch zmiennych: PESEL klienta i jego bankowy identyfikator. W pierwszej ukryta jest informacja o wieku, a w drugiej – o czasie życia klienta w banku. Oba czasy życia dają z reguły około 40% Giniego, ale model taki jest dość trudny do przedstawienia audytowi.

Dlaczego zatem w sytuacji realnej zmienna ma inną własność? Dzieje się tak dlatego, że w bankach zawsze istnieją reguły odrzucania klientów źle spłacających. Jeśli posiada się złą historię kredyto-

Tabela 16: Kategorie zmiennej ACT_CCSS_SENIORITY przy pełnej akceptacji

Numer	Warunek	Ryzyko (%)	Procent	Liczba
1	$25 < \text{ACT\_CCSS\_SENIORITY} \leqslant 57$	71,50	19,42	2 684
2	$18 < \text{ACT\_CCSS\_SENIORITY} \leqslant 25$	68,74	6,50	899
3	$57 < \text{ACT\_CCSS\_SENIORITY} \leqslant 67$	61,40	6,00	829
4	$67 < \text{ACT\_CCSS\_SENIORITY} \leqslant 140$	59,66	37,00	5 114
5	$140 < \text{ACT\_CCSS\_SENIORITY}$	54,86	17,55	2 426
6	$\text{ACT\_CCSS\_SENIORITY} \leqslant 18$	49,47	6,14	849
7	Missing	34,90	7,38	1 020
	Razem	59,36	100,00	13 821

Źródło: opracowanie własne.

Tabela 17: Kategorie zmiennej ACT_CCSS_SENIORITY przy strategii 3

Numer	Warunek	Ryzyko (%)	Procent	Liczba
1	$18 < \text{ACT\_CCSS\_SENIORITY} \leqslant 41$	59,73	16,34	1 125
2	$41 < \text{ACT\_CCSS\_SENIORITY} \leqslant 53$	47,97	3,94	271
3	$\text{ACT\_CCSS\_SENIORITY} \leqslant 18$	46,14	11,30	778
4	$53 < \text{ACT\_CCSS\_SENIORITY} \leqslant 142$	42,51	37,42	2 576
5	$142 < \text{ACT\_CCSS\_SENIORITY} \leqslant 184$	34,53	12,12	834
6	Missing	31,65	15,24	1 049
7	$184 < \text{ACT\_CCSS\_SENIORITY}$	25,10	3,65	251
	Razem	42,69	100,00	6 884

Źródło: opracowanie własne.

wą, to nie otrzyma się już nowego kredytu. Zatem klient, który ma długą historię, musi być wiarygodny. Potwierdza to też zestaw nierówności zmiennej czasu życia klienta w banku w przypadku strategii 3, gdzie klienci ze zdarzeniami default w ciągu ostatnich 12 miesięcy są odrzucani, co pokazano w tabeli 17. Kategorie znacząco się zmieniły. Klient z najdłuższą historią tym razem jest najmniej ryzykowny.

Rozważmy teraz sytuację, w której model estymowany na danych z zaakceptowanych wniosków według strategii 3 będzie wdrożony do systemu i wyprze istniejącą tam regułę odrzucania klientów z historycznym default. Może tak się stać, jeśli analityk nie będzie dostatecznie zdawał sobie sprawy z poprzedniej strategii i z wpływu wniosków odrzuconych. Jeśli zatem do tego by doszło, to model będzie odrzucał, być może, trochę za dużo nowych klientów, a akceptował dwa rodzaje klientów z długą historią. Pierwszy rodzaj jest pożądany, bo są to klienci, którzy raz na jakiś czas biorą kredyt i go spłacają. Drugi jest mniej bezpieczny, ponieważ to klienci, którzy wpadli w pętlę przekredytowania i coraz bardziej się zadłużają. Przy niektórych już kredytach klienci zaczynają wpadać w opóźnienia, ale model patrzy na czas życia klienta, więc takich też zaakceptuje. Jest to oczywiście sytuacja skrajna. Niemniej pokazuje ona bardzo wyraźnie, że wszelkie estymacje są uwarunkowane, że ważna jest stara strategia i ważne jest testowanie, jak nowa strategia, oparta na nowym modelu, zmieni akceptowany portfel. Jeśli będzie to zbyt duża zmiana, nikt nie da gwarancji sukcesu. Metody uwzględniania wpływu wniosków odrzuconych, ang. Reject Inference (Huang, 2007; Anderson *et al.*, 2009; Hand i Henley, 1994; Verstraeten i den Poel, 2005; Finlay, 2010; Banasik i Crook, 2003, 2005, 2007), są już znane od wielu lat. Jednak ma się wrażenie, że dopiero wykorzystanie danych symulacyjnych spowoduje ich dalszy rozwój.

3.4. Zewnętrzne bazy minimalizujące wpływ wniosków odrzuconych

Jednym z lepszych narzędzi, które pomagają minimalizować problem wpływu wniosków odrzuconych, jest korzystanie z zewnętrznych baz zbierających informacje o kredytach, długach i negatyw-

nych informacjach dla klientów rynku polskiego. Są to między innymi: Biura Informacji Gospodarczej (BIG), takie jak: BIG InfoMonitor, Krajowy Rejestr Długów (KRD) i Rejestr Dłużników ERIF. Wygodne także mogą być bazy: Międzybankowa Informacja Gospodarcza – Dokumenty Zastrzeżone (MIG–DZ) oraz Bankowy Rejestr (MIG–BR).

Najpotężniejszą bazą danych, zawierającą informacje o kredytach, zarówno pozytywne, jak i negatywne, jest Biuro Informacji Kredytowej (BIK).

Brak pełniej historii kredytowej klientów w BIK jest bardzo groźnym w skutkach problemem. Istnienie parabanków w Polsce i ich obecna ekspansja może spowodować poważne braki w danych, a tym samym w poprawnej estymacji ryzyka, jeśli nie będą one raportować swoich danych kredytowych do BIK. Wszelkie rozważania w książce wyraźnie wskazują, że wpływ wniosków odrzuconych nie jest do końca przewidywalny. Dane gromadzone w BIK, jeśli reprezentują cały rynek bankowy, to zaczynają być podobne do rajskiej strategii, czyli pełnej akceptacji całego rynku. Modele PD budowane na takich danych nie zmieniają tak szybko mocy predykcyjnej, ich wdrożenie do systemu nie powoduje istotnych zmian w populacji.

Na podstawie przeprowadzonych analiz można śmiało sformułować dwie rekomendacje:

1. Jeśli to tylko możliwe, należy wykorzystywać dane z BIK do budowy własnych modeli.

2. Należy z całą rzetelnością zadbać o jakość danych w BIK, by zawierały pełną informację o rynku kredytowym w Polsce. Troskę tę powinny wyrazić instytucje organu państwa, gdyż niekoniecznie jest to interes prywatnych spółek. W szczególności duże banki mogą być tym mniej zainteresowane, gdyż, raportując do BIK, pomagają konkurencji poprawnie estymować ryzyko detaliczne.

4. Budowa modelu aplikacyjnego

W rozdziale przedstawiona jest budowa modelu aplikacyjnego we wszystkich najważniejszych szczegółach. Ze względu na dane symulacyjne wszystkie pojawiające się po drodze wyniki mogą być bez problemu prezentowane i tym samym rozdział ten staje się specyficzny, gdyż z reguły dokumentacja modelu jest jedną z największych tajemnic strzeżonych przez bank. Wszystkie prezentowane wyniki są efektem przetwarzania i analizowania danych w Systemie SAS 9.3 przez własnoręcznie napisane przez autora kody w języku SAS 4GL. Kody te stanowią spójną całość i pozwalają w prosty i wygodny sposób wykonać wszystkie etapy budowy modelu.

Zanim powstanie model, warto określić w banku, lub innej instytucji finansowej, cykl życia modelu. Jest to o tyle ważne, że porządkuje proces i pomaga gromadzić wszystkie ważne dokumenty.

- Wniosek: Klient wewnętrzny lub zewnętrzny, zwany zamawiającym, powinien złożyć wniosek o budowę modelu. Model buduje się przez pewien czas (od kilku tygodni do kilku miesięcy), zatem taki wniosek uzasadnia wykonanie pierwszej pracy. We wniosku powinny być określone zarówno przyczyny budowy nowego modelu, jak i wymagania, czyli czego oczekuje biznes od jego budowy. Powinny być też określone podstawowe role: kto jest właścicielem modelu, kto będzie budował, kto walidował, kto wdrażał i kto odbierał cały projekt. Dobrze jest, aby wniosek ten był formalnym dokumentem z odpowiednimi podpisami.

- Budowa modelu: Wszystkie etapy przedstawione są w tym rozdziale. Należy tu dodać, że wiele z nich może kończyć się także pośrednim dokumentem lub spotkaniem z zamawiającym w celu przedstawienia już wykonanych prac i uzyskania akceptacji. Zamawiający może czasem mieć wpływ na proces budowy. Może np. nie zgodzić się na pewną zmienną, gdyż z jego doświadczenia wynika, że zmienna jest źle wprowadzana do systemu. Nawet jeśli posiada ona dobre własności predykcyjne, może być zanegowana. Ustalenie, na ile zamawiają-

cy może się wtrącać, powinno być dobrze określone, aby obie strony mogły wykonać swoją pracę. Etap ten musi się skończyć dokumentacją.

- Walidacja: Ten etap jest szczególnie ważny ze względu na rekomendacje i wytyczne Basel II/III. Jest to etap, gdzie zespół budujących model wyjaśnia i broni swoich racji przed osobą zatwierdzającą. Osoba ta oczywiście może uczestniczyć w budowie modelu od początku i podobnie jak zamawiający może wyrażać swoje opinie i sugestie. Etap ten też musi zakończyć się dokumentem. Możliwe jest także, że niektóre elementy trzeba powtórzyć, gdyż osoba zatwierdzająca może np. kwestionować niektóre zmienne, które nie mają sensownej interpretacji. Tego typu wymaganie jest bardzo ważne, gdyż inaczej nie przekonamy szerokiego grona w firmie do użycia tego modelu.

- Decyzja o wdrożeniu: Jeśli wszystko przebiegło pomyślnie, model może być zaakceptowany przez podjęcie stosownej decyzji. Musi być jednoznacznie określone, kto taką decyzję podejmuje i w jaki sposób.

- Wdrożenie: Musi być sporządzony dokument z tego etapu, powinien zawierać różnego rodzaju testy, w szczególności UAT (ang. User Acceptance Tests). Jest możliwe, że podczas wdrożenia niektóre wzory, czy reguły muszą być lekko zmodyfikowane, ze względu na specyfikę systemu. Tego typu zmiany muszą być staranie udokumentowane i czasem powinny przejść dodatkowy etap walidacji.

- Monitoring: Od wdrożenia regularnie wykonywany jest monitoring modelu. Oczywiście z reguły w pierwszych miesiącach po wdrożeniu wykonuje się wiele dodatkowych testów, a potem ustala się standardowy ich zestaw i wykonuje cyklicznie. Każdy cykl raportowy powinien być przyjęty przez właściwą osobę zatwierdzającą, gdyż w pewnym momencie życia modelu podejmuje się decyzję o sprawdzeniu możliwości budowy nowego lub zamianie na inny. Wtedy proces zaczyna się od początku.

4.1. Analiza aktualnego procesu i przygotowanie danych

Przypuśćmy, że nasz model budowany jest na bazie procesu produkcyjnego związanego ze strategią 2; patrz tabela 13. Należy zbudować model akceptacyjny dla kredytu ratalnego. Obecny proces akceptuje $60{,}5\%$ wniosków. Ryzyko akceptowanych jest na poziomie $3{,}9\%$, natomiast odrzuconych – $26{,}9\%$. Innymi słowy 40% populacji przychodzącej generuje bardzo duże ryzyko. Tego ryzyka oczywiście w rzeczywistości nie jesteśmy w stanie zmierzyć. Rzeczywisty proces dostarcza nam tylko procent akceptacji i ryzyko zaakceptowanych. Dodatkowo w obecnym procesie wdrożony jest model PD Ins, który na części zaakceptowanej posiada moc $41{,}7\%$. Z reguły nigdy nie ma etapu w systemie produkcyjnym bez jakiegokolwiek modelu. Albo model jest zakupiony od firmy zewnętrznej, przy starcie produktu, albo jest to już model budowany przez zespół w banku i model ten po prostu się dewaluuje. Wartość mocy $41{,}7\%$ jest na granicy decyzji, niestety wartości $69{,}04\%$ (dla całej populacji) w rzeczywistości nie da się zaobserwować, więc bank może podjąć decyzję o budowie nowego. Model dla kredytów ratalnych używany jest głównie w celu minimalizacji straty, by nie udzielać kredytów klientom, którzy już z tym kredytem wpadną w zadłużenia i nie powinni korzystać później z kolejnych. Jedyne zatem kryterium, które można określić, to aby model nie miał gorszej mocy predykcyjnej od aktualnego, czyli od $41{,}7\%$. Możemy też przypuścić, że aktualny model nie może być dłużej stosowany ze względu na specyficzne uwarunkowania licencyjne, albo ze względu na zmienną modelową, z której bank postanowił zrezygnować ze względu na rezygnacje klientów zniecierpliwionych zbyt długim procesem akceptacji. Z drugiej strony model PD Ins jest zbudowany na danych pełnej strategii i posiada moc $73{,}37\%$ na zbiorze treningowym; patrz tabela 44. Jest modelem, który nie musiał uwzględniać wniosków odrzuconych, czyli w momencie modelowania była dostępna pełna informacja o klientach. Taki model może być rozważony tylko na danych symulacyjnych i może być traktowany jako rodzaj ideału, do którego inne modele mogą tylko dążyć, ale nigdy go nie osiągną.

Ze względu na dość małą liczbę zaakceptowanych wniosków, około 14 tysięcy, w okresie modelowym 1975–1987, rozszerzono okres danych, aby uzyskać pożądane liczby obserwacji w zbiorach walidacyjnym i treningowym.

4.1.1. Definicja zdarzenia default

Definicja pojawiła się już w podrozdziale 1.3.10. Wszystkie modelowania opisane w książce robione są na definicji zdarzenia niewywiązania się z zobowiązań (ang. default), czyli wejścia w zadłużenie więcej niż 90 dni (90+) w okresie 12 miesięcy. Natomiast w zależności od potrzeby raz używana jest dwuwartościowa kolumna $\mathrm{default}_{12}^{GB}$ tylko ze statusami **Dobry** lub **Zły**, a raz z trzema z dodatkowym statusem **Nieokreślony** – kolumna $\mathrm{default}_{12}$. Z reguły modele buduje się, by odróżnić istotnie złych od istotnie dobrych, czyli pomija się nieokreślonych. Ryzyko jednak liczy się jako udział złych, czyli iloraz złych do sumy dobrych, złych i nieokreślonych. Przy kalibracji zatem używa się specjalnej zmiennej, gdzie nieokreślony jest oznaczony jako dobry. Ze względu na dalsze etapy budowy modelu i upraszczając modelowanie lepiej jest stan nieokreślony uznać za dobry i dalej postępować wyłącznie z dwiema wartościami zmiennej celu. W innym przypadku trzeba byłoby oddzielnie prognozować prawdopodobieństwo bycia nieokreślonym na zbiorze odrzuconych. Innymi słowy wszelkie modele budowane w tym rozdziale związane są z funkcją celu $\mathrm{default}_{12}^{GB}$. Można też wybrać krótszy okres obserwacji, nie 12 miesięcy tylko 9 lub nawet 6. Zależne jest to od rozkładu zdarzeń względem okresu obserwacji. Jeśli większość zdarzeń pojawia się w 12 miesiącach, to lepiej wybrać 12. Może też być to niemożliwe ze względu na zbyt krótką historię danych. Czasem lepiej pracować z 6 miesiącami, gdyż model bardziej koncentruje się na najświeższych przypadkach.

4.1.2. Dostępne dane

Dane dla wniosków zaakceptowanych w przedziałach pięcioletnich zostały przedstawione w tabeli 18, a dla wszystkich przychodzących w tabeli 19. Nowy model ma zastąpić stary, zatem musi umieć rozróżniać dobrych od złych klientów na całej populacji przychodzącej,

a nie tylko na zaakceptowanej przez stary model. To jest największą trudnością w budowie modelu akceptacyjnego, by umieć wyciągnąć wnioski z tego, co jest dane i zastosować także na tym, co dane nie jest. Dlatego najbardziej znaną metodą jest budowanie dwóch modeli. Pierwszy zwany jest KGB (ang. known good bad). Na jego bazie buduje się drugi na całej populacji (ang. ALL). Często przy analizie dostępnych danych dokonuje się przeglądu wszystkich reguł procesu akceptacji. W typowym procesie istnieją zarówno reguły oparte na modelach skoringowych, jak i twarde, związane z różnymi bazami nierzetelnych klientów (ang. black list) i wieloma innymi. Sytuacja może być podobna do strategii 3 i 4; patrz tabele 14 i 15. Decyzja o usunięciu jakiejś reguły może także bardzo zakłócić wnioskowanie statystyczne obarczone wnioskami odrzuconymi. Najlepiej, jak usuwa się regułę po regule i testuje w produkcyjnym środowisku, np. w formie champion challengera. Kolejnym bardzo ważnym tematem związanym z przygotowaniem danych jest wybranie jak najlepszego układu danych historycznych, aby był zbliżony do danych produkcyjnych, na których skoring będzie używany. Często usuwa się z historii wiele przypadków, gdyż ze względu na nowe strategie nie powinny się już zdarzać w przyszłości.

Tabela 18: Dane modelowe dla zaakceptowanych wniosków

Pięciolatka	Razem	Liczba wniosków			Procent		
		Dobry	Nieokreślony	Zły	Dobry	Nieokreślony	Zły
1975	5 190	4 628	277	285	89,17	5,34	5,49
1980	5 613	5 270	169	174	93,89	3,01	3,10
1985	5 861	5 530	162	169	94,35	2,76	2,88
1990	6 108	5 407	319	382	88,52	5,22	6,25
1995	4 741	4 146	278	317	87,45	5,86	6,69
Razem	27 513	24 981	1 205	1 327	90,80	4,38	4,82

Źródło: opracowanie własne.

4.1.3. Próby losowe

Modelowanie na podstawie modeli regresyjnych, w szczególności regresji logistycznej, wymaga stworzenia minimalnie dwóch zbio-

Tabela 19: Dane modelowe wszystkich wniosków

Pięciolatka	Razem	Liczba wniosków			Procent		
		Dobry	Nieokreślony	Zły	Dobry	Nieokreślony	Zły
1975	9 122	7 014	609	1 499	76,89	6,68	16,43
1980	9 108	7 602	467	1 039	83,47	5,13	11,41
1985	9 053	7 668	414	971	84,70	4,57	10,73
1990	9 139	7 152	611	1 376	78,26	6,69	15,06
1995	7 177	5 482	526	1 169	76,38	7,33	16,29
Razem	43 599	34 918	2 627	6 054	80,09	6,03	13,89

Źródło: opracowanie własne.

rów: treningowego i walidacyjnego. Inne modele, takie jak drzewa decyzyjne czy sieci neuronowe, często do poprawności wykrycia właściwych reguł predykcyjnych potrzebują także zbioru testującego. Dzielić można różnymi metodami, najbardziej znane to: zwykłe losowe (ang. random sampling), cały zbiór modelowy dzielimy losowo, np. w proporcjach 60% / 40%. Takie modelowanie sprzyja budowie modeli na dany czas, czy też dopasowanych do aktualnej historii (ang. point in time). Drugi sposób to czasowe próbkowanie (ang. time sampling), gdzie zbiór walidacyjny reprezentuje starszą paczkę danych niż treningowy, co umożliwia tworzenie modeli bardziej stabilnych w czasie dla całego cyklu koniunkturalnego (ang. through the cycle). W naszym przypadku dla budowanego modelu wybrano zwykłe losowanie. Zbiór treningowy ma 16 414 obserwacji, a walidacyjny 11 099.

Cały proces przygotowania danych, włączając definicję zdarzenia default, przegląd reguł akceptacji, wybór historii podobnej do danych obecnych i przyszłych oraz próby losowe, wspólnie można nazwać bazą modelową. Czas poświęcony bazie jest bardzo ważny i nie należy go bagatelizować. Ten etap w dobie ery Big Data może być zaniedbany, a w konsekwencji może prowadzić do złych decyzji i strat finansowych. Nie da się tego etapu wykonać automatycznie i analityk po prostu musi dokładnie go przemyśleć, przeanalizować i często przedyskutować, konfrontując swoje założenia z wieloma środowiskami.

4.2. Budowa modelu KGB dla zaakceptowanych

Budowa modelu aplikacyjnego wymaga wykonania kilku modeli pośrednich. Wynika to z prostego faktu, że model aplikacyjny, inaczej akceptacyjny, musi umieć dyskryminować klientów na całej populacji przychodzącej (aplikującej o kredyty), a nie tylko na aktualnie akceptowanej. Z drugiej strony jedyne obserwowane informacje o spłacalności kredytów pochodzą tylko z zaakceptowanej części. Dlatego pierwszym modelem jest karta skoringowa budowana tylko na części akceptowanej przez stary model. Jak już wspomniano wcześniej, model ten często oznacza się jako KGB.

4.2.1. Tworzenie kategorii zmiennych lub grupowanie (ang. binning)

Każda zmienna ciągła czy nominalna, jest transformowana do nominalnej lub porządkowej, powstałej przez grupowanie zmiennej jakościowej lub kategoryzację ciągłej. Najczęściej wykorzystuje się tu algorytmy drzew decyzyjnych (Krzyśko *et al.*, 2008) oparte na statystyce Entropii lub indeksie Giniego. Dość istotnym problemem jest tu ustalenie, na ile kategorii ma być podzielona przestrzeń wartości i na ile kategorie te są reprezentatywne. W naszym przypadku definiuje się to przez proste dwa parametry: minimalny udział kategorii to 3% i maksymalna liczba kategorii to 6.

Na początku, z reguły ze względu na dość dużą liczbę charakterystyk, która w dobie Big Data może się jeszcze znacząco powiększać, uruchamia się automatyczne procesy, które tworzą kategorie, głównie koncentrując się na celu statystycznym, maksymalizacji mocy predykcyjnej. Innymi słowy dla zmiennych ciągłych szukane są takie punkty podziałowe, które znacząco różnicują wartości ryzyka na kategoriach.

4.2.2. Wstępna selekcja zmiennych (preselekcja)

Wszystkich charakterystyk na początku procesu jest 201. Po automatycznym tworzeniu kategorii i policzeniu pierwszej statystyki, współczynnika Giniego, liczącego moc predykcyjną na zbiorze treningowym oraz ustaleniu punktu odcięcia powyżej 5% pozostało tylko 18

zmiennych. Akurat dla tego modelu tak mało zmiennych udało się uzyskać przy tylko jednym kryterium. Wynika to z dwóch powodów: założenia danych były tak dobierane, że ryzyko kredytów ratalnych nie było bardzo zależne od historycznych kredytów, w szczególności od kredytów gotówkowych, stąd mnóstwo zmiennych behawioralnych posiada Giniego poniżej 5%. Drugi powód to fakt, że wiele zmiennych behawioralnych zbiera raczej negatywną informację niż pozytywną. Jest to ogólnie bardzo ciekawy problem do badań naukowych: jak przygotowywać dobre zmienne? Zmienne, takie jak: ile razy klient był opóźniony, ile miał źle spłacanych kredytów itp., są bardzo predyktywne, ale w portfelu, w którym mamy dużo złych klientów i dużo dobrych. Wtedy owe zmienne potrafią odróżnić jednych od drugich. W naszym przypadku mamy jednak model budowany na części zaakceptowanej stanowiącej tylko $60{,}5\%$ całej populacji wnioskującej i tu już są klienci lepsi z ryzykiem na poziomie $3{,}9\%$ w odróżnieniu od odrzuconych z ryzykiem $26{,}9\%$, co powoduje, że zmienne behawioralne raczej nie mogą już bardzo pomóc, bo większość klientów zaakceptowanych poprzednie kredyty miała raczej w dobrych stanach. Jest to już pierwszy niepokojący sygnał informujący o wpływie wniosków odrzuconych, gdyż z modelu na zaakceptowanych nie uda się wyprowadzić własności klientów odrzucanych pod kątem ich ryzyka.

Dokładamy kilka kolejnych kryteriów dla zmiennych. Głównie chodzi tu o pomiar stabilności zmiennych. W tym wypadku stabilność rozumiemy jako porównywalne własności pomiędzy zbiorem treningowym a walidacyjnym. Pierwszą statystyką mierzącą stabilność jest $\mathrm{AR}_{\mathrm{diff}}$ (inaczej delta Gini), różnica względna pomiędzy statystyką Giniego na wymienionych zbiorach:

$$\mathrm{AR}_{\mathrm{diff}} = \frac{\mathrm{ABS}(\mathrm{AR}_{\mathrm{Train}} - \mathrm{AR}_{\mathrm{Valid}})}{\mathrm{AR}_{\mathrm{Train}}},$$

gdzie funkcja $\mathrm{ABS}()$ jest wartością bezwzględną.

Kolejną statystyką jest odległość Kullbacka–Leiblera (BIS–WP14, 2005):

$$\mathrm{KL} = \sum_i t_i \ln\left(\frac{t_i}{v_i}\right),$$

gdzie t_i i v_i są udziałami i-tej kategorii odpowiednio w zbiorze tre-

ningowym i walidacyjnym. Statystykę tę możemy też obliczyć, rozważając tylko rozkłady złych (ang. Bad) klientów, tworząc w ten sposób nową KLB, żeby mierzyć stabilność ogólnych rozkładów i ryzyka na kategoriach. Określamy zatem kryteria: $\mathrm{AR}_{\mathrm{diff}} < 20\%$, $\mathrm{KL} < 0{,}1$ i $\mathrm{KLB} < 0{,}1$. Pozostaje tylko 7 zmiennych; patrz tabela 20. Wszystkie wymienione kryteria preselekcji zmiennych mogą być określane na różne sposoby. Jest to pewnego rodzaju kwestia wiedzy eksperckiej analityka i danych, z którymi się pracuje. Inne pomocne statystyki często używane jako kryteria to: KS – Kolmogorov–Smirnov, IS – Index stability, IV – Information value, wszystkie opisane w (BIS–WP14, 2005).

Tabela 20: Wybrane zmienne z etapu preselekcji do dalszego modelowania

Zmienna	$\mathrm{AR}_{\mathrm{Train}}$ (%)
ACT_CINS_N_STATB	21,3
APP_CHAR_JOB_CODE	20,4
APP_CHAR_GENDER	16,8
APP_INCOME	14,9
ACT_CINS_N_STATC	13,4
ACT_CINS_SENIORITY	12,9
APP_SPENDINGS	11,6

Źródło: opracowanie własne.

4.2.3. Dalsza selekcja zmiennych, ręczne poprawki kategorii

W badaniu tworzonych kategorii nie wolno zapomnieć o stabilności w czasie. Powinny być zarówno stabilne rozkłady, czyli udziały kategorii w czasie, jak i wartości ryzyka dla poszczególnych kategorii. Brak stabilności może powodować kolejne niestabilności już na poziomie wskaźników finansowych banku. Niestabilność rozkładów może zaburzyć proces akceptacji, utrudni to bardzo pracę sprzedawców, gdyż z reguły godzą się oni na pewien procent odrzutu, ale są bardzo czuli na jego wzrosty ze zrozumiałych powodów. Brak stabilności ryzyka nie wpłynie od razu na wskaźniki sprzedaży, ale po

pewnym czasie na bilans banku i na opłacalność procesu. W skrajnym przypadku może po czasie być podjęta decyzja o istotnej zmianie procesu, czyli typowo o zmniejszeniu procentu akceptacji. Jest to często mechanizm wpadnięcia w pętlę błędnych decyzji, gdyż niestabilny proces zniechęca sprzedawców i samych klientów, wnioskują zatem klienci mocniej zdeterminowani sytuacją życiową, czyli bardziej ryzykowni, a to powoduje zwiększenie ryzyka, co z kolei wymusza zmniejszenie procentu akceptacji, by nadal utrzymać odpowiedni udział złych długów. Mamy tu przykład, że Credit Scoring to także pewne elementy psychologii tłumów, co także musi być w pewien sposób rozważone podczas dobierania parametrów strategii. Trzeba rozumieć zachowania i preferencje danego społeczeństwa, zanim zacznie się interpretować dane i modelować procesy.

Tabela 21: Raport kategorii zmiennej ACT_CINS_SENIORITY – liczby miesięcy od pierwszego kredytu ratalnego

Numer	Warunek (definicja kategorii)	Liczba	Procent	Ryzyko (%)
1	Missing	11 564	70,5	5,6
2	167 < ACT_CINS_SENIORITY $\leqslant$ 227	1 165	7,1	4,1
3	ACT_CINS_SENIORITY $\leqslant$ 96	848	5,2	2,9
4	227 < ACT_CINS_SENIORITY	1 427	8,7	2,8
5	96 < ACT_CINS_SENIORITY $\leqslant$ 132	710	4,3	2,0
6	132 < ACT_CINS_SENIORITY $\leqslant$ 167	700	4,3	1,1

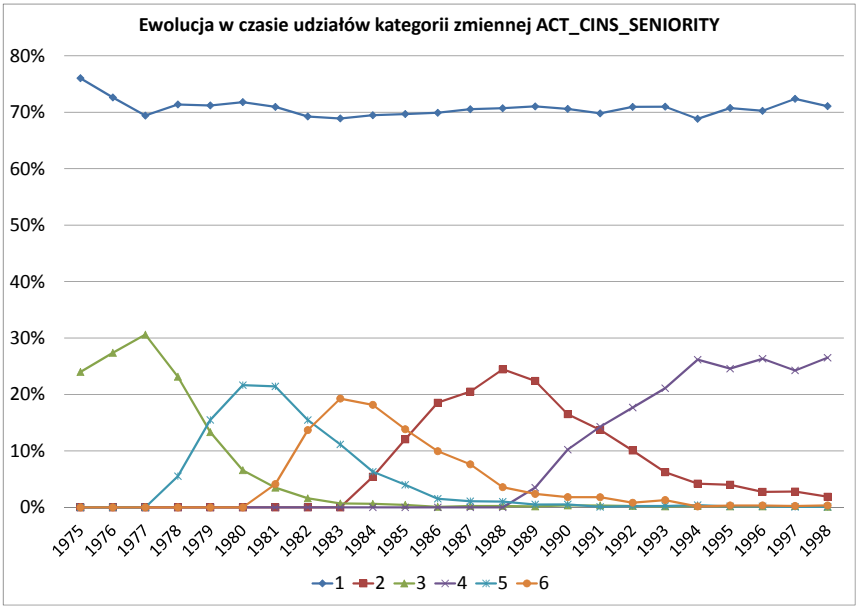
Źródło: opracowanie własne.

Stabilność zmiennych wstępnie jest badana na etapie preselekcji, gdzie sprawdza się własności na zbiorach treningowym i walidacyjnym. Kiedy jednak lista zmiennych jest już mniejsza, możliwe są głębsze studia każdej z nich. Na tym polega właśnie modelowanie, że pewne etapy są wykonywane automatycznie, a niektóre wymagają pracy analityka i ręcznej obróbki.

Pierwszą analizą może być raport kategorii, ich udziałów i wartości ryzyka, patrz tabela 21. Niestety nie mamy tu typowej monotoniczności zmiennej, czyli że wraz ze wzrostem liczby miesięcy niekoniecznie maleje lub wzrasta ryzyko. Możemy ocenić, że klient z brakiem historii kredytowej jest najbardziej ryzykowny, a klient

z dłuższą – znacząco mniej, ale widać wyraźnie pewne zakłócenia monotoniczności. Raport ten nie jest dokładny, gdyż nie ujmuje ewolucji w czasie. Na rysunku 17 przedstawiono raport bardziej szczegółowy, dzięki któremu możemy się dowiedzieć, że udziały kategorii nie są stabilne w czasie. Wymaga to poprawienia ręcznego. Raport ewolucji w czasie wykonuje się także dla wartości ryzyka, który pozwala upewnić się, czy ryzyko nie zmienia swojego porządku w czasie. Zdarza się np., że w jednym okresie kategoria 5 ma najmniejsze ryzyko, a w innym kategoria 6. Wtedy obie te kategorie się łączy w jedną, stabilniejszą w czasie. Można stworzyć algorytm automatyczny poprawiający stabilności poprzez łączenie kategorii, ale nie jest on wcale taki prosty. Można też wykazać, że monotoniczność jest cechą gwarantującą większą stabilność zmiennych.

Rysunek 17: Ewolucja w czasie udziałów kategorii



Źródło: opracowanie własne.

Po wielu próbach okazało się, że zmienną ACT_CINS_SENIORITY można podzielić tylko na dwie kategorie: pusta historia kre-

dytowa i niepusta. Wtedy zmienna ta uzyskała stabilność w czasie. Niestety nawet po poprawkach zmienna ta finalnie nie znalazła się w modelu, gdyż przestała być istotna.

4.2.4. Estymacja modelu, metoda LOG, liczenie oceny punktowej

Karta skoringowa budowana jest metodą LOG, która w literaturze znana jest pod nazwą WoE z ang. Weight of Evidence (Siddiqi, 2005). Standardowo dla każdej kategorii zmiennej oblicza się statystykę WoE. Jest to miernik bardzo podobny do logitu. Zachodzi następujące prawo:

$$\mathrm{WoE}_k = \ln\left(\frac{G_k/G}{B_k/B}\right) =$$

$$= \ln\left(\frac{G_k}{B_k}\right) - \ln\left(\frac{G}{B}\right),$$

czyli:

$$\mathrm{WoE}_k = \mathrm{Logit}_k - \mathrm{Logit},$$

gdzie k – oznacza dowolną kategorię zmiennej, G, B – liczby dobrych i złych klientów w całej populacji, a G_k i B_k w kategorii. Mamy zatem zależność, że Weight of Evidence dla kategorii jest różnicą logitu kategorii i logitu całej populacji. Dlatego w dalszej części nazywamy metodę budowy modelu LOG i wyliczamy logity zamiast WoE.

Każda zmienna wybrana do modelu jest transformowana do kawałkami stałej na podstawie obliczonych logitów każdej z jej kategorii. Estymacja regresji logistycznej w ogólnym zapisie określona jest wzorem:

$$\mathrm{Logit}(p_n) = X_n\beta,$$

gdzie p_n jest prawdopodobieństwem, że klient jest dobry, inaczej $p_n = P(Y = \mathbf{Dobry})$ dla n-tej obserwacji, a β reprezentuje wektor

współczynników regresji. Macierz X_n można szczegółowo rozpisać w następujący sposób:

$$X_n = l_{ij}\delta_{ijn},$$

gdzie l_{ij} jest logitem j-tej kategorii i-tej zmiennej a δ_{ijn} jest macierzą zerojedynkową przyjmującą wartość jeden, gdy n-ta obserwacja należy do j-tej kategorii i-tej zmiennej. Dodatkowo przyjęto uproszczone założenie, że każda zmienna ma tyle samo kategorii, aby nie wprowadzać większej liczby indeksów oraz że liczba kategorii jest taka sama jak liczba zmiennych i wynosi v.

Zauważmy dodatkowo, że w przypadku tylko jednej zmiennej w modelu, mamy dość prostą sytuację, przy której współczynnik $\beta = 1$. Wynika to z faktu, że po obu stronach równania stoją te same logity. Oznacza to także, że założenie o liniowej zależności funkcji celu i predyktorów jest z góry spełnione. W przypadku wielu zmiennych w modelu współczynniki powinny być wszystkie dodatnie, gdyż taką monotoniczność mają logity oraz numery kategorii zmiennych. Jeśli numer kategorii zmiennej rośnie, to ryzyko maleje. Tak zdefiniowane są wszystkie zmienne i ich logity na kategoriach. Jeśli zatem dochodzi do zmiany znaku współczynnika regresji, to mamy do czynienia z silną współliniowością lub przypadkiem zmiennej zakłóceniowej (ang. confounding variable). Zliczanie ujemnych współczynników jest zatem jedną z metod detekcji współliniowości.

Iloczyn macierzy X i wektora β stojący po prawej stronie równania regresji jest wartością oceny punktowej (ang. score) dla danej obserwacji. Ocena ta nie jest skalibrowana i ciężko ją interpretować. Zwykle wykonuje się kilka prostych przekształceń, aby nadać jej bardziej użyteczną formę. Zauważmy, że jeśli wartość prawdopodobieństwa p_n rośnie, to jego logit także, a zatem ocena punktowa również będzie rosnąć. Czyli im większa ocena punktowa, tym większe prawdopodobieństwo spłacenia kredytu. Najczęściej kalibruje się wartość oceny punktowej poprzez prostą funkcję liniową:

$$\mathrm{Logit}(p_n) = \ln\left(\frac{p_n}{1-p_n}\right) = S_n = aS_n^{\mathrm{New}} + b,$$

gdzie S_n^{New} jest nową oceną, a S_n starą, natomiast a i b są współczynnikami. Wyznacza się je tak, aby uzyskać dodatkową własność,

którą w książce definiuje się w następujący sposób: dla wartości 300 punktów szansa (ang. odds) bycia dobrym klientem powinna wynosić 50, a gdy szansa zwiększy się dwukrotnie (ang. double to odds), czyli będzie wynosić 100, to ocena powinna być równa 320. Szansę definiuje się jako iloraz liczby dobrych do złych klientów, lub jako stosunek $\frac{p_n}{1-p_n}$. Szansa 50 reprezentuje zatem segment klientów, gdzie na jednego złego przypada 50 dobrych klientów. Należy zatem rozwiązać układ równań (Siddiqi, 2005):

$$\ln(50) = a\ 300 + b,$$

$$\ln(100) = a\ 320 + b.$$

Jego rozwiązaniem są wartości:

$$a = \frac{\ln\left(\frac{100}{50}\right)}{20} = \frac{\ln(2)}{20},$$

$$b = \ln(50) - \frac{300 \ln\left(\frac{100}{50}\right)}{20} = \ln\left(\frac{50}{2^{15}}\right).$$

Drugą czynnością skalującą wartość oceny punktowej jest zadbanie, by wszystkie oceny cząstkowe pierwszej kategorii miały taką samą liczbę punktów. Pierwsza kategoria reprezentowana jest przez grupę najbardziej ryzykownych klientów. Ostatnia reprezentuje najlepszych. Jeśli oceny cząstkowe zaczynają się zawsze od tej samej wartości, to ta zmienna, która ma największą wartość oceny cząstkowej, może być interpretowana jako najważniejsza w modelu.

Mamy dalej:

$$S_n = \sum_{i,j=1}^{v} \beta_i l_{ij} \delta_{ijn} + \beta_0.$$

Możemy wydzielić człon związany z najgorszym klientem:

$$\gamma = \sum_{i=1}^{v} \beta_i l_{i1},$$

i dzięki temu wyraz wolny rozdzielić na dwa składniki:

$$\beta_0 = \sum_{i=1}^{v} \frac{\beta_0 + \gamma}{v} - \sum_{i=1}^{v} \beta_i l_{i1}.$$

W ten sposób powstaje ocena cząstkowa (ang. partial score):

$$P_{ij} = \beta_i l_{ij} + \frac{\beta_0 + \gamma}{v} - \beta_i l_{i1}.$$

Zauważmy, że dla każdej zmiennej i mamy:

$$P_{i1} = \frac{\beta_0 + \gamma}{v},$$

czyli oceny cząstkowe zaczynają się od tej samej wartości.

Mamy dalej:

$$S_n = \sum_{i,j=1}^{v} P_{ij}\delta_{ijn},$$

oraz finalnie:

$$S_n^{\text{New}} = \frac{S_n - b}{a} = \sum_{i,j=1}^{v} P_{ij}^{\text{New}}\delta_{ijn},$$

gdzie

$$P_{ij}^{\text{New}} = \frac{1}{a}P_{ij} - \frac{b}{v}.$$

Ostateczną wartość oceny cząstkowej często zaokrągla się do najbliższej wartości całkowitej. W ten sposób otrzymuje się kartę skoringową z naliczonymi punktami przy każdej z kategorii ze zmiennych wybranych do modelu.

W efekcie można mówić o dwóch rodzajach kalibracji. Pierwszej, transformującej wartość prawdopodobieństwa p_n do oceny punktowej, gdzie bierze udział przekształcenie logitowe oraz z oceny punktowej do prawdopodobieństwa, gdzie używa się funkcji odwrotnej do logitu postaci:

$$p_n = \frac{1}{1 + e^{-(\omega_s S_n^{\text{New}} + \omega_0)}},$$

gdzie ω_s i ω_0 są współczynnikami.

4.2.5. Finalna postać modelu KGB

W tym wypadku ze względu na małą liczbę zmiennych nie były konieczne zaawansowane metody selekcji zmiennych, gdyż od razu otrzymaliśmy finalny model KGB; patrz tabela 22.

Wszystkie karty skoringowe prezentowane w książce są tak budowane, że dość łatwo można rozpoznać, która zmienna jest najistotniejsza w modelu. Mianowicie zmienne można uporządkować po maksymalnej ocenie cząstkowej. Zmienna z najwyższą wartością jest najważniejsza w modelu (porównaj podrozdział 4.2.4).

Podane mierniki modelu to, oprócz Giniego (na zbiorze treningowm i walidacyjnym), także wygodne statystyki LiftX, które informują, ile razy model jest lepszy od losowego w wyborze modelowanego zdarzenia na podzbiorze $X\%$ (czyli na danym percentylu), wybranym jako obserwacje z największym prawdopodobieństwem modelowanego zdarzenia. Statystyki te są szczególnie ważne przy modelach akceptacyjnych, gdzie właśnie chodzi o ustalenie odpowiedniego punktu odcięcia, by odrzucić jak najwięcej niepożądanych wniosków.

Tabela 22: Model KGB

Gini w % na zbiorze		Lift1	Lift5	Lift10	Lift20
treningowym	walidacyjnym				
37,92	33,58	9,72	4,14	2,77	2,44

Karta skoringowa

Zmienna	Warunek	Ocena cząstkowa
ACT_CINS_N_STATB	$0 <$ ACT_CINS_N_STATB	86
	Missing	102
	ACT_CINS_N_STATB $\leqslant 0$	138
APP_CHAR_GENDER	Male	86
	Female	105
APP_CHAR_JOB_CODE	Owner company	86
	Permanent	90
	Retired	103
APP_SPENDINGS	$960 <$ APP_SPENDINGS	86
	APP_SPENDINGS $\leqslant 960$	94

Źródło: opracowanie własne.

4.3. Estymacja ryzyka odrzuconych wniosków

Etap modelowania związany z estymacją ryzyka odrzuconych wniosków jest jednym z najtrudniejszych w całym procesie budowy modelu. Nie chodzi tu o skomplikowane wzory, czy teorie, ale o świadomość błędu, jaki może być popełniony. Jeśli udział odrzuconych stanowi istotną część portfela przychodzącego, czy wnioskującego, to błąd może być rzędu kilkuset procent, a tym samym możemy bank narazić na bardzo poważne straty.

4.3.1. Porównanie modeli: KGB i PD Ins

Na wstępie oceńmy poprawność dyskryminowania nowego modelu, mierząc jego moc także na części odrzuconej. Ponieważ wszelkie analizy wykonujemy na danych symulacyjnych, to znamy też statusy zmiennej $\mathrm{default}_{12}^{GB}$ także dla wniosków odrzuconych. Dodatkowo obliczymy też moce predykcyjne na zmiennej $\mathrm{default}_{12}$; patrz tabela 23. Porównanie obu modeli i mocy predykcyjnych na dwóch różnych definicjach default daje ważne informacje. Po pierwsze widać wyraźnie, że moc modelu spada, jeśli za dobrego uznaje się także nieokreślonego. Istnienie nieokreślonych, szczególnie przy ich dużym udziale, sprawia większy problem w uzyskaniu zadowalających wskaźników predykcyjności. Zatem powinno się ich uwzględniać w próbach modelowych i stosować definicję $\mathrm{default}_{12}^{GB}$, gdyż takie ujęcie bezpośrednio przekłada się potem na estymację ryzyka. Model stary, PD Ins, na odrzuconych wnioskach zdecydowanie zachowuje się lepiej niż nasz nowy KGB. Oznacza to, że nie możemy uważać, że profil odrzuconego klienta jest podobny do zaakceptowanego. Model estymowany na bazie próby zaakceptowanych nie potrafi dobrze rozróżniać klientów w części odrzuconej. Trzeba zatem wykonać więcej zabiegów, by uzyskać lepsze wyniki.

4.3.2. Analiza rozkładów i ryzyka dla zmiennych

Posiadając model KGB dla zaakceptowanych, możemy wyznaczyć teoretyczne wartości ryzyka dla wniosków odrzuconych. Do tego celu kalibrujemy nasze nowe oceny punktowe do wartości prawdopodobieństwa na dostępnej próbie zaakceptowanych. Pamiętajmy, że

Tabela 23: Porównanie mocy predykcyjnych modeli KGB i PD Ins

Zmienna	Segment / Gini modelu (%)	KGB	PD Ins
default_{12}^{GB}	Odrzucone	14,09	48,29
	Zaakceptowane	36,15	41,29
	Wszystkie	24,73	65,55
default_{12}	Odrzucone	15,17	50,70
	Zaakceptowane	37,34	42,77
	Wszystkie	26,12	67,60

Źródło: opracowanie własne.

w rzeczywistości statusy default nie są znane na części odrzuconej, zatem stary model też jest kalibrowany tylko na części zaakceptowanej. Mamy zatem wzory składni języka SAS 4GL:

```
pd_ins_new=1/(1+exp(-(-0.032397939
        *risk_ins_score_new+9.5756814988)));
pd_ins_old=1/(1+exp(-(-0.037844597
        *risk_ins_score+11.929922366)));
```

gdzie `risk_ins_score_new`, `risk_ins_score` są ocenami punktowymi odpowiednio modeli nowego i starego.

W dokumentacji starego modelu wyznaczono kalibrację dla całej populacji (włączając także wnioski odrzucone), gdyż model ten był budowany przy stu procentowej akceptacji (porównaj podrozdział 5.2.1):

```
pd_ins=1/(1+exp(-(-0.032205144
        *risk_ins_score+9.4025558419)));
```

W części zaakceptowanej, dysponując obserwowanymi statusami default, nie musimy już nic zmieniać, by budować modele. W części odrzuconej brakuje nam statusów default przypisanych każdemu wnioskowi oddzielnie. Na razie wykonajmy porównanie estymowanego ryzyka poprzez wykorzystanie kalibrowanych ocen punktowych. Tworzymy trzy dodatkowe kolumny estymacji ryzyka E_Nowa, E_Stara i E_Cała w następujący sposób: jeśli jest to wnio-

sek zaakceptowany, to przyjmuje obserwowane wartości default, jeśli jest odrzucony, to przyjmuje wartość prawdopodobieństwa odpowiednio: pd_ins_new, pd_ins_old i pd_ins.

Tabela 24: Rozkłady zmiennej ACT_CINS_N_STATB

Numer	Warunek / Procent	Z	O	W
1	0 < ACT_CINS_N_STATB	4,71	12,2	7,47
2	Missing	70,72	76,68	72,92
3	ACT_CINS_N_STATB $\leqslant$ 0	24,57	11,12	19,61

Źródło: opracowanie własne.

Tabela 25: Zmienna ACT_CINS_N_STATB. Ryzyko oraz jego estymacje dla kategorii

Numer	Ryzyko (%)			E_Nowa (%)		E_Stara (%)		E_Cała (%)	
	Z	O	W	O	W	O	W	O	W
1	9,50	35,25	25,02	8,19	8,71	37,46	26,35	26,85	19,95
2	5,61	27,69	14,17	5,71	5,65	33,35	16,37	22,53	12,17
3	1,67	34,68	8,58	1,48	1,63	51,13	12,02	38,26	9,33
Razem	4,82	29,39	13,89	5,54	5,09	35,83	16,26	24,80	12,20

Źródło: opracowanie własne.

Przeprowadźmy studia rozkładów wybranej zmiennej: ACT_CINS_N_STATB mówiącej o liczbie wszystkich kredytów ratalnych klienta źle zakończonych (czyli niespłaconych, w statusie B) w celu lepszego rozumienia estymacji ryzyka na części odrzuconej. Zauważmy po pierwsze, że udziały grup posiadają inne rozkłady w zależności od części: O – odrzuconej, Z – zaakceptowanej i W – wszystkich razem. Udziały gorszych segmentów, w szczególności pierwszego, muszą być większe po stronie odrzuconych; patrz tabela 24. Porównując ryzyko i estymacje musimy pamiętać, że wszystkie wartości ryzyka na części zaakceptowanej są takie same, dlatego w tabeli 25 ryzyko dla grupy Z jest przedstawione tylko raz. Widać wyraźnie, że grupa pierwsza, najbardziej ryzykowna, dla części odrzuconej

jest bardzo źle estymowana przez nowy model. Stary model przeszacowuje ryzyko, a estymacja całości jest najlepsza, choć lekko je zaniża. Estymacja całości z reguły w praktyce nie jest dana. Szczególnie, że związana jest z czasem budowy starego modelu, czyli może już nie reprezentować obecnego portfela klientów. Estymacja stara jest metodą odświeżenia informacji, ale niestety kosztem kalibracji tylko na części zaakceptowanej. Pomimo tego kosztu model stary ma w sobie możliwości identyfikowania różnych profili klientów: i na części zaakceptowanej, i na odrzuconej. Kalibrowanie tylko do podzbioru zaakceptowanych bardzo nie zakłóca jego zdolności do estymacji odrzuconych. W dalszej części będziemy zatem wykorzystywać już tylko estymację starą, pomijając całości. Widać wyraźnie, że nie ma tu idealnego rozwiązania. Każda estymacja, jak sama nazwa wskazuje, jest tylko przybliżeniem obserwowanego ryzyka, którego niestety w praktyce nie znamy. Do jego estymacji na części odrzuconej musimy zatem podejść bardzo subtelnie.

4.3.3. Analiza ocen punktowych i kalibracji

Wartości ocen punktowych nowego modelu podzielone są na 11 grup. Dla każdej grupy obliczono wartości obserwowanego ryzyka oraz estymowanego na sposób nowy i stary. Na rysunku 18 dodatkowo pokazano wartości prawdopodobieństwa wyznaczone z nowego modelu. Dla części zaakceptowanej wszystkie krzywe mają zbliżone własności. Zupełnie inaczej wygląda to na części odrzuconej, patrz rysunek 19. Nowa estymacja jest znacząco niedoszacowana. Ten wykres pokazuje istotne znaczenie wpływu wniosków odrzuconych. Wyraźnie da się zauważyć, że profil klienta odrzuconego ma zupełnie inne własności od zaakceptowanego. Nie mamy monotonicznej własności ryzyka na odrzuconych, wzrost ocen punktowych nie powoduje zmniejszania się wartości obserwowanego ryzyka, co szczególnie widać dla oceny 423. Estymacja stara i obserwowane ryzyko, którego niestety w rzeczywistości nie można obserwować, są zbliżone do siebie. Kształty krzywych ryzyka i estymacji starej są na tyle odmienne od estymacji nowej, że można by mówić o odwróceniu profilu klienta, a mianowicie kiedy ocena punktowa rośnie, to przy dużych jej wartościach ryzyko rośnie zamiast maleć. Jest to

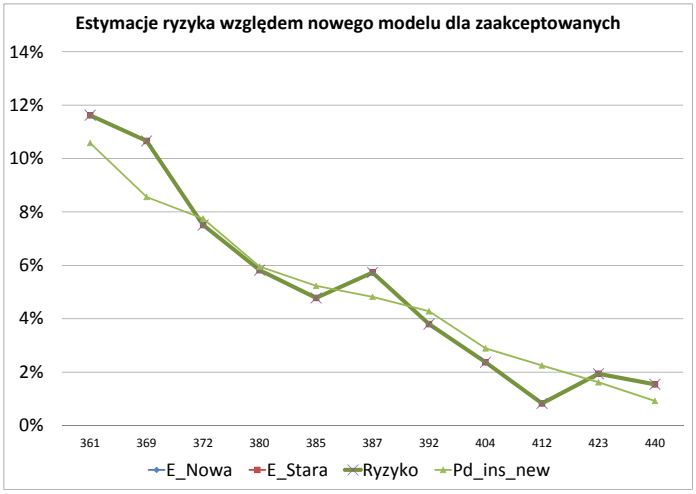
interesujący fakt obserwowany na danych symulacyjnych dla portfela odrzuconych wniosków. W rzeczywistości dla tego zbioru nie jest znane prawdziwe ryzyko, więc problem ten nie może być poprawnie zbadany. Jeśli jednak przy danych symulacyjnych pojawia się tego typu zjawisko, to należy je mieć w pamięci i zdawać sobie sprawę, że klient odrzucony przez stary model może mieć zupełnie inne, prawie odwrotne, własności od klienta zaakceptowanego. Na rysunku 20 przedstawiono wartości ryzyka całej próby modelowej. Ryzyko obserwowane potrafi być prawie 3 razy większe od nowej estymacji. Co więcej mamy tu do czynienia z kilkoma nieregularnościami krzywej obserwowanego ryzyka, które nie występowały na części zaakceptowanej.

Wszystkie przedstawione liczby i wykresy wyraźnie przekonują nas o poważnej trudności w estymacji ryzyka na części odrzuconej. Temat wniosków odrzuconych jest dziś już bardzo znany, w literaturze mnożą się metody, a każdy autor przekonuje nas, że jego sposób jest najlepszy. Nie wolno jednak dać się ponieść idei myślenia pozytywnego i bezkrytycznie stosować jakąś wybraną metodę. Właśnie dane symulacyjne mogą stać się dobrym narzędziem rozstrzygającym o poprawności różnych metod oraz być może pomogą w badaniach nad jeszcze lepszymi, a może najlepszymi. W kontekście Big Data nie wolno zapomnieć o rozważanych w książce przykładach, by nie bagatelizować tego tematu i nie stosować gotowych narzędzi typu „czarna skrzynka".

4.3.4. Finalna estymacja na bazie krzywych logitowych

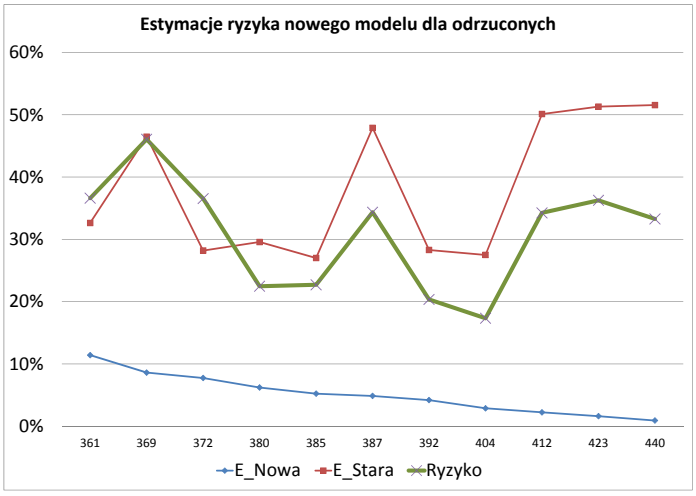
Biznes nie lubi określenia niewykonalne lub niemożliwe. Zawsze trzeba jednak problem rozwiązać. Należy uświadomić sobie dane, które posiadamy: ryzyko na zaakceptowanych i dwa modele stary i nowy (PD Ins i KGB). Stary model powinien był być poprawnie skalibrowany i umieć, przynajmniej w części, estymować ryzyko odrzuconych. W naszym przypadku tak jest w zupełności, gdyż model ten był budowany na całej populacji. W rzeczywistości dla typowego banku stary model był kalibrowany metodami wpływu wniosków odrzuconych jakiś czas temu, kiedy był budowany. Z reguły, co gorsze, model ten traci swoją moc predykcyjną, co jest właśnie powo-

Rysunek 18: Model KGB. Estymacja ryzyka dla zaakceptowanych



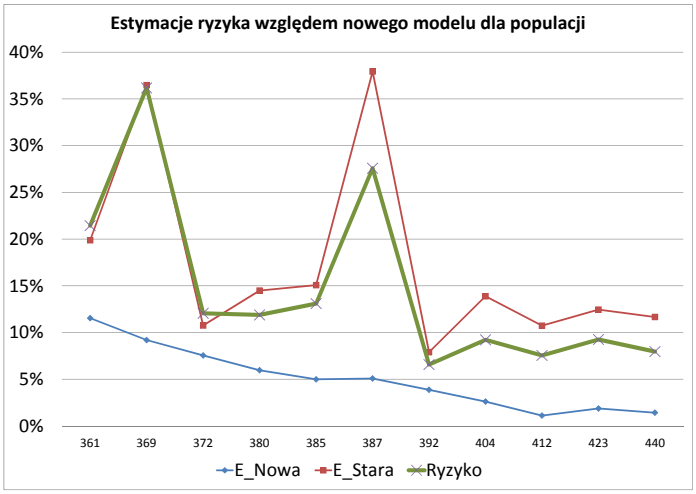
Źródło: opracowanie własne.

Rysunek 19: Model KGB. Estymacja ryzyka dla odrzuconych



Źródło: opracowanie własne.

Rysunek 20: Model KGB. Estymacja ryzyka dla całej próby modelowej



Źródło: opracowanie własne.

dem budowy nowego. Mamy zatem dylemat, czy wierzyć staremu i słabemu modelowi, czy nowemu. Oba modele coś nam mówią swoistego i jedynego. Jak rozstrzygnąć, który lepiej opisuje profil odrzuconego klienta?

Rozważmy krzywe logitowe obliczone dla ryzyka i estymacji względem starych ocen punktowych. Ponieważ logity z prawdopodobieństw są liniową kombinacją ocen punktowych, to krzywe te powinny być liniami prostymi (często nazywane są wykresami szans, z ang. odds chart). Dla segmentów starych ocen punktowych wykonano rysunek 21. LogitW (logit wszystkich: zaakceptowanych i odrzuconych) reprezentuje tu krzywa o równaniu:

```
[Logit(pd_ins_old)=] y=-0.037844597
       *risk_ins_score+11.929922366;
```

które jest wynikiem wcześniejszej kalibracji starego modelu na części zaakceptowanej.

Krzywa LogitZ odpowiada obserwowanym logitom na części zaakceptowanej.

Jeśli dodatkowo obliczymy logity od estymacji nowej, to okaże się, że jej punkty układają się w zupełnie inną krzywą nierównoległą do już narysowanych. Możemy zatem przypuścić, że nasza poprawna estymacja ryzyka jest kombinacją następujących członów:

$$\text{Dopasowanie} = \text{logit}(p) = \alpha\,\text{Logit}(\text{E\_Nowa}) + \beta\,\text{Logit}(\text{E\_Stara}).$$

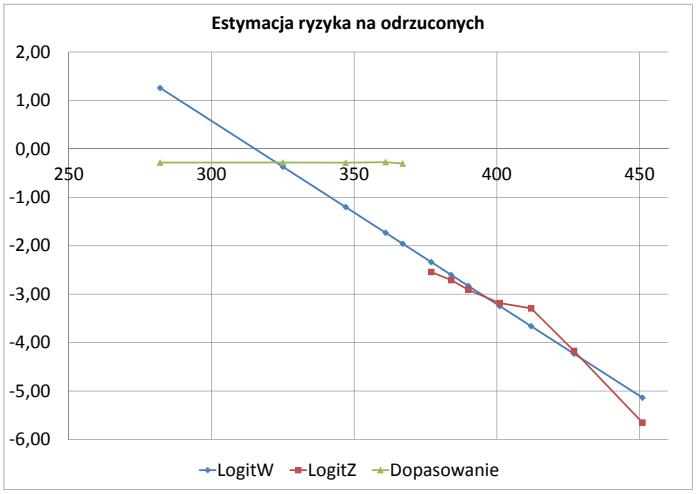
Problem w tym, że nikt nie jest w stanie podać wartości obu współczynników[5]. W naszym przykładzie jest to prostsze, bo znamy obserwowane ryzyko i możemy zbadać, który człon estymacji nowej czy starej jest bardziej adekwatny. W rzeczywistości tego nie możemy określić. Jeśli jednak udałoby się pozyskać dosłownie kilka punktów, a być może tylko dwa, to zadanie byłoby już prostsze. Znalezienie takich punktów jest możliwe dzięki narzędziom systemów decyzyjnych, które potrafią akceptować np. co tysięcznego klienta poniżej punktu odcięcia. Trzeba tylko umieć obliczyć, jak dużo powinniśmy takich klientów zaakceptować i jaki będzie tego koszt, wynikający z dużej straty. W przypadku bankowości dla polskiego rynku można się też posłużyć raportami Biura Informacji Kredytowej (BIK), gdzie możemy znaleźć informacje o spłacanych kredytach w innych bankach, pomimo naszej odmowy udzielenia kredytu.

Dla wartości $\beta = 0$ i $\alpha = 0{,}1$ logity wyznaczone z nowej estymacji nie pokrywają się z linią LogitW, przecinają się i wyraźnie mają inne nachylenie. Oznacza to, że informacja wzięta z profili zaakceptowanych klientów jest niewystarczająca do poprawnego estymowania ryzyka dla odrzuconych. Gdy dodatkowo ustawi się $\beta = 1$, (patrz rysunek 22), to nowo powstała linia jest znacząco lepsza. Sposób ten pokazuje dość klarownie, w jakim obszarze możemy się poruszać, szukając najlepszej możliwej estymacji ryzyka dla odrzuconych.

W naszym przypadku finalnie wybieramy współczynniki: $\alpha = 0$ i $\beta = 1$. W ten sposób zostało określone prawdopodobieństwo zdarzenia default dla wniosków odrzuconych; patrz tabela 26. Ryzyko na odrzuconych jest lekko przeszacowane, ale w ujęciu ostrożnościowym jest to lepsze rozwiązanie niż odwrotne.

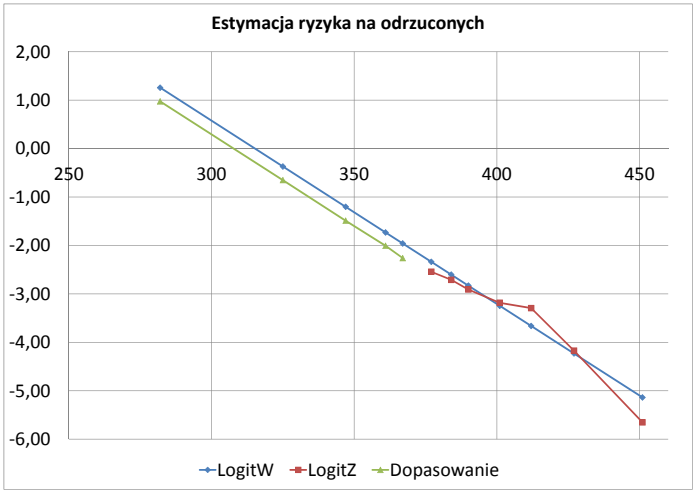
[5] Można także dodać wyraz wolny i próbować dobrać trzy współczynniki.

Rysunek 21: Dobór współczynników, $\alpha = 0{,}1$ i $\beta = 0$



Źródło: opracowanie własne.

Rysunek 22: Dobór współczynników, $\alpha = 0{,}1$ i $\beta = 1$



Źródło: opracowanie własne.

Tabela 26: Finalna estymacja ryzyka odrzuconych

Segment	Ryzyko (%)	Estymacja (%)
Z	4,82	4,82
O	29,39	35,36
W	13,89	16,09

Źródło: opracowanie własne.

4.4. Model ALL dla całej populacji

Na podstawie estymacji ryzyka dla odrzuconych możliwe jest teraz zbudowanie modelu dla wszystkich wniosków zarówno zaakceptowanych, jak i odrzuconych. Rozważmy tu dwa modele. Pierwszy oparty na tych samych zmiennych co model KGB, tylko skorygowanych w myśl zasady, że skoro tylko takie zmienne zostały wybrane do modelu na zaakceptowanych, czyli na bazie prawdziwie obserwowanych faktów, to już nie powinno się dobierać innych zmiennych na bazie estymowanego ryzyka dla odrzuconych, czyli na podstawie teoretycznych wartości. Drugi, w którym ponownie przeprowadza się selekcję wszystkich możliwych zmiennych, pozwalając w finalnym modelu uzyskać zupełnie inną listę zmiennych niż w modelu KGB. Okazuje się, że drugie podejście jest właściwsze, choć pozornie wydaje się zabiegiem mieszającym świat obserwowany z teoretycznym. Po wnikliwym przeanalizowaniu stwierdza się jednak, że idea jest poprawna. Otóż ze względu na estymację ryzyka na odrzuconych dochodzi nowa informacja o profilu złego klienta odrzucanego. Ten rodzaj danych w modelu KGB nie mógł być uwzględniony i zestaw zmiennych tego modelu dobierany był tylko, by identyfikować profil złego klienta zaakceptowanego. Możliwe jest zatem, że wiele pozostałych zmiennych jest w stanie zdefiniować reguły dla klientów odrzuconych i nie powinno się tej szansy pomijać. Pozostaje jedynie pewna wątpliwość, na ile błędna estymacja ryzyka odrzuconych zaburzy świat obserwowany. Niestety nie mamy wyjścia, wszystkie wcześniejsze rozumowania wyraźnie świadczyły, że model KGB nie może być stosowany na całej populacji, z wyjątkiem

sytuacji dużego udziału zaakceptowanych, np. 80% lub 90%, gdzie wpływ odrzuconych może być pominięty. Zatem tak czy inaczej popełniony jest błąd: albo stosujemy model KGB i zdajemy sobie sprawę z błędu niedoszacowania globalnego ryzyka nawet do kilkuset procent, albo próbujemy to ryzyko lepiej oszacować przez analizy odrzuconych i tworzymy model na teoretycznych danych, który też, niestety, obarczony jest błędem. Oszacowanie – w którym przypadku popełniony jest większy błąd – jest praktycznie niemożliwe. W naszym przypadku danych symulacyjnych okazuje się, że modele ALL są lepsze i błąd estymacji ryzyka jest znacząco mniejszy.

4.4.1. Przygotowanie danych, nowa definicja default

Nowe próby modelowe do budowy modelu ALL muszą zawierać statusy zdarzenia default zarówno dla wniosków zaakceptowanych, jak i dla odrzuconych. W pierwszym przypadku dane są obserwowane i gotowe, w drugim dana jest tylko estymacja ryzyka, czyli prawdopodobieństwo. Musimy zatem zdefiniować nową binarną funkcję celu, by potem zastosować model binarnej regresji logistycznej. Rozważmy jeden dowolny odrzucony wniosek. Przypuśćmy, że estymacja ryzyka wynosi 5%. Istnieją dwa najprostsze sposoby stworzenia nowego zbioru danych z nowym zdarzeniem default, który oznaczamy $\mathrm{default N}_{12}^{GB}$.

Ważenie obserwacji

W nowym zbiorze nasz wiersz reprezentowany jest przez dwa wiersze. Jeden z wagą 5% i wartością $\mathrm{default N}_{12}^{GB} = \mathbf{Zły}$, a drugi z wagą 95% i wartością $\mathrm{default N}_{12}^{GB} = \mathbf{Dobry}$. Wszystkie zaakceptowane wnioski wstawiane są z wagą 100% i z niezmienioną wartością default, czyli $\mathrm{default N}_{12}^{GB} = \mathrm{default}_{12}^{GB}$.

Powtórzenie obserwacji, 100 razy więcej

W nowym zbiorze nasz wiersz reprezentowany jest przez 100 wierszy. Pierwsze 5%, czyli 5 wierszy, z wartością $\mathrm{default N}_{12}^{GB} = \mathbf{Zły}$, a kolejne 95%, czyli 95 wierszy, z wartością $\mathrm{default N}_{12}^{GB} = \mathbf{Dobry}$. Wszystkie zaakceptowane wnioski są powtórzone 100 razy z niezmienioną wartością default, czyli $\mathrm{default N}_{12}^{GB} = \mathrm{default}_{12}^{GB}$.

Obie metody są równoważne, aczkolwiek druga jest bardziej czasochłonna, ze względu na większe wielkości danych, jak i trochę

mniej dokładna, gdyż transformuje wartości prawdopodobieństw tylko z dokładnością do jednego procenta. Z punktu widzenia technicznego jest jednak najprostszą, gdyż nie wymaga jakichkolwiek przeróbek kodów SAS 4GL, w szczególności wstawiania dodatkowej instrukcji `weight` do każdej procedury agregującej.

4.4.2. Model ALL1, lista zmiennych taka, jak w modelu KGB

Tabela 27: Model ALL1

Gini w % na zbiorze		Lift1	Lift5	Lift10	Lift20
treningowym	walidacyjnym				
31,14	31,43	3,91	3,76	2,76	2,26

Karta skoringowa

Zmienna	Warunek	Ocena cząstkowa
ACT_CINS_N_STATB	0 $<$ ACT_CINS_N_STATB or Missing	58
	ACT_CINS_N_STATB $\leqslant$ 0	67
APP_CHAR_GENDER	Male	58
	Female	62
APP_CHAR_JOB_CODE	Contract	58
	Owner company	103
	Retired	109
	Permanent	120
APP_INCOME	APP_INCOME $\leqslant$ 1503 or 4923 $<$ APP_INCOME	58
	1503 $<$ APP_INCOME $\leqslant$ 4923	62
APP_SPENDINGS	APP_SPENDINGS $\leqslant$ 240 or 2220 $<$ APP_SPENDINGS	58
	240 $<$ APP_SPENDINGS $\leqslant$ 2220	64

Źródło: opracowanie własne.

Pierwsze podejście do budowy modelu ALL, oznaczmy je jako model ALL1, związane jest z zachowaniem tej samej listy zmiennych jak w modelu KGB. Dokonuje się tu jedynie poprawek ze względu na wnioski odrzucone. Innymi słowy tworzone są nowe kategorie dla każdej ze zmiennych oraz ponownie kalibrowane są wartości ocen punktowych. W tabeli 31 (strona 130) przedstawiono moce predykcyjne wszystkich budowanych modeli. Widać wyraźnie, że moc modelu ALL1 znacząco spadła na zaakceptowanych, a wrosła dla odrzuconych. Co gorsze, model kalibrowany do wartości prawdopodobieństwa na całej populacji znacząco zawyża estymowane ryzyko na części zaakceptowanej, co oznacza, że niepotrzebnie zostanie

zmniejszona liczba akceptowanych dobrych wniosków. Finalna karta skoringowa modelu ALL1 przedstawiona jest w tabeli 27.

4.4.3. Nowa preselekcja zmiennych

W celu budowy kolejnego modelu, oznaczmy go przez ALL2, proces rozpoczynamy od rozważenia wszystkich możliwych zmiennych, nawet jeśli nie były w modelu KGB. Pierwszy etap preselekcji zmiennych pozostawił aż 116 zmiennych. W tabeli 28 przedstawiono tylko pierwszych 15 względem statystyki Giniego ($\mathrm{AR}_{\mathrm{Train}}$).

Tabela 28: Preselekcja zmiennych dla modelu ALL2, pierwsze 15 zmiennych

Zmienna	$\mathrm{AR}_{\mathrm{Train}}$ (%)
APP_CHAR_JOB_CODE	26,8
APP_LOAN_AMOUNT	26,3
ACT_LOANINC	25,0
APP_INSTALLMENT	22,1
ACT_CINS_MIN_SENIORITY	19,7
APP_NUMBER_OF_CHILDREN	19,3
ACT_CALL_CC	18,9
APP_CHAR_MARITAL_STATUS	18,2
APP_N_INSTALLMENTS	15,6
ACT_CC	15,2
APP_INCOME	13,3
APP_SPENDINGS	11,8
AGS9_MAX_CMAXI_DUE	11,2
AGS9_MEAN_CMAXI_DUE	11,2
AGS9_MIN_CMAXI_DUE	11,2

Źródło: opracowanie własne.

4.4.4. Wielowymiarowa selekcja zmiennych – generator modeli

Ze względu na dość dużą liczbę zmiennych pozostałych po etapie preselekcji wykonana jest selekcja zmiennych w modelu regresji logistycznej metodą krokową i uruchomiony jest algorytm z opcjami:

`selection=FORWARD SLSTAY=0.45 SLENTRY=0.45 START=0 STOP=70`.

W ten sposób otrzymujemy 43 zmienne.

W procedurze Logistic istnieje metoda selekcji oparta na heurystyce branch and bound (Furnival i Wilson, 1974). Jest to bardzo wygodna metoda, gdyż pozwala wygenerować wiele modeli, w tym wypadku 10, po 5 najlepszych z modeli o 7 zmiennych i o 8 zmiennych (opcja: `selection=SCORE START=7 STOP=8 BEST=5`).

Po analizie wybranych modeli, zbadaniu ich różnych statystyk, takich jak: stabilność, współliniowość i istotność zmiennych oraz po niewielkich poprawkach ręcznych zostaje wytypowana finalna lista zmiennych w modelu ALL2.

- ACT_CC – stosunek raty kredytowej do aktualnego wynagrodzenia klienta podanego we wniosku.
- ACT_CINS_MIN_SENIORITY – liczba miesięcy od ostatniego wniosku o kredyt ratalny.
- ACT_CINS_N_STATC – liczba spłaconych kredytów ratalnych (w statusie C).
- APP_CHAR_JOB_CODE – kod zawodu.
- APP_CHAR_MARITAL_STATUS – status małżeński.
- APP_LOAN_AMOUNT – wnioskowana kwota kredytu.
- APP_NUMBER_OF_CHILDREN – liczba dzieci będących na utrzymaniu wnioskującego.

4.4.5. Model ALL2, szeroka lista zmiennych

Finalną postać karty skoringowej modelu ALL2 przedstawiono w tabeli 29. Dodatkowo w tabeli 30 przedstawiono dwa mierniki ważności zmiennych w modelu. Pierwszy oparty jest na statystyce Giniego ($\mathrm{AR}_{\mathrm{Train}}$) liczonej oddzielnie dla każdej ze zmiennych. Drugi jest związany bezpośrednio z ocenami punktowymi; liczony jest tu

udział rozstępu ocen cząstkowych danej zmiennej w rozstępie finalnych ocen. Nie zawsze obie te miary nadają ten sam porządek zmiennym. Wydaje się, że udział oceny punktowej jest najlepszą miarą, gdyż uwzględnia dodatkowo wszelkie zależności pomiędzy zmiennymi, które razem zostały wzięte do modelu.

Tabela 29: Model ALL2

Gini w % na zbiorze		Lift1	Lift5	Lift10	Lift20
treningowym	walidacyjnym				
59,51	59,39	5,62	4,50	3,83	2,93

Karta skoringowa

Zmienna	Warunek	Ocena cząstkowa
ACT_CC	$1.00 < \text{ACT\_CC}$	24
	$0.85 < \text{ACT\_CC} \leqslant 1.00$	43
	$0.25 < \text{ACT\_CC} \leqslant 0.85$	52
	$\text{ACT\_CC} \leqslant 0.25$	64
ACT_CINS_MIN_SENIORITY	Missing or $\text{ACT\_CINS\_MIN\_SENIORITY} \leqslant 22$	24
	$22 < \text{ACT\_CINS\_MIN\_SENIORITY}$	53
ACT_CINS_N_STATC	$\text{ACT\_CINS\_N\_STATC} \leqslant 0$	24
	Missing	49
	$0 < \text{ACT\_CINS\_N\_STATC} \leqslant 2$	56
	$2 < \text{ACT\_CINS\_N\_STATC}$	65
APP_CHAR_JOB_CODE	Contract	24
	Owner company	83
	Retired	91
	Permanent	105
APP_CHAR_MARITAL_STATUS	Single, Divorced	24
	Married, Widowed	44
APP_LOAN_AMOUNT	$11376 < \text{APP\_LOAN\_AMOUNT}$	24
	$8880 < \text{APP\_LOAN\_AMOUNT} \leqslant 11376$	45
	$4824 < \text{APP\_LOAN\_AMOUNT} \leqslant 8880$	57
	$1968 < \text{APP\_LOAN\_AMOUNT} \leqslant 4824$	71
	$\text{APP\_LOAN\_AMOUNT} \leqslant 1968$	81
APP_NUMBER_OF_CHILDREN	$\text{APP\_NUMBER\_OF\_CHILDREN} \leqslant 0$	24
	$0 < \text{APP\_NUMBER\_OF\_CHILDREN} \leqslant 1$	48
	$1 < \text{APP\_NUMBER\_OF\_CHILDREN}$	88

Źródło: opracowanie własne.

W tabeli 31 pokazano moce predykcyjne modelu ALL2 w porównaniu z pozostałymi. Widać, że model ALL2 może już konkurować z modelem starym (PD Ins). Nie uzyskano wprawdzie aż tak dobrej mocy predykcyjnej jak model wzorcowy, ale i tak jest to zadowalający wynik. Z drugiej strony, niestety, osłabienie modelu o około 10% wartości Giniego może spowodować miesięczne milionowe spadki w zyskach z procesu akceptacji kredytowej. Niestety nie jest

Tabela 30: Lista zmiennych w modelu ALL2

Zmienna	$\mathrm{AR}_{\mathrm{Train}}$ (%)	Udział oceny punktowej (%)
APP_CHAR_JOB_CODE	26,8	24,40
APP_LOAN_AMOUNT	26,2	17,17
APP_NUMBER_OF_CHILDREN	19,3	19,28
APP_CHAR_MARITAL_STATUS	15,9	6,02
ACT_CC	15,0	12,05
ACT_CINS_MIN_SENIORITY	11,4	8,73
ACT_CINS_N_STATC	9,4	12,35

Źródło: opracowanie własne.

możliwe zbudowanie modelu porównywalnego do PD Ins na bazie procesu, który akceptuje $60{,}5\%$ kredytów ratalnych. Brak pełnej informacji o kliencie, nawet przy najlepszych metodach analiz wniosków odrzuconych, nigdy nie zostanie nadrobiony i zawsze będzie to kosztem mocy predykcyjnej nowego modelu. Można wykonać podobne analizy i wykresy jak dla modelu KGB; analogiczne do rysunków 18, 19, czy 20. Niestety w przypadku modelu ALL2 okazuje się podobnie, że estymacja ryzyka dla zaakceptowanych jest nieco zawyżona, wykresy są zdecydowanie lepsze w stosunku do modelu KGB, ale i tak pozostawiają jeszcze wiele do życzenia.

Z reguły w dokumentacji modelu podaje się jeszcze wiele szczegółowych raportów. Nie stanowią one już nowej treści poznawczej, ze względu na ich występowanie w innych wcześniejszych rozdziałach. Ograniczymy się jedynie do wymienienia ich rodzajów.

- Krzywe ROC, LIFT i CAP na zbiorach treningowym i walidacyjnym (Krzyśko *et al.*, 2008; BIS–WP14, 2005).
- Stabilność zmiennych w czasie: ewolucja udziałów kategorii (lub ocen cząstkowych) w czasie i ewolucja ryzyka dla kategorii w czasie.
- Stabilność modelu w czasie: ewolucja Giniego w czasie liczona na różnych statusach default, zarówno krótko-, jak i długookresowych.

- Stabilność w czasie kategorii ocen punktowych (z ang. score band), zarówno ich udziałów, jak i wartości ryzyk.
- Mierniki istotności zmiennych, wartości współczynników z regresji logistycznej.
- Raporty typu Vintage (najczęściej 90+) w rozbiciu na kategorie ocen punktowych.
- Mierniki współliniowości, takie jak: VIF – współczynnik inflacji wariancji, CI – indeks warunkujący, czasem też macierz korelacji Pearsona i Spearmana (Koronacki i Mielniczuk, 2001; Welfe, 2003; Belsley *et al.*, 1980; Belsley, 1991), włączając w to także sprawdzenie, ile współczynników regresji zmieniło znak.
- Przedziały ufności dla Giniego (BIS–WP14, 2005) i ewentualnie przedziały uzyskane metodą bootstrapową.

Stosuje się też jeszcze inne diagnostyki modeli, takie jak: Brier Score, Pietra Index czy Hosmer–Lemeshow (BIS–WP14, 2005). Ciekawe koncepcje badania modeli zaproponowane są przez (Scallan, 2011), gdzie proponuje się statystyki MIV (ang. Marginal Information Value), czy MKS (ang. Marginal Kolmogorov–Smirnov) (Scallan, 2013).

Tabela 31: Porównanie mocy predykcyjnych modeli: KGB, ALL1, ALL2 i PD Ins

Zmienna	Segment / Gini modeli (%)	KGB	ALL1	ALL2	PD Ins
$\mathrm{default}_{12}^{GB}$	Odrzucone	14,09	17,14	24,93	48,29
	Zaakceptowane	36,15	3,85	32,81	41,29
	Wszystkie	24,73	31,30	54,08	65,55
$\mathrm{default}_{12}$	Odrzucone	15,17	17,64	26,53	50,70
	Zaakceptowane	37,34	4,17	34,11	42,77
	Wszystkie	26,12	32,23	56,13	67,60

Źródło: opracowanie własne.

4.5. Segmentacja portfela. Jeden model kontra kilka

Jednym z dość często występujących problemów podczas budowy modeli jest pytanie, czy budować jeden wspólny model dla całego portfela, czy też kilka dedykowanych do każdego wyszczególnionego segmentu. Odpowiedź nie jest taka łatwa, gdyż związana jest z całym procesem biznesowym. Pierwszym argumentem dość prostym w rozstrzygnięciu jest aspekt złożoności procesu, w szczególności zarządzania regułami w systemie decyzyjnym. Jeśli budujemy kilka modeli, to musimy wszystkie wdrożyć do systemu, co oznacza więcej pracy, więcej dodatkowych parametrów i ogólnie ryzyko operacyjne jest większe od przypadku wdrożenia tylko jednego modelu.

Niestety ryzyko kredytowe najczęściej jest większe od ryzyka operacyjnego złożoności implementacji wielu modeli. Poza tym istnienie dedykowanych modeli dla wybranych segmentów z reguły pozwala lepiej zarządzać portfelem, co w efekcie przynosi tylko same korzyści. Jedynym minusem jest większa praca i złożoność rozwiązania.

Rozważmy zatem dwa przypadki. Pierwszy, gdy zbudowany jest tylko jeden model na całym dostępnym zbiorze wniosków kredytowych. Ograniczmy się do sytuacji strategii akceptującej wszystkie wnioski i rozważmy już omówiony model PD Ins, związany z prognozowaniem ryzyka kredytu ratalnego, (patrz tabela 10 ze strony 83 oraz tabela 44, strona 147). Model ten kalibrowany jest do wartości PD formułą podaną w podrozdziale 5.2.1. Ze względu na kilka oznaczeń wartości prawdopodobieństwa zdarzenia default (PD) w tym rozdziale przyjmijmy, że dla tego modelu oznaczamy je jako PD całości.

W drugim przypadku czeka nas większa praca, bo trzeba zbudować dwa modele. Wcześniej musimy określić dwa segmenty. Z reguły dobór segmentów jest związany z procesami biznesowymi i pewną wiedzą ekspercką. Typowe podziały są związane z produktami kredytowymi, czy kombinacjami tych produktów. Często portfel kart kredytowych dzieli się na te często używane i rzadko. Podziały te najczęściej powodują znacznie lepsze własności wybieranych zmiennych z ABT. Jeśli mamy kartę kredytową z wieloma transakcjami,

to da się tu policzyć wiele ciekawych zmiennych opisujących typy transakcji, trendy, wielkości i częstotliwości itp. W przypadku małej liczby transakcji zmienne te będą miały najczęściej braki danych i nie będą użyteczne w modelowaniu. Można zatem mówić o kryterium analitycznym, aby dzielić populację na takie segmenty, by zmienne były dobrze określone. W naszym przypadku segmentacja związana jest z pojęciem znanego i nieznanego klienta. Jeśli klient w momencie aplikowania o kredyt ratalny ma już jakąś historię w naszym banku, to możemy go nazwać znanym, a w przeciwnym wypadku – nieznanym. W naszym portfelu znanego klienta możemy zdefiniować na podstawie zmiennej ACT_CINS_SENIORITY, która oblicza liczbę miesięcy od pierwszego wniosku kredytu ratalnego do aktualnie wnioskowanego. Jeśli zmienna ta jest niepusta, to mamy do czynienia ze znanym klientem, który stanowi około 30% wszystkich wniosków.

Podczas budowy modelu dla segmentu znanego klienta etap preselekcji zmiennych wytypował aż 155 zmiennych. Jest to wynikiem faktu, że klient miał już wcześniej kredyty, więc wiele zmiennych behawioralnych było niepustych i uzyskały istotne moce predykcyjne, większe niż w przypadku klienta nieznanego.

Należy tu zwrócić uwagę na pewną subtelność. Typowe procesy bankowe dla znanego klienta obliczają model skoringowy, zwany behawioralnym, budowany dla trochę innej definicji zdarzenia default i ogólnie mierzący inne prawdopodobieństwo. Jest to typowy model PD wynikający z rekomendacji Basel II i III służący do oceny klientów, czyli do pomiaru ich ryzyka pod warunkiem, że posiadają kredyty, takie jak na moment analizy, by wyznaczać w ten sposób wymogi kapitałowe. W naszym wypadku budowany jest model, który posiada cechy behawioralne, ale mierzy ryzyko, że klient, posiadając już kredyty w momencie aplikowania, nie wywiąże się z zobowiązań, biorąc dodatkowo jeszcze jeden nowy kredyt. Mamy tu zatem inne warunkowanie statystyczne. Niestety ze względu na rekomendację, by modele PD do wymogów kapitałowych były stosowane w procesach bankowych, często zapomina się o różnicy w warunkowaniu i stosuje wspomniane modele także do akceptacji nowych kredytów. Z reguły traci się na mocy predykcyjnej takich modeli. Nic nie stoi na przeszkodzie, by w procesie akceptacji były

używane różne modele i analizowane były zarówno wymogi kapitałowe, jak i inne miary ryzyka, zysku, czy CLTV (ang. Customer LifeTime Value), by jak najlepiej podejmować decyzje kredytowe.

Porównajmy model dla znanego klienta z modelem PD Ins. Oba modele mają jedną wspólną zmienną ACT_CINS_N_STATC, oznaczającą liczbę już zamkniętych kredytów ratalnych. Zobaczmy zatem, czym różnią się rozkłady tej zmiennej w zależności od próby modelowej.

Na próbie modelowej znanego klienta zmienna ta posiada moc predykcyjną o wartości $27{,}47\%$, natomiast w przypadku modelu PD Ins tylko $10{,}28\%$. Wyraźne różnice w rozkładach można zauważyć, studiując tabelę 32. Brak danych stanowi $71{,}2\%$ w przypadku modelu na całości, co właśnie determinuje znacząco mniejszą moc predykcyjną. Teraz staje się jasne, że właściwy dobór segmentów poprawia własności zmiennych. W drugim przypadku, modelu znanego klienta, w ogóle nie ma braku danych, dzięki temu kategorii może być więcej i może ta zmienna dzielić próbę na znacznie subtelniejsze podgrupy.

W przypadku modelu nieznanego klienta preselekcja zmiennych wybrała 13 zmiennych. Tylko tak mało, gdyż można było tu wybierać tylko wśród zmiennych typowo aplikacyjnych, czyli takich, które albo deklaruje klient na wniosku, albo od nich pochodnych. Jedna zmienna jest wspólna z modelem na całości: APP_CHAR_JOB_CODE. Różnice w ich rozkładach przedstawiono w tabeli 33. W tym wypadku nie mamy znaczącej różnicy, gdyż jest to zmienna aplikacyjna. Moc predykcyjna na próbie z całości wynosi $23{,}86\%$, natomiast dla nieznanego $29{,}14\%$, co tłumaczy, że zmienne behawioralne mocniej dyskryminują i dlatego wśród znanych i nieznanych razem czysta zmienna aplikacyjna musi mieć mniejszą moc. Można także zauważyć, co jest zgodne z intuicją, że klienci bardziej ryzykowni, czyli np. kontraktowcy w próbie nieznanego klienta są bardziej liczni i mają tu większe ryzyko niż na próbie z całości. Wynika to z prostego faktu, że ta część klientów będzie z reguły miała mały procent akceptacji, czyli będzie więcej po stronie nieznanego niż znanego klienta.

Wreszcie każde dwa nowe modele zostały skalibrowane do prawdopodobieństwa PD, każdy model osobno na innej próbie. Warto-

Tabela 32: Własności zmiennej ACT_CINS_N_STATC dla modeli znanego klienta i PD Ins

Rozkłady zmiennej dla modelu znanego klienta

Warunek	Liczba	Procent	Ryzyko (%)
ACT_CINS_N_STATC $\leqslant$ 0	666	16,3	28,5
0 < ACT_CINS_N_STATC $\leqslant$ 2	2 616	63,8	13,2
2 < ACT_CINS_N_STATC $\leqslant$ 3	367	9,0	9,3
3 < ACT_CINS_N_STATC $\leqslant$ 4	222	5,4	5,4
4 < ACT_CINS_N_STATC	227	5,5	2,2

Rozkłady zmiennej dla modelu PD Ins

Warunek	Liczba	Procent	Ryzyko (%)
ACT_CINS_N_STATC $\leqslant$ 0	535	4,7	29,0
0 < ACT_CINS_N_STATC $\leqslant$ 1	1 528	13,4	12,6
Missing	8 105	71,2	12,4
1 < ACT_CINS_N_STATC $\leqslant$ 2	604	5,3	11,4
2 < ACT_CINS_N_STATC	607	5,3	6,1

Źródło: opracowanie własne.

Tabela 33: Własności zmiennej APP_CHAR_JOB_CODE dla modeli nieznanego klienta i PD Ins

Rozkłady zmiennej dla modelu nieznanego klienta

Warunek	Liczba	Procent	Ryzyko (%)
Contract	823	8,2	43,1
Owner company	1 236	12,3	15,0
Retired	4 276	42,5	10,3
Permanent	3 725	37,0	8,8

Rozkłady zmiennej dla modelu PD Ins

Warunek	Liczba	Procent	Ryzyko (%)
Contract	768	6,7	42,1
Owner company	1 265	11,1	15,3
Retired	5 754	50,6	10,5
Permanent	3 592	31,6	9,4

Źródło: opracowanie własne.

ści tych prawdopodobieństw oznaczmy jako PD segmentów. Mamy zatem możliwość porównania PD całości z PD segmentów; patrz tabela 34. Oba modele są poprawnie skalibrowane na całej próbie i pokazują 13%, tak jak obserwowane ryzyko. Na segmentach mamy jednak już różnicę; model całości zaniża prognozowane ryzyko nieznanego klienta, a zawyża znanego. Ten fakt jest istotnym powodem dzielenia populacji i budowania oddzielnych modeli. Jeśli chcemy, aby prognozy PD odzwierciedlały obserwowane wartości dla krytycznych dla biznesu segmentów, bo np. spodziewamy się na nich istotnie małego lub dużego ryzyka, to lepiej budować modele oddzielnie. Oczywiście segmenty nie mogą być zbyt małoliczne, gdyż wtedy też nie uda się dobrze estymować ryzyka. Druga informacja w tabeli prezentuje moce predykcyjne. Wyraźnie widać, że moce dla modeli dedykowanych, budowanych dla segmentów, mają większe wartości i sumaryczna ich wartość Giniego na całej próbie także jest większa. Jeśli przypomnimy sobie wnioski z opłacalności procesu (patrz podrozdział 2.1), to jasne stanie się, że im większa moc pre-

dykcyjna, tym więcej zarabiamy. Zatem zarówno lepiej estymujemy ryzyko na segmentach, jak i ogólnie dyskryminujemy klientów, a to finalnie powoduje większe pomnażanie kapitału przedsiębiorstwa.

Jeśli moce predykcyjne jednego modelu na segmentach są zadowalające, to pośrednim rozwiązaniem może być kalibrowanie modelu do każdego segmentu. Wtedy finalnie implementowany będzie jeden model z różnymi wzorami na PD dla segmentów.

Tabela 34: Porównanie wskaźników pojedynczego modelu z dwoma na segmentach

Ryzyko obserwowane – oczekiwane (PD)

Segmenty	Liczba	Procent	Ryzyko (%)	PD całości (%)	PD segmentów (%)
Całość	23 637	100,00	13,00	13,00	13,00
Nieznany klient	16 827	71,19	12,61	11,34	12,61
Znany klient	6 810	28,81	13,96	17,09	13,96

Moce predykcyjne obu modeli po segmentach

Segmenty	PD całości (%)	PD segmentów (%)
Całość	73,11	78,07
Nieznany klient	65,51	71,02
Znany klient	87,72	89,78

Źródło: opracowanie własne.

5. Szczegółowe informacje i dokumentacje

5.1. Tabela analityczna, opisy wszystkich zmiennych

Opisy wszystkich zmiennych zostały stworzone automatycznym programem w SAS w języku angielskim. Wszystkie zmienne przedstawiono w tabelach 35–43. Pominięto tu jedynie funkcje celu, gdyż poświęcony jest nim specjalny podrozdział 1.3.10.

Tabela 35: Wszystkie zmienne ABT, część 1

Nr	Nazwa zmiennej	Opis (w języku angielskim)
1	cid	Id of application
2	aid	Id of Cust.
3	period	Year, month in format YYYYMM
4	act_age	Actual Cust. age
5	act_cc	Actual credit capacity (installment plus spendings) over income
6	act_loaninc	Loan amount over income
7	app_income	Cust. income
8	app_loan_amount	Loan amount
9	app_n_installments	Number of installments
10	app_number_of_children	Number of children
11	app_spendings	Spendings
12	app_installment	Installment amount
13	app_char_branch	Branch
14	app_char_gender	Gender
15	app_char_job_code	Job code
16	app_char_marital_status	Marital status
17	app_char_city	City type
18	app_char_home_status	Home status
19	app_char_cars	Cars

Źródło: opracowanie własne.

Tabela 36: Wszystkie zmienne ABT, część 2

Nr	Nazwa zmiennej	Opis (w języku angielskim)
20	act_call_n_loan	Actual Cust. loan number
21	act_ccss_n_loan	Actual Cust. loan number of Css product
22	act_cins_n_loan	Actual Cust. loan number of Ins product
23	act_ccss_maxdue	Cust. actual maximal due installments on product css
24	act_cins_maxdue	Cust. actual maximal due installments on product ins
25	act_ccss_n_loans_act	Cust. actual number of loans on product css
26	act_cins_n_loans_act	Cust. actual number of loans on product ins
27	act_ccss_utl	Cust. actual utilization rate on product css
28	act_cins_utl	Cust. actual utilization rate on product ins
29	act_call_cc	Cust. credit capacity (all installments plus spendings) over income
30	act_ccss_cc	Cust. credit capacity (installment plus spendings) over income on product css
31	act_cins_cc	Cust. credit capacity (installment plus spendings) over income on product ins
32	act_ccss_dueutl	Cust. due installments over all installments rate on product css
33	act_cins_dueutl	Cust. due installments over all installments rate on product ins
34	act_cus_active	Cust. had active (status=A) loans one month before
35	act_ccss_n_statB	Cust. historical number of finished loans with status B on product css
36	act_cins_n_statB	Cust. historical number of finished loans with status B on product ins
37	act_ccss_n_statC	Cust. historical number of finished loans with status C on product css
38	act_cins_n_statC	Cust. historical number of finished loans with status C on product ins
39	act_ccss_n_loans_hist	Cust. historical number of loans on product css
40	act_cins_n_loans_hist	Cust. historical number of loans on product ins
41	act_ccss_min_lninst	Cust. minimal number of left installments on product css
42	act_cins_min_lninst	Cust. minimal number of left installments on product ins
43	act_ccss_min_pninst	Cust. minimal number of paid installments on product css
44	act_cins_min_pninst	Cust. minimal number of paid installments on product ins
45	act_ccss_min_seniority	Cust. minimal seniority on product css
46	act_cins_min_seniority	Cust. minimal seniority on product ins
47	act3_n_arrears	Cust. number in arrears on last 3 mths on all loans
48	act6_n_arrears	Cust. number in arrears on last 6 mths on all loans
49	act9_n_arrears	Cust. number in arrears on last 9 mths on all loans
50	act12_n_arrears	Cust. number in arrears on last 12 mths on all loans

Źródło: opracowanie własne.

Tabela 37: Wszystkie zmienne ABT, część 3

Nr	Nazwa zmiennej	Opis (w języku angielskim)
51	act3_n_arrears_days	Cust. number on last 3 mths of days greter than 15 on all loans
52	act6_n_arrears_days	Cust. number on last 6 mths of days greter than 15 on all loans
53	act9_n_arrears_days	Cust. number on last 9 mths of days greter than 15 on all loans
54	act12_n_arrears_days	Cust. number on last 12 mths of days greter than 15 on all loans
55	act3_n_good_days	Cust. number of days lower than 15 on last 3 mths on all loans
56	act6_n_good_days	Cust. number of days lower than 15 on last 6 mths on all loans
57	act9_n_good_days	Cust. number of days lower than 15 on last 9 mths on all loans
58	act12_n_good_days	Cust. number of days lower than 15 on last 12 mths on all loans
59	act_ccss_seniority	Cust. seniority on product css
60	act_cins_seniority	Cust. seniority on product ins
61	ags12_Max_CMaxC_Days	Max calc. on last 12 mths on max Cust. days for Css product
62	ags12_Max_CMaxI_Days	Max calc. on last 12 mths on max Cust. days for Ins product
63	ags12_Max_CMaxA_Days	Max calc. on last 12 mths on max Cust. days for all product
64	ags12_Max_CMaxC_Due	Max calc. on last 12 mths on max Cust. due for Css product
65	ags12_Max_CMaxI_Due	Max calc. on last 12 mths on max Cust. due for Ins product
66	ags12_Max_CMaxA_Due	Max calc. on last 12 mths on max Cust. due for all product
67	agr12_Max_CMaxC_Days	Max calc. on last 12 mths on unmissing max Cust. days for Css product
68	agr12_Max_CMaxI_Days	Max calc. on last 12 mths on unmissing max Cust. days for Ins product
69	agr12_Max_CMaxA_Days	Max calc. on last 12 mths on unmissing max Cust. days for all product
70	agr12_Max_CMaxC_Due	Max calc. on last 12 mths on unmissing max Cust. due for Css product
71	agr12_Max_CMaxI_Due	Max calc. on last 12 mths on unmissing max Cust. due for Ins product
72	agr12_Max_CMaxA_Due	Max calc. on last 12 mths on unmissing max Cust. due for all product
73	ags3_Max_CMaxC_Days	Max calc. on last 3 mths on max Cust. days for Css product
74	ags3_Max_CMaxI_Days	Max calc. on last 3 mths on max Cust. days for Ins product

Źródło: opracowanie własne.

Tabela 38: Wszystkie zmienne ABT, część 4

Nr	Nazwa zmiennej	Opis (w języku angielskim)
75	ags3_Max_CMaxA_Days	Max calc. on last 3 mths on max Cust. days for all product
76	ags3_Max_CMaxC_Due	Max calc. on last 3 mths on max Cust. due for Css product
77	ags3_Max_CMaxI_Due	Max calc. on last 3 mths on max Cust. due for Ins product
78	ags3_Max_CMaxA_Due	Max calc. on last 3 mths on max Cust. due for all product
79	agr3_Max_CMaxC_Days	Max calc. on last 3 mths on unmissing max Cust. days for Css product
80	agr3_Max_CMaxI_Days	Max calc. on last 3 mths on unmissing max Cust. days for Ins product
81	agr3_Max_CMaxA_Days	Max calc. on last 3 mths on unmissing max Cust. days for all product
82	agr3_Max_CMaxC_Due	Max calc. on last 3 mths on unmissing max Cust. due for Css product
83	agr3_Max_CMaxI_Due	Max calc. on last 3 mths on unmissing max Cust. due for Ins product
84	agr3_Max_CMaxA_Due	Max calc. on last 3 mths on unmissing max Cust. due for all product
85	ags6_Max_CMaxC_Days	Max calc. on last 6 mths on max Cust. days for Css product
86	ags6_Max_CMaxI_Days	Max calc. on last 6 mths on max Cust. days for Ins product
87	ags6_Max_CMaxA_Days	Max calc. on last 6 mths on max Cust. days for all product
88	ags6_Max_CMaxC_Due	Max calc. on last 6 mths on max Cust. due for Css product
89	ags6_Max_CMaxI_Due	Max calc. on last 6 mths on max Cust. due for Ins product
90	ags6_Max_CMaxA_Due	Max calc. on last 6 mths on max Cust. due for all product
91	agr6_Max_CMaxC_Days	Max calc. on last 6 mths on unmissing max Cust. days for Css product
92	agr6_Max_CMaxI_Days	Max calc. on last 6 mths on unmissing max Cust. days for Ins product
93	agr6_Max_CMaxA_Days	Max calc. on last 6 mths on unmissing max Cust. days for all product
94	agr6_Max_CMaxC_Due	Max calc. on last 6 mths on unmissing max Cust. due for Css product
95	agr6_Max_CMaxI_Due	Max calc. on last 6 mths on unmissing max Cust. due for Ins product
96	agr6_Max_CMaxA_Due	Max calc. on last 6 mths on unmissing max Cust. due for all product
97	ags9_Max_CMaxC_Days	Max calc. on last 9 mths on max Cust. days for Css product
98	ags9_Max_CMaxI_Days	Max calc. on last 9 mths on max Cust. days for Ins product
99	ags9_Max_CMaxA_Days	Max calc. on last 9 mths on max Cust. days for all product
100	ags9_Max_CMaxC_Due	Max calc. on last 9 mths on max Cust. due for Css product
101	ags9_Max_CMaxI_Due	Max calc. on last 9 mths on max Cust. due for Ins product

Źródło: opracowanie własne.

Tabela 39: Wszystkie zmienne ABT, część 5

Nr	Nazwa zmiennej	Opis (w języku angielskim)
102	ags9_Max_CMaxA_Due	Max calc. on last 9 mths on max Cust. due for all product
103	agr9_Max_CMaxC_Days	Max calc. on last 9 mths on unmissing max Cust. days for Css product
104	agr9_Max_CMaxI_Days	Max calc. on last 9 mths on unmissing max Cust. days for Ins product
105	agr9_Max_CMaxA_Days	Max calc. on last 9 mths on unmissing max Cust. days for all product
106	agr9_Max_CMaxC_Due	Max calc. on last 9 mths on unmissing max Cust. due for Css product
107	agr9_Max_CMaxI_Due	Max calc. on last 9 mths on unmissing max Cust. due for Ins product
108	agr9_Max_CMaxA_Due	Max calc. on last 9 mths on unmissing max Cust. due for all product
109	ags12_Mean_CMaxC_Days	Mean calc. on last 12 mths on max Cust. days for Css product
110	ags12_Mean_CMaxI_Days	Mean calc. on last 12 mths on max Cust. days for Ins product
111	ags12_Mean_CMaxA_Days	Mean calc. on last 12 mths on max Cust. days for all product
112	ags12_Mean_CMaxC_Due	Mean calc. on last 12 mths on max Cust. due for Css product
113	ags12_Mean_CMaxI_Due	Mean calc. on last 12 mths on max Cust. due for Ins product
114	ags12_Mean_CMaxA_Due	Mean calc. on last 12 mths on max Cust. due for all product
115	agr12_Mean_CMaxC_Days	Mean calc. on last 12 mths on unmissing max Cust. days for Css product
116	agr12_Mean_CMaxI_Days	Mean calc. on last 12 mths on unmissing max Cust. days for Ins product
117	agr12_Mean_CMaxA_Days	Mean calc. on last 12 mths on unmissing max Cust. days for all product
118	agr12_Mean_CMaxC_Due	Mean calc. on last 12 mths on unmissing max Cust. due for Css product
119	agr12_Mean_CMaxI_Due	Mean calc. on last 12 mths on unmissing max Cust. due for Ins product
120	agr12_Mean_CMaxA_Due	Mean calc. on last 12 mths on unmissing max Cust. due for all product
121	ags3_Mean_CMaxC_Days	Mean calc. on last 3 mths on max Cust. days for Css product
122	ags3_Mean_CMaxI_Days	Mean calc. on last 3 mths on max Cust. days for Ins product

Źródło: opracowanie własne.

Tabela 40: Wszystkie zmienne ABT, część 6

Nr	Nazwa zmiennej	Opis (w języku angielskim)
123	ags3_Mean_CMaxA_Days	Mean calc. on last 3 mths on max Cust. days for all product
124	ags3_Mean_CMaxC_Due	Mean calc. on last 3 mths on max Cust. due for Css product
125	ags3_Mean_CMaxI_Due	Mean calc. on last 3 mths on max Cust. due for Ins product
126	ags3_Mean_CMaxA_Due	Mean calc. on last 3 mths on max Cust. due for all product
127	agr3_Mean_CMaxC_Days	Mean calc. on last 3 mths on unmissing max Cust. days for Css product
128	agr3_Mean_CMaxI_Days	Mean calc. on last 3 mths on unmissing max Cust. days for Ins product
129	agr3_Mean_CMaxA_Days	Mean calc. on last 3 mths on unmissing max Cust. days for all product
130	agr3_Mean_CMaxC_Due	Mean calc. on last 3 mths on unmissing max Cust. due for Css product
131	agr3_Mean_CMaxI_Due	Mean calc. on last 3 mths on unmissing max Cust. due for Ins product
132	agr3_Mean_CMaxA_Due	Mean calc. on last 3 mths on unmissing max Cust. due for all product
133	ags6_Mean_CMaxC_Days	Mean calc. on last 6 mths on max Cust. days for Css product
134	ags6_Mean_CMaxI_Days	Mean calc. on last 6 mths on max Cust. days for Ins product
135	ags6_Mean_CMaxA_Days	Mean calc. on last 6 mths on max Cust. days for all product
136	ags6_Mean_CMaxC_Due	Mean calc. on last 6 mths on max Cust. due for Css product
137	ags6_Mean_CMaxI_Due	Mean calc. on last 6 mths on max Cust. due for Ins product
138	ags6_Mean_CMaxA_Due	Mean calc. on last 6 mths on max Cust. due for all product
139	agr6_Mean_CMaxC_Days	Mean calc. on last 6 mths on unmissing max Cust. days for Css product
140	agr6_Mean_CMaxI_Days	Mean calc. on last 6 mths on unmissing max Cust. days for Ins product
141	agr6_Mean_CMaxA_Days	Mean calc. on last 6 mths on unmissing max Cust. days for all product
142	agr6_Mean_CMaxC_Due	Mean calc. on last 6 mths on unmissing max Cust. due for Css product
143	agr6_Mean_CMaxI_Due	Mean calc. on last 6 mths on unmissing max Cust. due for Ins product
144	agr6_Mean_CMaxA_Due	Mean calc. on last 6 mths on unmissing max Cust. due for all product
145	ags9_Mean_CMaxC_Days	Mean calc. on last 9 mths on max Cust. days for Css product
146	ags9_Mean_CMaxI_Days	Mean calc. on last 9 mths on max Cust. days for Ins product
147	ags9_Mean_CMaxA_Days	Mean calc. on last 9 mths on max Cust. days for all product
148	ags9_Mean_CMaxC_Due	Mean calc. on last 9 mths on max Cust. due for Css product

Źródło: opracowanie własne.

Tabela 41: Wszystkie zmienne ABT, część 7

Nr	Nazwa zmiennej	Opis (w języku angielskim)
149	ags9_Mean_CMaxI_Due	Mean calc. on last 9 mths on max Cust. due for Ins product
150	ags9_Mean_CMaxA_Due	Mean calc. on last 9 mths on max Cust. due for all product
151	agr9_Mean_CMaxC_Days	Mean calc. on last 9 mths on unmissing max Cust. days for Css product
152	agr9_Mean_CMaxI_Days	Mean calc. on last 9 mths on unmissing max Cust. days for Ins product
153	agr9_Mean_CMaxA_Days	Mean calc. on last 9 mths on unmissing max Cust. days for all product
154	agr9_Mean_CMaxC_Due	Mean calc. on last 9 mths on unmissing max Cust. due for Css product
155	agr9_Mean_CMaxI_Due	Mean calc. on last 9 mths on unmissing max Cust. due for Ins product
156	agr9_Mean_CMaxA_Due	Mean calc. on last 9 mths on unmissing max Cust. due for all product
157	ags12_Min_CMaxC_Days	Min calc. on last 12 mths on max Cust. days for Css product
158	ags12_Min_CMaxI_Days	Min calc. on last 12 mths on max Cust. days for Ins product
159	ags12_Min_CMaxA_Days	Min calc. on last 12 mths on max Cust. days for all product
160	ags12_Min_CMaxC_Due	Min calc. on last 12 mths on max Cust. due for Css product
161	ags12_Min_CMaxI_Due	Min calc. on last 12 mths on max Cust. due for Ins product
162	ags12_Min_CMaxA_Due	Min calc. on last 12 mths on max Cust. due for all product
163	agr12_Min_CMaxC_Days	Min calc. on last 12 mths on unmissing max Cust. days for Css product
164	agr12_Min_CMaxI_Days	Min calc. on last 12 mths on unmissing max Cust. days for Ins product
165	agr12_Min_CMaxA_Days	Min calc. on last 12 mths on unmissing max Cust. days for all product
166	agr12_Min_CMaxC_Due	Min calc. on last 12 mths on unmissing max Cust. due for Css product
167	agr12_Min_CMaxI_Due	Min calc. on last 12 mths on unmissing max Cust. due for Ins product
168	agr12_Min_CMaxA_Due	Min calc. on last 12 mths on unmissing max Cust. due for all product
169	ags3_Min_CMaxC_Days	Min calc. on last 3 mths on max Cust. days for Css product
170	ags3_Min_CMaxI_Days	Min calc. on last 3 mths on max Cust. days for Ins product
171	ags3_Min_CMaxA_Days	Min calc. on last 3 mths on max Cust. days for all product
172	ags3_Min_CMaxC_Due	Min calc. on last 3 mths on max Cust. due for Css product
173	ags3_Min_CMaxI_Due	Min calc. on last 3 mths on max Cust. due for Ins product
174	ags3_Min_CMaxA_Due	Min calc. on last 3 mths on max Cust. due for all product

Źródło: opracowanie własne.

Tabela 42: Wszystkie zmienne ABT, część 8

Nr	Nazwa zmiennej	Opis (w języku angielskim)
175	agr3_Min_CMaxC_Days	Min calc. on last 3 mths on unmissing max Cust. days for Css product
176	agr3_Min_CMaxI_Days	Min calc. on last 3 mths on unmissing max Cust. days for Ins product
177	agr3_Min_CMaxA_Days	Min calc. on last 3 mths on unmissing max Cust. days for all product
178	agr3_Min_CMaxC_Due	Min calc. on last 3 mths on unmissing max Cust. due for Css product
179	agr3_Min_CMaxI_Due	Min calc. on last 3 mths on unmissing max Cust. due for Ins product
180	agr3_Min_CMaxA_Due	Min calc. on last 3 mths on unmissing max Cust. due for all product
181	ags6_Min_CMaxC_Days	Min calc. on last 6 mths on max Cust. days for Css product
182	ags6_Min_CMaxI_Days	Min calc. on last 6 mths on max Cust. days for Ins product
183	ags6_Min_CMaxA_Days	Min calc. on last 6 mths on max Cust. days for all product
184	ags6_Min_CMaxC_Due	Min calc. on last 6 mths on max Cust. due for Css product
185	ags6_Min_CMaxI_Due	Min calc. on last 6 mths on max Cust. due for Ins product
186	ags6_Min_CMaxA_Due	Min calc. on last 6 mths on max Cust. due for all product
187	agr6_Min_CMaxC_Days	Min calc. on last 6 mths on unmissing max Cust. days for Css product
188	agr6_Min_CMaxI_Days	Min calc. on last 6 mths on unmissing max Cust. days for Ins product
189	agr6_Min_CMaxA_Days	Min calc. on last 6 mths on unmissing max Cust. days for all product
190	agr6_Min_CMaxC_Due	Min calc. on last 6 mths on unmissing max Cust. due for Css product
191	agr6_Min_CMaxI_Due	Min calc. on last 6 mths on unmissing max Cust. due for Ins product
192	agr6_Min_CMaxA_Due	Min calc. on last 6 mths on unmissing max Cust. due for all product
193	ags9_Min_CMaxC_Days	Min calc. on last 9 mths on max Cust. days for Css product
194	ags9_Min_CMaxI_Days	Min calc. on last 9 mths on max Cust. days for Ins product
195	ags9_Min_CMaxA_Days	Min calc. on last 9 mths on max Cust. days for all product
196	ags9_Min_CMaxC_Due	Min calc. on last 9 mths on max Cust. due for Css product
197	ags9_Min_CMaxI_Due	Min calc. on last 9 mths on max Cust. due for Ins product
198	ags9_Min_CMaxA_Due	Min calc. on last 9 mths on max Cust. due for all product
199	agr9_Min_CMaxC_Days	Min calc. on last 9 mths on unmissing max Cust. days for Css product

Źródło: opracowanie własne.

Tabela 43: Wszystkie zmienne ABT, część 9

Nr	Nazwa zmiennej	Opis (w języku angielskim)
200	agr9_Min_CMaxI_Days	Min calc. on last 9 mths on unmissing max Cust. days for Ins product
201	agr9_Min_CMaxA_Days	Min calc. on last 9 mths on unmissing max Cust. days for all product
202	agr9_Min_CMaxC_Due	Min calc. on last 9 mths on unmissing max Cust. due for Css product
203	agr9_Min_CMaxI_Due	Min calc. on last 9 mths on unmissing max Cust. due for Ins product
204	agr9_Min_CMaxA_Due	Min calc. on last 9 mths on unmissing max Cust. due for all product

Źródło: opracowanie własne.

5.2. Dokumentacje modeli ocen punktowych

Wszystkie modele kart skoringowych zostały wykonane przez autora książki na podstawie własnych algorytmów w SAS 4GL. Budowane są metodą LOG opisaną w podrozdziałach 4.2.4 i 2.2.3. Modele te służą tylko jako przykłady. Posiadają dobre własności, aczkolwiek powinno się jeszcze poprawić niektóre reguły i być może wymienić pewne zmienne. Dla wielu z nich powinniśmy otrzymać relacje monotoniczne pomiędzy wzrastaniem ich argumentów a wzrastaniem lub zmniejszaniem się oceny cząstkowej. Brak takiej relacji oznacza, że w naturalny sposób algorytm tworzący atrybuty nie identyfikuje takiej relacji, gdyż głównym jego celem jest podzielenie na kategorie jednorodne różniące się pomiędzy sobą wartościami ryzyka.

W dokumentacji, oprócz karty skoringowej, prezentowane są podstawowe mierniki modeli: Gini i LiftX; patrz podrozdział 4.2.5. Zauważmy, że statystyki LiftX przy modelach dla portfela gotówkowego (PD Css i Cross PD Css) mają małe wartości. Wynika to z bardzo dużego ogólnego ryzyka tych portfeli.

Modele PD są zbudowane z funkcją celu $\mathrm{default}_{12}$, natomiast model PR ze zdarzeniem response w okresie 6 miesięcy obserwacji.

5.2.1. Model ryzyka dla kredytu ratalnego (PD Ins)

Podstawowe informacje o modelu przedstawione są w tabeli 44.

Wartość PD_Ins kalibrowana jest w następujący sposób:

```
pd_ins=1/(1+exp(-(-0.032205144
       *risk_ins_score+9.4025558419)));
```

5.2.2. Model ryzyka dla kredytu gotówkowego (PD Css)

Podstawowe informacje o modelu przedstawione są w tabeli 45.

Wartość PD_Css kalibrowana jest w następujący sposób:

```
pd_css=1/(1+exp(-(-0.028682728
       *risk_css_score+8.1960829753)));
```

5.2.3. Model ryzyka dla kredytu gotówkowego w momencie aplikowania o kredyt ratalny (Cross PD Css)

Podstawowe informacje o modelu przedstawione są w tabeli 46.

Wartość Cross_PD_Css kalibrowana jest w następujący sposób:

```
cross_pd_css=1/(1+exp(-(-0.028954669
       *cross_css_score+8.2497434934)));
```

5.2.4. Model skłonności skorzystania z kredytu gotówkowego w momencie aplikowania o kredyt ratalny (PR Css)

Podstawowe informacje o modelu przedstawione są w tabeli 47.

Wartość PR_Css kalibrowana jest w następujący sposób:

```
pr_css=1/(1+exp(-(-0.035007455
       *response_score+10.492092793)));
```

Tabela 44: Model PD Ins

Gini w % na zbiorze		Lift1	Lift5	Lift10	Lift20
treningowym	walidacyjnym				
73,37	73,37	7,62	5,59	4,52	3,34

Karta skoringowa

Zmienna	Warunek	Ocena cząstkowa
ACT_CC	1.0535455861 $<$ ACT_CC	–1
	0.857442348 $<$ ACT_CC $\leqslant$ 1.0535455861	29
	0.3324658426 $<$ ACT_CC $\leqslant$ 0.857442348	40
	0.248125937 $<$ ACT_CC $\leqslant$ 0.3324658426	49
	ACT_CC $\leqslant$ 0.248125937	61
ACT_CINS_MIN_SENIORITY	ACT_CINS_MIN_SENIORITY $\leqslant$ 22	–1
	22 $<$ ACT_CINS_MIN_SENIORITY $\leqslant$ 36	50
	Missing	53
	36 $<$ ACT_CINS_MIN_SENIORITY $\leqslant$ 119	76
	119 $<$ ACT_CINS_MIN_SENIORITY	99
ACT_CINS_N_LOAN	1 $<$ ACT_CINS_N_LOAN	–1
	ACT_CINS_N_LOAN $\leqslant$ 1	57
ACT_CINS_N_STATC	ACT_CINS_N_STATC $\leqslant$ 0	–1
	0 $<$ ACT_CINS_N_STATC $\leqslant$ 1	49
	Missing	49
	1 $<$ ACT_CINS_N_STATC $\leqslant$ 2	54
	2 $<$ ACT_CINS_N_STATC	87
APP_CHAR_JOB_CODE	Contract	–1
	Owner company	58
	Retired	76
	Permanent	81
APP_CHAR_MARITAL_STATUS	Single	–1
	Divorced	40
	Married	55
	Widowed	57
APP_LOAN_AMOUNT	11376 $<$ APP_LOAN_AMOUNT	–1
	8880 $<$ APP_LOAN_AMOUNT $\leqslant$ 11376	21
	7656 $<$ APP_LOAN_AMOUNT $\leqslant$ 8880	30
	4824 $<$ APP_LOAN_AMOUNT $\leqslant$ 7656	35
	1920 $<$ APP_LOAN_AMOUNT $\leqslant$ 4824	51
	APP_LOAN_AMOUNT $\leqslant$ 1920	57
APP_NUMBER_OF_CHILDREN	APP_NUMBER_OF_CHILDREN $\leqslant$ 0	–1
	0 $<$ APP_NUMBER_OF_CHILDREN $\leqslant$ 1	23
	1 $<$ APP_NUMBER_OF_CHILDREN	57

Źródło: opracowanie własne.

Tabela 45: Model PD Css

Gini w % na zbiorze		Lift1	Lift5	Lift10	Lift20
treningowym	walidacyjnym				
74,06	74,06	1,77	1,67	1,63	1,64

Karta skoringowa

Zmienna	Warunek	Ocena cząstkowa
ACT_AGE	$50 < \text{ACT\_AGE} \leqslant 61$	24
	$62 < \text{ACT\_AGE} \leqslant 68$	34
	$\text{ACT\_AGE} \leqslant 50$	33
	$68 < \text{ACT\_AGE} \leqslant 80$	44
	$61 < \text{ACT\_AGE} \leqslant 62$	46
	$80 < \text{ACT\_AGE}$	70
ACT_CALL_CC	$1.5775700935 < \text{ACT\_CALL\_CC} \leqslant 2.0091145833$	24
	$2.0091145833 < \text{ACT\_CALL\_CC}$	34
	$1.4502074689 < \text{ACT\_CALL\_CC} \leqslant 1.5775700935$	43
	$1.1900674433 < \text{ACT\_CALL\_CC} \leqslant 1.4502074689$	51
	$\text{ACT\_CALL\_CC} \leqslant 1.1900674433$	64
ACT_CCSS_DUEUTL	$0.0416666667 < \text{ACT\_CCSS\_DUEUTL} \leqslant 0.21875$	24
	$0.21875 < \text{ACT\_CCSS\_DUEUTL}$	29
	$0.025 < \text{ACT\_CCSS\_DUEUTL} \leqslant 0.0416666667$	33
	$0.0208333333 < \text{ACT\_CCSS\_DUEUTL} \leqslant 0.025$	41
	$\text{ACT\_CCSS\_DUEUTL} \leqslant 0.0208333333$	54
	Missing	57
ACT_CCSS_MIN_LNINST	$1 < \text{ACT\_CCSS\_MIN\_LNINST} \leqslant 7$	24
	$7 < \text{ACT\_CCSS\_MIN\_LNINST} \leqslant 11$	26
	$\text{ACT\_CCSS\_MIN\_LNINST} \leqslant 0$	31
	$11 < \text{ACT\_CCSS\_MIN\_LNINST}$	32
	$0 < \text{ACT\_CCSS\_MIN\_LNINST} \leqslant 1$	40
	Missing	47
ACT_CCSS_N_STATC	$0 < \text{ACT\_CCSS\_N\_STATC} \leqslant 4$	24
	$\text{ACT\_CCSS\_N\_STATC} \leqslant 0$	32
	$4 < \text{ACT\_CCSS\_N\_STATC} \leqslant 10$	34
	$10 < \text{ACT\_CCSS\_N\_STATC} \leqslant 21$	51
	Missing	53
	$21 < \text{ACT\_CCSS\_N\_STATC}$	82
AGS3_MEAN_CMAXA_DUE	$1.333 < \text{AGS3\_MEAN\_CMAXA\_DUE}$	24
	$0.666 < \text{AGS3\_MEAN\_CMAXA\_DUE} \leqslant 1.333$	47
	$0.333 < \text{AGS3\_MEAN\_CMAXA\_DUE} \leqslant 0.666$	62
	$\text{AGS3\_MEAN\_CMAXA\_DUE} \leqslant 0.333$	73
APP_NUMBER_OF_CHILDREN	$\text{APP\_NUMBER\_OF\_CHILDREN} \leqslant 0$	24
	$0 < \text{APP\_NUMBER\_OF\_CHILDREN} \leqslant 1$	33
	$1 < \text{APP\_NUMBER\_OF\_CHILDREN}$	57

Źródło: opracowanie własne.

Tabela 46: Model Cross PD Css

Gini w % na zbiorze		Lift1	Lift5	Lift10	Lift20
treningowym	walidacyjnym				
73,77	73,77	1,80	1,50	1,52	1,48

Karta skoringowa

Zmienna	Warunek	Ocena cząstkowa
ACT12_N_GOOD_DAYS	$4 < \text{ACT12\_N\_GOOD\_DAYS} \leqslant 8$	29
	$3 < \text{ACT12\_N\_GOOD\_DAYS} \leqslant 4$	34
	$8 < \text{ACT12\_N\_GOOD\_DAYS}$	34
	$2 < \text{ACT12\_N\_GOOD\_DAYS} \leqslant 3$	37
	$\text{ACT12\_N\_GOOD\_DAYS} \leqslant 2$	45
	Missing	54
ACT_CCSS_MAXDUE	$1 < \text{ACT\_CCSS\_MAXDUE} \leqslant 4$	29
	$4 < \text{ACT\_CCSS\_MAXDUE}$	37
	$0 < \text{ACT\_CCSS\_MAXDUE} \leqslant 1$	45
	$\text{ACT\_CCSS\_MAXDUE} \leqslant 0$	59
	Missing	71
ACT_CCSS_N_STATC	$\text{ACT\_CCSS\_N\_STATC} \leqslant 4$	29
	$4 < \text{ACT\_CCSS\_N\_STATC} \leqslant 6$	40
	$6 < \text{ACT\_CCSS\_N\_STATC} \leqslant 15$	54
	Missing	79
	$15 < \text{ACT\_CCSS\_N\_STATC} \leqslant 26$	85
	$26 < \text{ACT\_CCSS\_N\_STATC}$	125
ACT_CCSS_UTL	$\text{ACT\_CCSS\_UTL} \leqslant 0.4083333333$	29
	$0.4083333333 < \text{ACT\_CCSS\_UTL} \leqslant 0.4479166667$	32
	$0.4479166667 < \text{ACT\_CCSS\_UTL} \leqslant 0.4895833333$	36
	$0.4895833333 < \text{ACT\_CCSS\_UTL} \leqslant 0.5208333333$	40
	$0.5208333333 < \text{ACT\_CCSS\_UTL} \leqslant 0.5347222222$	44
	Missing	57
	$0.5347222222 < \text{ACT\_CCSS\_UTL}$	58
AGS3_MEAN_CMAXA_DUE	$1 < \text{AGS3\_MEAN\_CMAXA\_DUE} \leqslant 3$	29
	$3 < \text{AGS3\_MEAN\_CMAXA\_DUE}$	33
	$0.6666666667 < \text{AGS3\_MEAN\_CMAXA\_DUE} \leqslant 1$	39
	$\text{AGS3\_MEAN\_CMAXA\_DUE} \leqslant 0.6666666667$	49
	Missing	60
APP_INCOME	$\text{APP\_INCOME} \leqslant 411$	29
	$411 < \text{APP\_INCOME} \leqslant 573$	42
	$3872 < \text{APP\_INCOME}$	52
	$573 < \text{APP\_INCOME} \leqslant 1049$	60
	$1049 < \text{APP\_INCOME} \leqslant 3872$	77

Źródło: opracowanie własne.

Tabela 47: Model PR Css

Gini w % na zbiorze		Lift1	Lift5	Lift10	Lift20
treningowym	walidacyjnym				
86,22	86,22	4,36	3,03	2,79	2,60

Karta skoringowa

Zmienna	Warunek	Ocena cząstkowa
ACT_CCSS_MIN_SENIORITY	ACT_CCSS_MIN_SENIORITY $\leqslant 4$	51
	$4 <$ ACT_CCSS_MIN_SENIORITY $\leqslant 7$	60
	$30 <$ ACT_CCSS_MIN_SENIORITY	64
	$7 <$ ACT_CCSS_MIN_SENIORITY $\leqslant 9$	65
	$9 <$ ACT_CCSS_MIN_SENIORITY $\leqslant 30$	76
	Missing	96
ACT_CCSS_N_LOAN	$5 <$ ACT_CCSS_N_LOAN	51
	$4 <$ ACT_CCSS_N_LOAN $\leqslant 5$	59
	$2 <$ ACT_CCSS_N_LOAN $\leqslant 4$	80
	$1 <$ ACT_CCSS_N_LOAN $\leqslant 2$	108
	ACT_CCSS_N_LOAN $\leqslant 0$	124
	$0 <$ ACT_CCSS_N_LOAN $\leqslant 1$	131
ACT_CCSS_N_STATC	$12 <$ ACT_CCSS_N_STATC	51
	$7 <$ ACT_CCSS_N_STATC $\leqslant 12$	60
	$3 <$ ACT_CCSS_N_STATC $\leqslant 7$	68
	ACT_CCSS_N_STATC $\leqslant 3$	78
	Missing	106
ACT_CINS_N_STATC	$5 <$ ACT_CINS_N_STATC	51
	$3 <$ ACT_CINS_N_STATC $\leqslant 5$	54
	$1 <$ ACT_CINS_N_STATC $\leqslant 3$	59
	$0 <$ ACT_CINS_N_STATC $\leqslant 1$	64
	ACT_CINS_N_STATC $\leqslant 0$	75
	Missing	103

Źródło: opracowanie własne.

5.3. Parametry modelu biznesowego: akwizycja – sprzedaż krzyżowa

W podrozdziałach od 5.3.1 do 5.3.4 przedstawione są konkretne algorytmy tworzące przykładowy zestaw danych losowych. Niektóre z nich najwygodniej jest przedstawić w napisanych kodach języka SAS 4GL.

5.3.1. Parametry ogólne

Na wstępnie należy podać metodę użycia generatorów liczb losowych: funkcja `ranuni(&seed)` zwracająca liczbę losową z rozkładu jednostajnego z przedziału $(0, 1)$ oraz `rannor(&seed)` zwracająca liczbę losową z rozkładu normalnego standaryzowanego, czyli mniej więcej z przedziału $(-4, 4)$. Parametr `&seed` jest ziarnem generatora, powodującym, że liczby losowe będą się tworzyć, zaczynając zawsze od tej samej wartości. Wielokrotne uruchomienie kodu daje ten sam zestaw danych. Natomiast zmiana ziarna powoduje stworzenie już innego układu.

**** SAS 4GL ****
```
%let seed=1;
%let n_day=4; – średnia liczba wniosków na każdy dzień produkcji
%let percent_dec=1.1; – procent powiększenia produkcji grudniowej
%let n_terms=3; – parametr cyklu koniunkturalnego
%let n_terms_vars=2; – drugi parametr cyklu koniunkturalnego
%let s_date='01jan1970'd; – data początku produkcji
%let e_date='31jan2001'd; – data końca produkcji
%let percent_repeat=0.2; – szansa powtórzonego kredytu ratalnego
dla tego samego klienta
```
*** SAS 4GL ****

5.3.2. Parametry poziomu klienta

Zmienne aplikacyjne mają słowniki określone procedurą format:

**** SAS 4GL ****
```
proc format;
value jobc – słownik kodu zawodu
1 = 'Contract'
2 = 'Owner company'
3 = 'Permanent'
4 = 'Retired';
```

```
value martials – słownik statusu małżeńskiego
1 = 'Single'
2 = 'Divorced'
3 = 'Widowed'
4 = 'Married';
value homes – słownik statusu mieszkaniowego
1 = 'With parents'
2 = 'Rental'
3 = 'Owner';
value city – słownik typu miasta
1 = 'Small'
2 = 'Medium'
3 = 'Big'
4 = 'Large';
value cars – słownik posiadania samochodu
1 = 'No'
2 = 'Owner';
value gender
1 = 'Male' – słownik płci
2 = 'Female';
value branch – słownik branży
1 = 'Computers'
2 = 'Radio-TV'
3 = 'Fenitures'
4 = 'DiY'
other='Empty';
run;
*** SAS 4GL ****
```

Najważniejsze tworzenie podstawowych struktur danych, kredytu ratalnego i danych klientów wykonywane jest przez kilka datastepów (są to podstawowe bloki języka 4GL).

```
**** SAS 4GL ****
data data.clients1(keep=cid data_of_birth gender)
 data.production1(keep=aid cid app_date period
 data_of_birth installment
 n_installments loan_amount branch)
 data.production2(keep=aid cid app_date period
 installment n_installments
 loan_amount branch);
length aid $16 cid $10;
n=1;
do app_date=&s_date to &e_date;
   max=&n_day*(1+rannor(&seed)/20);
```

```
  if month(app_date)=12 then max=max*&percent_dec;
  period=put(app_date,yymmn6.);
  term=intck('month',&s_date,&e_date)+1;
  j=intck('month',&s_date,input(period,yymmn6.));
  pr=(0.01+(1.5+sin(&n_terms_vars*3.1412*j/term)
    +rannor(&seed)/5)/8)/0.36;
  do i=1 to max;
    aid='ins'||put(app_date,yymmddn8.)
      ||put(i,z4.)||'1'; – tworzenie identyfikatora rachunku
    x=int((75-18)*(rannor(&seed)+4)/7 + 10 +
      20*ranuni(&seed)+5*pr); – zmienna potrzebna do
```
wyznaczenia wieku klienta
```
    if x<18 then x=18;
    if x>75 then x=75;
    data_of_birth=int(app_date-x*365.5);
    cid=put(n,z10.); – tworzenie identyfikatora klienta
    gender=(ranuni(&seed)>(0.45+pr/20))+1;
    if gender>2 then gender=2;
    if gender<1 then gender=1;
    installment=int(abs(rannor(&seed))*200+60+50*pr);
    n_installments=12;
    if ranuni(&seed)<(0.3+pr/50)
      then n_installments=24;
    if ranuni(&seed)<(0.2-pr/50)
      then n_installments=36;
    loan_amount=n_installments*installment;
    branch=int(ranuni(&seed)*3+1.5+pr/10);
    if branch>4 then branch=4;
    if branch<1 then branch=1;
    output data.production1;
    output data.clients1;
    if ranuni(&seed)<&percent_repeat then do; – 20 razy na
```
100 warunek jest spełniony i tworzone są dodatkowe wnioski klientów
```
      aid='ins'||put(app_date,yymmddn8.)
        ||put(i,z4.)||'2';
      m=int(n*abs(rannor(&seed))/5);
      if m<1 then m=1;
      if m>n then m=n;
      cid=put(m,z10.);
      installment=int(abs(rannor(&seed))*200+60);
      n_installments=12;
      if ranuni(&seed)<0.3 then n_installments=24;
      if ranuni(&seed)<0.2 then n_installments=36;
      loan_amount=n_installments*installment;
      branch=int(ranuni(&seed)*3+1.5);
```

```
      if branch>4 then branch=4;
      if branch<1 then branch=1;
      output data.production2;
    end;
    if ranuni(&seed)<&percent_repeat then do;
      aid='ins'||put(app_date,yymmddn8.)
        ||put(i,z4.)||'3';
      m=int(n*abs(rannor(&seed))/5);
      if m<1 then m=1;
      if m>n then m=n;
      cid=put(m,z10.);
      installment=int(abs(rannor(&seed))*200+60);
      n_installments=12;
      if ranuni(&seed)<0.3 then n_installments=24;
      if ranuni(&seed)<0.2 then n_installments=36;
      loan_amount=n_installments*installment;
      branch=int(ranuni(&seed)*3+1.5);
      if branch>4 then branch=4;
      if branch<1 then branch=1;
      output data.production2;
    end;
    n=n+1;
  end;
end;
format app_date data_of_birth yymmdd10. ;
run;
*** SAS 4GL ****
```

Drugi data-step tworzy historię zmian cech klientów, zakładając, że zmiany dokonują się raz na rok.

```
**** SAS 4GL ****
%let income_i_spendings=%str(
income=int((5000-500)*abs(rannor(&seed))/4+500);
if job_code=3 then
income=int((7000-1500)*abs(rannor(&seed))/4+1500);
if job_code=4 then
income=int((4000-300)*abs(rannor(&seed))/4+300);
if job_code=2 then
income=int((17000-3000)*abs(rannor(&seed))/4+3000);
spendings=20*int(income*(abs(rannor(&seed))
  +home_status+cars-2)/(8*20));
```

`);` – ciąg instrukcji powodujących ponowne wyznaczenie wynagrodzenia zależnego od kodu zawodu oraz przeliczenie wydatków

```
data data.clients_all;
set data.clients1;
year_s=year(&s_date);
year_e=year(&e_date);
yearb=year(data_of_birth);
do year=yearb+18 to year_e; – pętla rozpoczyna się od 18 lat
  age=year-yearb;
  if age=18 then do; – ustawianie początkowych cech klienta w wieku
18 lat
    job_code=1+(ranuni(&seed)<0.4)*2;
    number_of_children=0;
    marital_status=1;
    city=int(ranuni(&seed)*3+1.5);
    if city>4 then city=4;
    if city<1 then city=1;
    home_status=int(ranuni(&seed)*2+1.5);
    if home_status>3 then home_status=3;
    if home_status<1 then home_status=1;
    cars=1;
    &income_i_spendings;
  end; else do;
    if marital_status=1 and age<60 and
      ranuni(&seed)<0.1 then
      marital_status=4; – warunki zmiany statusu małeńskiego
    if number_of_children<1 and marital_status=4
      and ranuni(&seed)<0.1 and age<45 then
        number_of_children=number_of_children+1;
– warunki na zmianę liczby dzieci
    if number_of_children=1 and marital_status=4
      and ranuni(&seed)<0.05 and age<45 then
        number_of_children=number_of_children+1;
    if number_of_children=2 and marital_status=4
      and ranuni(&seed)<0.01 and age<45 then
        number_of_children=number_of_children+1;
    if number_of_children>0 and age>45
      and ranuni(&seed)<0.1 then
      number_of_children=number_of_children-1;
    if marital_status=4 and ranuni(&seed)<0.01
      then marital_status=2;
    if marital_status=4 and age>60
      and ranuni(&seed)<0.1
      then marital_status=3;
    if (marital_status=4 or age>25) and home_status=1
      and ranuni(&seed)<0.7 then home_status=2;
    if (marital_status=4 or age>25) and home_status=1
```

```
        and ranuni(&seed)<0.2 then home_status=3;
      if home_status=2 and ranuni(&seed)<0.05
        then home_status=3;
      if ranuni(&seed)<0.005 then do;
        city=int(ranuni(&seed)*3+1.5);
      if city>4 then city=4;
      if city<1 then city=1;
      end;
      if cars=1 and 20<age<=60 and ranuni(&seed)<0.05
        then cars=2;
      if cars=2 and ranuni(&seed)<0.001 then cars=1;
      if job_code ne 4 and age>50 and ranuni(&seed)<0.1
      then do;  – warunki na zmiany kodów zawodów
        job_code=4;
        &income_i_spendings;
      end;
      if job_code ne 4 and age>70 then do;
        job_code=4;
        &income_i_spendings;
      end;
      if job_code=1 and ranuni(&seed)<0.05 then do;
        job_code=3;
        &income_i_spendings;
      end;
      if job_code in (3,1) and ranuni(&seed)<0.01
      then do;
        job_code=2;
        &income_i_spendings;
      end;
      if job_code=2 and ranuni(&seed)<0.01 then do;
        job_code=3;
        &income_i_spendings;
      end;
      if job_code in (3,2) and ranuni(&seed)<0.005
      then do;
        job_code=1;
        &income_i_spendings;
      end;
    end;
    if year>=year_s then output;
end;
drop
year_s year_e yearb;
run;
*** SAS 4GL ****
```

Należy podkreślić, że wszystkie parametry zostały ustalone ekspercko, dlatego nie wydaje się słuszne omawianie każdej linii kodu szczegółowo. Istnieje nadzieja, że przyszłe projekty naukowe uzupełnią model o szczegółowe obliczenia współczynników, czyniąc je zgodnymi z rzeczywistością.

5.3.3. Parametry kredytu ratalnego

Macierz wykorzystana do kroków iteracyjnych:

$$M_{ij} = \begin{bmatrix} & j=0 & j=1 & j=2 & j=3 & j=4 & j=5 & j=6 & j=7 \\ i=0 & 0{,}970 & 0{,}030 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 \\ i=1 & 0{,}200 & 0{,}650 & 0{,}150 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 \\ i=2 & 0{,}040 & 0{,}240 & 0{,}190 & 0{,}530 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 \\ i=3 & 0{,}005 & 0{,}025 & 0{,}080 & 0{,}100 & 0{,}790 & 0{,}000 & 0{,}000 & 0{,}000 \\ i=4 & 0{,}000 & 0{,}000 & 0{,}010 & 0{,}080 & 0{,}090 & 0{,}820 & 0{,}000 & 0{,}000 \\ i=5 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}020 & 0{,}030 & 0{,}950 & 0{,}000 \\ i=6 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}010 & 0{,}010 & 0{,}980 \end{bmatrix}$$

W każdym kroku iteracyjnym obliczane są dwa skoringi (równania 1.1 i 1.2, ze strony 41). Im większa wartość oceny punktowej, tym mniejsze ryzyko dla danego rachunku.

Ocena punktowa do segmentacji na podportfele bardziej i mniej wrażliwe na cykl koniunkturalny jest obliczana jako suma poniższych składowych:

```
score_cycle=sum(
1*app_income,
1*app_nom_branch,
1*app_nom_gender,
2*app_nom_job_code,
1*app_number_of_children,
1*app_nom_marital_status,
1*app_nom_city,
1*app_nom_home_status,
1*app_nom_cars),
```

gdzie punkt odcięcia jest równy 0 po wykonaniu standaryzacji otrzymanej oceny punktowej. Innymi słowy: rachunki z większą od zera wartością standaryzowanej oceny punktowej są bardziej podatne na

koniunkturę. Z doświadczeń autora wynika, że jeśli w ocenie punktowej jest wziętych kilka cech, to oscylacja ryzyka widoczna jest dla każdego podportfela, ekspansja kryzysu dosięga wszystkich klientów. Wybór współczynników do tej oceny punktowej nie jest istotny. Jest tu tylko pokazana możliwość dodatkowego zróżnicowania ryzyka w podgrupach. Dla kredytów gotówkowych używany jest ten sam mechanizm.

Drugi skoring oparty jest na następujących członach (użyte są tu nieco inne oznaczenia zmiennych niż opisane w ABT w podrozdziale 5.1, gdyż kredyty ratalne spłacane są niezależnie od gotówkowych, zatem agregaty na kliencie związane są tylko z jednym produktem):

`score_main=sum(`
`-1*act_cus_utl,` – jak mało spłacone, liczba spłaconych rat do wszystkich rat tego kredytu, to większe ryzyko
`-1*act_cus_dueutl,` – jak duży udział zaległych rat w całej kwocie kredytu, to większe ryzyko
`-1*act_cus_cc,` – im większe obciążenie klienta, suma rat i wydatków do dochodu, tym większe ryzyko
`-1*act_cus_n_loans_act,` – im więcej klient posiada spłacanych kredytów, tym większe ryzyko
`3*act_cus_seniority,` – im dłużej klient posiada kredyty ratalne, tym mniejsze ryzyko
`-5*act_cus_loan_number,` – im większy numer kolejnego kredytu, tym większe ryzyko
`5*(act_cus_loan_number=1),` – dla pierwszego kredytu, promocja, ten kredyt powinien mieć małe ryzyko
`6*act_cus_n_statC,` – im więcej spłaconych kredytów ratalnych klienta (zakończone statusami C) tym mniejsze ryzyko
`-3*act_cus_n_statB,` – im więcej zakończonych statusem B, czyli źle spłaconych, tym większe ryzyko
`1*app_nom_branch,` – mniejsze ryzyko dla wyższego numeru branży, najmniejsze ryzyko dla sklepów z kategorii „Zrób to Sam” (DiY)
`3*app_nom_gender,` – kobiety mają mniejsze ryzyko
`6*app_nom_job_code,` – emeryci mają najmniejsze ryzyko

-3*(app_nom_job_code=4 and
app_nom_marital_status in (2,3)
 and app_nom_gender=2), – potrzebne do interakcji, emerytki rozwiedzione lub owdowiałe mają większe ryzyko
 2*(app_nom_job_code=4 and
app_nom_marital_status in (2,3)
 and app_nom_gender=1), – potrzebne do interakcji, emeryci rozwiedzeni lub owdowiali mają mniejsze ryzyko
 6*app_number_of_children, – im więcej dzieci tym mniejsze ryzyko
 3*app_nom_marital_status, – zamężni lub żonaci mają najmniejsze ryzyko
 1*app_nom_city, – w dużych miastach mniejsze ryzyko
 1*app_nom_home_status, – właściciele mieszkań mają mniejsze ryzyko
 1*app_nom_cars, – właściciele samochodów mają mniejsze ryzyko
-1*app_spendings, – im większe wydatki, tym większe ryzyko
-5*act_days, – im później klient wpłaca ratę, tym większe ryzyko
-4*act_utl, – im mniej spłaconych rat dla tego rachunku, tym większe ryzyko
-6*act_dueutl , – im więcej opóźnionych rat do wszystkich, tym większe ryzyko
-2*act_due, – im więcej opóźnionych rat, tym większe ryzyko
 4*act_age, – im starszy klient, tym mniejsze ryzyko
-2*act_cc, – im większe obciążenie ratami do wynagrodzenia, tym większe ryzyko
-1*act_dueinc, – im większa suma opóźnionych rat do dochodu, tym większe ryzyko
-2*act_loaninc, – im większy stosunek kwota kredytu do dochodu, tym większe ryzyko
 2*app_income, – im większe wynagrodzenie, tym mniejsze ryzyko
-1*app_loan_amount, – im większa kwota kredytu, tym większe ryzyko

`-4*app_n_installments,` – im więcej rat, tym większe ryzyko
`-2*agr3_Mean_Due,` – wszystkie zmienne od tego miejsca i poniżej dotyczą albo danego rachunku, albo wszystkich kredytów ratalnych klienta i wprowadzają zmianę ryzyka w zależności od opóźnień, ogólnie im więcej opóźnień, tym większe ryzyko
`-3*ags3_Mean_Days,`
`-3*agr6_Mean_Due,`
`-3*ags6_Mean_Days,`
`-2*agr9_Mean_Due,`
`-3*ags9_Mean_Days,`
`-2*agr12_Mean_Due,`
`-3*ags12_Mean_Days,`
`-3*ags3_Max_CMax_Due,`
`-2*ags12_Max_CMax_Due,`
`-2*ags9_Max_CMax_Days,`
`-1*act12_n_cus_arrears,`
`-2*ags12_Max_CMax_Due,`
`5*(ags12_Max_CMax_Due=0))` – dodatkowa promocja dla klientów bez złej historii z ostatnich 12 miesięcy, im zmniejszamy ryzyko.

5.3.4. Parametry kredytu gotówkowego

W przypadku kredytu gotówkowego macierz musi być zmieniona, aby uzyskać znacznie większe ryzyko:

$$M_{ij} = \begin{bmatrix} & j=0 & j=1 & j=2 & j=3 & j=4 & j=5 & j=6 & j=7 \\ i=0 & 0{,}850 & 0{,}015 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 \\ i=1 & 0{,}250 & 0{,}450 & 0{,}300 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 \\ i=2 & 0{,}040 & 0{,}240 & 0{,}190 & 0{,}530 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 \\ i=3 & 0{,}005 & 0{,}025 & 0{,}080 & 0{,}100 & 0{,}790 & 0{,}000 & 0{,}000 & 0{,}000 \\ i=4 & 0{,}000 & 0{,}000 & 0{,}010 & 0{,}080 & 0{,}090 & 0{,}820 & 0{,}000 & 0{,}000 \\ i=5 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}020 & 0{,}030 & 0{,}950 & 0{,}000 \\ i=6 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}000 & 0{,}010 & 0{,}010 & 0{,}980 \end{bmatrix}$$

Ocena punktowa determinująca przejścia pomiędzy liczbą zaległych rat dla kredytu gotówkowego jest sumą następujących członów (tym razem nie podano już opisów zmiennych, gdyż są podobne do

sytuacji kredytu ratalnego i opisanych zmiennych ABT w rozdziale 5.1):

```
score_main_css=sum(
-1*act_ins_cus_maxdue,
```

– zmienna liczona na kredytach ratalnych klienta, im więcej opóźnionych rat na jakimkolwiek kredycie ratalnym, tym większe ryzyko kredytu gotówkowego

```
-1*act_cus_utl,
-1*act_cus_dueutl,
-1*act_cus_cc,
-1*act_cus_n_loans_act,
 3*act_cus_seniority,
-5*act_cus_loan_number,
 5*(act_cus_loan_number=1),
 6*act_cus_n_statC,
-2*act_cus_n_statB,
 3*app_nom_gender,
 6*app_nom_job_code,
-3*(app_nom_job_code=4 and
app_nom_marital_status in (2,3)
 and app_nom_gender=2),
 2*(app_nom_job_code=4 and
app_nom_marital_status in (2,3)
 and app_nom_gender=1),
 6*app_number_of_children,
 3*app_nom_marital_status,
 1*app_nom_city,
 1*app_nom_home_status,
 1*app_nom_cars,
-1*app_spendings,
-5*act_days,
-4*act_utl,
-5*act_dueutl,
-2*act_due,
 4*act_age,
-2*act_cc,
-1*act_dueinc,
-2*act_loaninc,
```

```
 2*app_income,
-2*agr3_Mean_Due,
-3*ags3_Mean_Days,
-3*agr6_Mean_Due,
-3*ags6_Mean_Days,
-2*agr9_Mean_Due,
-3*ags9_Mean_Days,
-2*agr12_Mean_Due,
-2*ags12_Mean_Days,
-2*ags3_Max_CMax_Due,
-2*ags12_Max_CMax_Due,
-1*ags9_Max_CMax_Days,
-1*act12_n_cus_arrears,
-2*ags12_Max_CMax_Due,
 5*(ags12_Max_CMax_Due=0)).
```

Zauważmy, że spłacalność kredytów gotówkowych zależy od kredytów ratalnych tylko ze względu na pierwszą zmienną w wypisanych członach. Innymi słowy opóźnienia klienta w kredytach ratalnych w ciągu ostatnich 12 miesięcy ciążą na kredytach gotówkowych.

Wyjątkowo w sytuacji kredytów gotówkowych pojawia się dodatkowy skoring, który określa pozytywną reakcję na kampanie gotówkowe (używane są tu oznaczenia ‘ins’ – kredyt ratalny, ‘css’ – kredyt gotówkowy):

```
score_response_css=sum(
-1*act_age,
-2*app_income,
 2*app_nom_gender,
 3*(app_nom_job_code in (1,4)),
 1*app_number_of_children,
 1*(app_nom_marital_status in (1,2)),
 1*(app_nom_city in (1,4)),
 1*(app_nom_home_status in (1,3)),
 1*app_nom_cars,
 1*app_spendings,
 1*act_cins_seniority,
```

– im dłuższa jest historia kredytów ratalnych klienta, tym bardziej chce kredyt gotówkowy

```
 1*act_cins_min_seniority,
```
– im mniej miesięcy minęło od ostatniego kredytu ratalnego, tym bardziej chce kredyt gotówkowy
```
 1*act_cins_n_loans_hist,
 2*act_cins_n_statC,
-2*act_cins_n_statB,
 1*act_cins_maxdue,
 4*(2<=act_cins_min_pninst<=7),
```
– jeśli spłacił już 2 do 7 rat kredytu ratalnego, to chce kredyt gotówkowy
```
 4*(1<=act_cins_min_lninst<=4),
```
– jeśli zostały mu 1 do 4 rat do spłacenia kredytu ratalnego, to chce kredyt gotówkowy
```
 2*act_ccss_seniority,
 2*act_ccss_min_seniority,
 1*act_ccss_n_loans_hist,
 3*act_ccss_n_statC,
-3*act_ccss_n_statB,
 1*act_ccss_n_loans_act,
 6*(2<=act_ccss_min_pninst<=5),
 6*(1<=act_ccss_min_lninst<=4),
 3*act_ccss_utl,
 2*act_ccss_cc).
```

Zwróćmy uwagę, że model marketingowy nie jest tożsamy z ryzykiem. Inne skoringi determinują opóźnienia, a inne – skłonności – do zaciągania kolejnych kredytów. Jest pewna część klientów, którzy biorą kolejny kredyt, gdyż mają poważne problemy finansowe i wpadają w pętlę kredytową, przekredytowując się i doprowadzając do bankructwa. Jest jednak i taka część klientów, która posiada mniejsze ryzyko i także jest zainteresowana kolejnymi kredytami. Jedyną rzeczą, jaką trzeba zarządzać, to umiejętność odróżniania jednych klientów od drugich, w czym bardzo pomocne są modele Credit Scoring.

Trzeba też pamiętać, że kredyt gotówkowy może się pojawić przy danym kliencie, tylko gdy miesiąc wcześniej miał jakieś aktywne rachunki w symulacyjnym portfelu bankowym.

Spis rysunków

Spis tabel

Bibliografia

Anderson B., Haller S., Siddiqi N. (2009), *Reject inference techniques implemented in Credit Scoring for SAS Enterprise Miner*, SAS Working paper, 305, http://support.sas.com/resources/papers/proceedings09/305-2009.pdf, dostęp: 2014.03.23.

Anderson R. (2007), *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press.

Banasik J., Crook J. (2003), *Lean models and reject inference*, Credit Scoring & Credit Control VIII, Edinburgh.

Banasik J., Crook J. (2005), *Credit Scoring, augmentation and lean models*, Journal of the Operational Research Society, 56(9), s. 1072–81.

Banasik J., Crook J. (2007), *Reject inference, augmentation, and sample selection*, European Journal of Operational Research, 183(3), s. 1582–94.

Belsley D.A. (1991), *Conditioning Diagnostics, Collinearity and weak data in regression*, John Wiley & Sons, New York.

Belsley D.A., Kuh E., Welsch R.E. (1980), *Regression Diagnostics*, John Wiley & Sons, New York.

Benmelech E., Dlugosz J. (2010), *The credit rating crisis*, NBER Macroeconomics Annual, 24, http://www.nber.org/chapters/c11794, dostęp: 2013.09.22.

Berman J.J. (2013), *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*, Morgan Kaufmann, Elsevier.

BIS (2009), *Principles for sound stress testing practices and supervision*, Raport techniczny, Basel Committee on Banking Supervision, Bank For International Settlements, http://www.bis.org, dostęp: 2014.06.20.

BIS–BASEL (2005), *International convergence of capital measurement and capital standards*, Raport techniczny, Basel Committee on Banking Supervision, Bank For International Settlements, http://www.bis.org, dostęp: 2012.08.30.

BIS–WP14 (2005), *Studies on validation of internal rating systems, working paper no. 14*, Raport techniczny, Basel Committee on Banking Supervision, Bank For International Settlements, http://www.bis.org, dostęp: 2012.08.30.

Blikle A.J. (1994), *Doktryna jakości – rzecz o skutecznym zarządzaniu*, In statu nascendi, http://www.moznainaczej.com.pl/Download/Doktryna Jakosci/DoktrynaJakosci.pdf, dostęp: 2014.02.19.

DeBonis J.N., Balinski E., Allen P. (2002), *Value-Based Marketing for Bottom-Line Success 5 Steps to Creating Customer Value*, American Marketing Association. The McGraw–HIll Companies, Inc.

Delen D., Walker G., Kadam A. (2005), *Predicting breast cancer survivability: a comparison of three data mining methods*, Artificial Intelligence in Medicine, 34, s. 113–127, http://seer.cancer.gov, dostęp: 2012.08.30.

Finlay S. (2010), *Credit Scoring, Response Modelling and Insurance Rating*, PALGRAVE MACMILLAN.

Frątczak E. (2012), *Zaawansowane Metody Aanaliz Statystycznych*, Oficyna Wydawnicza SGH, Warszawa.

Furnival G.M., Wilson R.W. (1974), *Regression by leaps and bounds*, Technometrics, 16, s. 499–511.

Gartner-Report (2001), *Application delivery strategies*, META Group Inc. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf, dostęp: 2014.06.17.

Goldenberg B.J. (2008), *CRM in Real Time: Empowering Customer Relationships*, Medford, New Jersey.

Hand D.J., Henley W.E. (1994), *Can reject inference ever work?* IMA Journal of Mathematics Applied in Business & Industry, 5, s. 45–55.

Huang E. (2007), *Scorecard specification, validation and user acceptance: A lesson for modellers and risk managers*, Credit Scoring Conference CRC, Edinburgh, http://www.business-school.ed.ac.uk/crc/conferences/conference-archive?a=45487, dostęp: 2012.08.30.

Huang E., Scott C. (2007), *Credit risk scorecard design, validation and user acceptance: A lesson for modellers and risk managers*, Credit Scoring Conference CRC, Edinburgh, http://www.business-school.ed.ac.uk/crc/conferences/conference-archive?a=45569, dostęp: 2012.08.30.

Kennedy K., Delany S.J., Namee B.M. (2011), *A framework for generating data to simulate application scoring*, Credit Scoring Conference CRC, Edinburgh, http://arrow.dit.ie/scschcomcon/96/, dostęp: 2013.09.23.

Kincaid C. (2013), *How to be a data scientist using SAS*, NESUG Proceedings, http://support.sas.com/resources/papers/proceedings14/1486-2014.pdf, dostęp: 2013.09.22.

Konopczak M., Sieradzki R., Wiernicki M. (2010), *Kryzys na światowych rynkach finansowych – wpływ na rynek finansowy w polsce oraz implikacje dla sektora realnego*, Bank i Kredyt, 41(6), s. 45–70, http://www.bankikredyt.nbp.pl, dostęp: 2013.09.22.

Koronacki J., Mielniczuk J. (2001), *Statystyka dla studentów kierunków technicznych i przyrodniczych*, WNT, Warszawa.

Krzyśko M., Wołyński W., Górecki T., Skorzybut M. (2008), *Systemy uczące się - rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*, WNT, Warszawa.

Kwiatkowska A.M. (2007), *Systemy wspomagania decyzji. Jak korzystać z WIEDZY i informacji*, WNT, Warszawa.

Lessmanna S., Seowb H.V., Baesenscd B., Thomasd L.C. (2013), *Benchmarking state–of–the–art classification algorithms for Credit Scoring: A ten–year update*, Credit Scoring Conference CRC, Edinburgh, http://www.business-school.ed.ac.uk/crc/conferences/, dostęp: 2013.09.23.

Malik M., Thomas L.C. (2009), *Modelling credit risk in portfolios of consumer loans: Transition matrix model for consumer credit ratings*, Credit Scoring Conference CRC, Edinburgh, http://www.business-school.ed.ac.uk/crc/conferences/conference-archive?a=45281, dostęp: 2012.08.30.

Matuszyk A. (2008), *Credit Scoring*, CeDeWu, Warszawa.

Mayer-Schonberger V., Cukier K. (2013), *Big Data: A revolution that will transform how we live, work and think*, An Eamon Dolan Book, Houghton Mifflin Harcourt, Boston, New York.

Mays E. (2009), *Systematic risk effects on consumer lending products*, Credit Scoring Conference CRC, Edinburgh, http://www.business-school.ed.ac.uk/crc/conferences/conference-archive?a=45269, dostęp: 2012.08.30.

Mester L.J. (1997), *What's the point of Credit Scoring?* Business Review, SEPTEMBER/OCTOBER 1997, Federal Reserve Bank of Philadelphia, s. 8–9.

Ogden D.C. (2009), *Customer lifetime value (CLV). A methodology for quantifying and managing future cash flows*, SAS Working paper. http://www.sas.com/offices/NA/canada/downloads/CValue11/Customer-Lifetime-Value-David-Ogden-Nov2009.pdf, dostęp: 2014.02.19.

Ohlhorst F.J. (2013), *Big Data Analytics: Turning Big Data into Big Money*, Wiley and SAS Business Series.

Payne A. (2005), *HANDBOOK OF CRM: Achieving Excellence in Customer Management*, Elsevier.

Poon M. (2007), *Scorecards and devices for consumer credits: The case of fair, isaac and company incorporated*, The Sociological Review, Issue Supplement S2, 55, s. 284–306.

Przanowski K. (2013), *Banking retail consumer finance data generator - Credit Scoring data repository*, e–FINANSE, 9(1), s. 44–59, http://arxiv.org/abs/1105.2968, dostęp: 2012.08.30.

Przanowski K. (2014), *Rola danych symulacyjnych w badaniach Credit Scoring*, Monografia, Statystyka w służbie biznesu i nauk społecznych, Wydawnictwo Wyższej Szkoły Menedżerskiej w Warszwie.

Przanowski K., Mamczarz J. (2012), *Generator danych consumer finance – metody porównywania technik modelowych w Credit Scoring*, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Poznaniu, 239, s. 97–111.

Scallan G. (2011), *Selecting characteristics and attributes in logistic regression*, Credit Scoring Conference CRC, Edinburgh, http://www.scoreplus.com/resources/reference/index.php, dostęp: 2012.08.30.

Scallan G. (2013), *Marginal Kolmogorov–Smirnov analysis: Measuring lack of fit in logistic regression*, Credit Scoring Conference CRC, Edinburgh, http://www.scoreplus.com/resources/reference/index.php, dostęp: 2014.02.14.

Siddiqi N. (2005), *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, Wiley and SAS Business Series.

Soubra D. (2012), *The 3vs that define big data*, Data Science Central, Web page for posting, http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data, dostęp: 2014.01.09.

Thomas L.C., Edelman D.B., Crook J.N. (2002), *Credit Scoring and Its Applications*, Society for Industrial and Applied Mathematics, Philadelfia.

Thonabauer G., Nosslinger B. (2004), *Guidelines on Credit Risk Management. Credit Approval Process and Credit Risk Management*, Oesterreichische Nationalbank and Austrian Financial Market Authority.

Verstraeten G., den Poel D.V. (2005), *The impact of sample bias on consumer Credit Scoring performance and profitability*, Journal of the Operational Research Society, 56, s. 981–992.

Welfe A. (2003), *Ekonometria*, PWE, Warszawa.

KAROL PRZANOWSKI – adiunkt w Instytucie Statystyki i Demografii Szkoły Głównej Handlowej w Warszawie. Absolwent matematyki teoretycznej Uniwersytetu Łódzkiego i doktor fizyki teoretycznej.

Naukowo zajmuje się teoretyczną stroną Credit Scoring. Posiada duże doświadczenie w analizowaniu portfela Consumer Finance i tworzeniu symulatorów danych odzwierciedlających procesy tego portfela. Jest ekspertem z Sytemu SAS, zaawansowanego programowania i analiz statystycznych. Jest autorem wielu własnych programów SAS 4GL do budowy modeli kart skoringowych. Opiekun Studenckiego Koła Naukowego Business Analytics. Prowadzi jedyne w swoim rodzaju zajęcia z „Credit Scoring i Makroprogramowania w SAS".

Odpowiedzialny w dużych bankach grup kapitałowych za budowanie, wdrażanie i monitoring modeli predykcyjnych, tworzenie zautomatyzowanych procesów CRM, zarządzanie kampaniami i ofertami, tworzenie automatycznych procesów budżetowania i planowania.

Z wielką pasją podchodzi do SAS 4GL i makroprogramowania, uważa dzień za stracony, jeśli nie napisze choć kilku linii kodu. Współautor podręcznika „Przetwarzanie danych w SAS". Autor rozdziałów z zaawansowanego makroprogramowania i kolejności uruchamiania kodu.

Przedstawione opracowanie dotyczy jednego z najważniejszych problemów praktycznych rozpatrywanych w finansach, mianowicie metod skoringowych. Uważam wybór obszaru badań za słuszny. Autor w książce koncentruje się na dwóch zadaniach. Pierwsze to zbudowanie generatora losowych danych podobnych do danych dotyczących kredytów konsumenckich, a drugie to konstrukcja modelu skoringowego. Autor opracowania wykonał sporą pracę w stworzeniu systemów, pokazując to szczegółowo w książce. Rozważania są dobrze umocowane w literaturze.

prof. dr hab. Krzysztof Jajuga
Uniwersytet Ekonomiczny we Wrocławiu

OFICYNA WYDAWNICZA
SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE
02-554 Warszawa, al. Niepodległości 162
tel. 22 564 94 77, fax 22 564 86 86
www.wydawnictwo.sgh.waw.pl
e-mail: wydawnictwo@sgh.waw.pl

ISBN 978-83-7378-922-7