KAROL PRZANOWSKI

CREDIT SCORING

Studia przypadków procesów biznesowych

110 lat 1906 2016 SGH

OFICYNA WYDAWNICZA
SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE

KAROL PRZANOWSKI

CREDIT SCORING

Studia przypadków procesów biznesowych

110 lat 1906 2016 SGH

OFICYNA WYDAWNICZA
SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE
WARSZAWA 2015

Recenzent
Jerzy Różański

Redaktor
Zofia Wydra

Wydanie I

ISBN 978-8330-051-4

Szkoła Główna Handlowa w Warszawie – Oficyna Wydawnicza
02-554 Warszawa, al. Niepodległości 162
tel. +48 22 564 94 77, 22 564 94 86
www.wydawnictwo.sgh.waw.pl
e-mail: wydawnictwo@sgh.waw.pl

Projekt 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 35/III/16

Jest to kolejna książka dr. Karola Przanowskiego o tematyce scoringowej, w której autor omawia zastosowania w biznesie analizy danych z obszaru Advanced Business Analytics. W pracy w prosty i zarazem przystępny sposób przedstawiono metody pomiaru przydatności modeli predykcyjnych w ujęciu finansowym. Przykłady te przemawiają do wyobraźni Czytelnika. Tego typu publikacji jest wciąż bardzo mało. Wielką zaletą książki jest dołączony zestaw gotowych arkuszy kalkulacyjnych, które samodzielnie można dostosowywać do konkretnych przypadków i modyfikować. Daje to nieocenione możliwości formułowania własnych scenariuszy obliczeniowych i tym samym „wciąga" w to, co powszechnie określa się inżynierią danych na użytek analityki.

Gorąco polecam i życzę owocnej lektury
dr hab. Ewa Frątczak,
prof. Szkoły Głównej Handlowej w Warszawie

***

Książka Karola Przanowskiego dotyczy jednego z najczęściej spotykanych w praktyce problemów finansowych, jakim jest Credit Scoring. Rozważania ujęte w pracy mają kilka ważnych cech. Jest to przystępny z punktu widzenia praktyki opis narzędzi, jak również pokazanie zastosowań Credit Scoringu także w innych obszarach, takich jak projektowanie kampanii reklamowych oraz relacje z potencjalnie odchodzącymi klientami. Unikalna zaleta książki to wskazanie znaczenia danych symulowanych, przybliżających trudniejsze do uzyskania dane rzeczywiste. Znalazły się tu również szczegółowe opisy praktycznych sytuacji. Książka może być polecona przede wszystkim praktykom, ale znajdzie też zastosowanie w procesie dydaktycznym.

prof. dr hab. Krzysztof Jajuga,
Uniwersytet Ekonomiczny we Wrocławiu

***

Karol Przanowski podejmuje temat obecnie bardzo ważny dla wielu instytucji finansowych, w tym przede wszystkim banków. Jest nim budowa modeli prognozujących zachowania klientów indywidualnych. Ze względu na swoje doświadczenie naukowe i zawodowe Autor skupia się na omówieniu zagadnień związanych ze scoringiem kredytowym – robi to w sposób kompleksowy i zarazem zwięzły, gdyż tematyka jest obszerna. Stara się naświetlić kwestie związane z budową modeli selekcjonujących kredytobiorców na różnych etapach budowania relacji z bankiem: w momencie składania pierwszego wniosku kredytowego, przy odnawianiu kredytu, w procesie przygotowywania dla klientów rozszerzonej oferty kredytowej czy też w przypadku pojawienia się opóźnień w spłatach i finalnie w procesie restrukturyzacji kredytowej i windykacji. Autor patrzy na zagadnienie w sposób szeroki – w kontekście nie tylko strat kredytowych, ale także osiąganego przez bank dochodu odsetkowego i prowizyjnego oraz ponoszonych kosztów kampanii marketingowych. Słusznie stara się zaprezentować podejście polegające na optymalizacji realizowanego dochodu, a nie wyłącznie na minimalizowaniu strat. Modele predykcyjne nie są remedium na wszystkie problemy dotyczące zarządzania biznesem i ryzykiem w banku, wręcz przeciwnie – są tylko jednym z narzędzi wspierających procesy decyzyjne. Modele potrzebują ciągłego rozwoju i – oczywiście – należy zgodzić się z Autorem, że nawet najlepiej skalibrowane na danych testowych formuły nie zadziałają w realnym świecie, gdy bank nie będzie dążył do budowania jak najlepszych procesów zarządczych oraz gromadził wysokiej jakości danych w ramach tych procesów.

Michał Sobiech,
członek zarządu CFO&CRO, Bank Pocztowy S.A.

***

Byłem promotorem dr. Karola Przanowskiego na studiach MBA w Instytucie Nauk Ekonomicznych Polskiej Akademii Nauk. Uważam, że jego sposób prezentowania zastosowań Advanced Analytics w biznesie jest godny uwagi. Jako osoba na co dzień związana z finansami gorąco polecam właśnie taki sposób argumentowania przydatności modeli predykcyjnych – podkreślanie nie tyle zawiłych statystycznych własności konstrukcji samego modelu, ile związanych z nim korzyści, czyli tego, jak i w jakim stopniu optymalizuje typowe wielkości, takie jak wynik finansowy, przychody czy koszty. Jestem przekonany, że materiał ten pomoże wielu środowiskom biznesowym lepiej poznać możliwości zastosowań zaawansowanej analizy danych, a tym samym przyczyni się do poprawy efektywności przedsiębiorstw, wspomagając je w zarządzaniu finansami.

dr Leszek Borowiec,
członek zarządu Poczty Polskiej Usługi Cyfrowe Sp. z o.o.,
dyrektor zarządzający Pionem Finansów Poczty Polskiej S.A.

***

Spis treści

Od autora

Pod koniec 2014 roku w Oficynie Wydawniczej SGH opublikowałem książkę *Credit Scoring w erze Big-Data – techniki modelowania z wykorzystaniem generatora losowych danych portfela Consumer Finance*. Prowadząc zajęcia ze studentami studiów dziennych i podyplomowych, uświadomiłem sobie jednak potrzebę uzupełnienia jej o rozdziały zaznamiające czytelnika z podstawowymi pojęciami, procesami i strukturami danych potrzebnymi do zrozumienia Credit Scoringu. Pomogły mi w tym także dyskusje prowadzone podczas spotkań Studenckiego Koła Naukowego Business Analytics, ważne było także wsparcie studentów, którzy pomogli mi przygotować pierwszą wersję arkusza kalkulacyjnego do symulacji, szczególną pomoc okazała studentka Agata Misiak.

Zostałem także zmotywowany do tego, by ukończyć podyplomowe studnia MBA, a następnie postanowiłem wykorzystać okazję do rozpowszechnienia tematu, który fascynuje mnie już od kilkunastu lat.

Jednym z ważnych powodów tych działań stała się także chęć przedstawienia metod Credit Scoring w sposób łatwy i wyraźnie odwołujący się do języka biznesowego, co pozwoli zarówno być ścisłym statystycznie, jak i jednocześnie posługiwać się wskaźnikami i liczbami związanymi bezpośrednio z przymnażaniem kapitału w przedsiębiorstwie. Drugim powodem była potrzeba rozszerzenia zakresu Credit Scoringu z dziedziny bankowej na inne, takie jak zarządzanie kampaniami reklamowymi czy utrzymanie odchodzących klientów telekomunikacji. Postanowiłem stworzyć proste symulacje w arkuszu kalkulacyjnym, by można było samodzielnie upewnić się, czy rzeczywiście w danym biznesowym procesie modele skoringowe przynoszą milionowe korzyści.

Dwa rozdziały musiałem skopiować z mojej poprzedniej książki, lekko je modyfikując i dodając nowe komentarze, aby nowe opracowanie stanowiło spójną całość. Można zatem czytać je jako oddzielną pracę albo traktować jako pierwszy krok do wejścia w imponujący świat Credit Scoringu, który nadal zachwyca i budzi respekt. Nawet w XXI w. musimy pogodzić się z faktem, że wiele jest jeszcze do

zrobienia i problemy modelowania predykcyjnego nadal potrafią zadziwić nawet doświadczonych analityków.

Przedstawiłem w książce wiele studiów przypadków, z których większość to proste symulacje w arkuszu kalkulacyjnym, a kilka to rozbudowane symulacje i procesy w zaawansowanym systemie SAS do przetwarzania i analizowania dużych zbiorów danych. Wszelkie parametry są zbliżone do rzeczywistości, ale nie są prawdziwe, nie reprezentują przypadku konkretnego przedsiębiorstwa. Pomimo że funkcjonujemy w epoce *Big Data*, trudno jest pozyskać niektóre dane. Sprzyja to niestety rozpowrzechnianiu się poglądu, że naukowiec zajmuje się tylko teorią, a konsultant praktyką. Można jednak na danych symulacyjnych, quasi-rzeczywistych, wykazać znajomość tematu i pokazać w bardzo praktyczny sposób całą inżynierię procesów biznesowych. Tak zebrane studia przypadków stają się wartościowym materiałem działań przedsprzedażowych (ang. *presale*) i pozwalają spojrzeć na zagadnienie okiem zarówno praktyka, jak i naukowca. W tym drugim przypadku ważna jest pewna doza zdrowej krytyki, by ostrożnie interpretować wyniki analiz, uwzględnić szersze spojrzenie, zdawać sobie sprawę ze słabości modeli matematycznych i nie popadać w swego rodzaju „naiwność" analityczną. Bardzo ważne staje się zatem uświadomienie sobie, że najlepszą praktyką jest dobra teoria.

Można odnieść wrażenie, że w dzisiejszym biznesie nadal wybiera się proste reguły biznesowe, najczęściej rekomendowane przez firmy konsultingowe jako najlepsze standardy, pomijając wykorzystywanie zaawansowanych modeli statystycznych. Ulega się mechanizmowi prostego kopiowania rozwiązań, dążąc do utrzymania *status quo*, zamiast twórczo pomnażać kapitał przedsiębiorstwa.

Wstęp

Obecne czasy najmocniej są związane z odkryciem roli danych w zarządzaniu przedsiębiorstwami. Dziś największymi aktywami stają się dane i modele statystyczne, które na ich bazie potrafią wspierać i podejmować automatyczne decyzje. Największa innowacyjność technologiczna to nie komputer czy sieć internetowa, ale przede wszystkim dane gromadzone w sieci i w przedsiębiorstwach podczas realizacji różnego rodzaju procesów biznesowych. Danych przybywa w tempie niewiarygodnym i powoduje to, że wiele instytucji zaczyna zupełnie inaczej traktować swoje produkty. Telefon przestaje służyć głównie do wykonywania połączeń telefonicznych, staje się źródłem cennych danych o kliencie, nie tylko dane bilingowe, ale przede wszystkim geolokacyjne, czyli związane z aktualnym położeniem i przemieszczaniem się użytkownika telefonu, także informacje o aplikacjach i ich użytkowaniu stają się okazją do uzyskania przewagi konkurencyjnej firm, które potrafią te dane przetworzyć w lepsze i bardziej dopasowane do potrzeb klientów produkty i usługi. Na tym właśnie polega nowa rewolucja, którą nazywa się *Big Data* (Przanowski, 2014a).

Jesteśmy świadkami istotnej zmiany postrzegania świata i procesów biznesowych. Na naszych oczach następuje całkowita digitalizacja wszystkiego i wszystkich. Zarówno produkty, jak i sami ludzie stają się obiektami produkującymi setki, miliony informacji. Tego tempa zmian nikt już nie zatrzyma. Daje ono okazję do podjęcia nowych wyzwań, w szczególności do coraz efektywniejszego usprawniania modeli i procesów biznesowych, które dziś stają się głównie masowe i automatyczne. Potrzebują one zatem wsparcia zaawansowanych narzędzi analitycznych, by nimi sterować i kontrolować ich jakość.

Podstawową metodą sterowania procesami jest ich optymalizacja, czyli szukanie takiego rozwiązania, które stanowi optimum. Maksymalizuje ona funkcję celu, którą najczęściej jest wielkość oparta na wskaźnikach finansowych. Dążymy do minimalizacji kosztów, maksymalizacji przychodów lub też minimalizujemy czas produkcji czy wykonania usługi itp.

Jednym z ważniejszych sposobów optymalizacji jest stosowanie modeli predykcyjnych, które potrafią prognozować badane zjawisko. Najczęściej chcemy przewidywać zachowanie klientów. Jeśli wiemy, ilu klientów kupi nasz produkt, to możemy dobrze zaplanować jego produkcję, a tym samym cały budżet. Prognozowanie zachowania klientów staje się zatem kluczowe w wielu obecnie rozwijanych biznesach. Im więcej klientów, tym bardziej trzeba ufać narzędziom statystycznym, które potrafią badać zjawiska masowe. Tylko analiza dużej liczby danych pozwala wychwycić subtelne różnice w zachowaniu każdego klienta. To bardzo ciekawa myśl, która staje się istotną dewizą towarzyszącą optymalizacji. Rozważmy szczególny przykład procesu biznesowego, jakim jest akceptacja kredytów w banku. Jeśli postanowimy wnioski kredytowe rozpatrywać indywidualnie i zatrudnimy wielu analityków kredytowych, to każdy z nich, na podstawie doświadczenia zdobytego w kontaktach z przychodzącymi do niego klientami, po jakimś czasie odnajdzie reguły rozpoznawania tych, którzy kredyty będą spłacać terminowo. Każdy z analityków będzie jednak posiadał inny zestaw reguł. Jeden odkryje, że młodsi klienci spłacają gorzej, drugi może nawet temu zaprzeczyć, gdyż akurat do niego ustawiali się w kolejce głównie młodzi, musiał zatem znaleźć zupełnie inne kryterium rozróżniające. Być może wieloletnie doświadczenie takiego analityka będzie już dawało bardzo dobre efekty, ale niestety znacznie lepsze można uzyskać przez centralizację procesu i analizę danych historycznych w całości. Tylko dzięki zgromadzeniu ich wszystkich w jednym miejscu i przeanalizowaniu bardzo szczegółowo możliwe jest właśnie prawdziwe wychwycenie subtelności. Wystarczy przy każdym wniosku mylić się średnio o kilka złotych mniej. Teraz pojawia się kolejny ważny czynnik optymalizacyjny. Owa maleńka różnica w poprawie decyzji w przypadku jednego klienta jest mnożona przez liczbę wszystkich klientów czy wniosków. Im więcej jest wniosków, tym większe efekty finansowe, oszczędności lub przychody daje zsumowana różnica.

W pracy podjęto głównie temat roli danych symulacyjnych w badaniu procesów biznesowych – w ich optymalizacji. W wielu dziedzinach, a w szczególności w bankowości, nie jest możliwe otrzymanie danych rzeczywistych. Wynika to głównie z ochrony tajemnicy przedsiębiorstwa, gdyż z danych takich można wyciągnąć wiele

istotnych wniosków dotyczących własności danego banku. Utrudnia to badania naukowe. Trzeba zatem sięgać po dane symulacyjne, by przynajmniej w przybliżeniu pokazywać i analizować problemy, które występują w rzeczywistości.

Niestety problem dostępu do danych istnieje także w innych sektorach biznesu, być może otrzymamy pozwolenie, by używać większych zakresów danych, ale również tu niektóre informacje są chronione, np. wybrane parametry produktów lub składowe kosztów.

Drugim ważnym celem pracy jest zaprezentowanie korzyści z wykorzystania modeli predykcyjnych w procesach biznesowych na podstawie przykładowych analiz finansowych. Nawet jeśli nie jest możliwe przeprowadzenie prawdziwego studium przypadku konkretnego procesu biznesowego, to i tak wykorzystanie danych symulacyjnych i zaprezentowanie podstawowych wskaźników finansowych danego procesu są już wystarczające do tego, by wyrobić sobie zdanie i by potem w zetknięciu się z prawdziwymi danymi wiedzieć, jak zarządzać procesem. Można by tu mówić o typowych i znanych modelach biznesowych, ale z tym pojęciem kojarzy się nam już pełny opis złożonego procesu i składowych finansowych. W naszym przypadku odwołujemy się wyłącznie do wskaźników takich, jak przychody i koszty. Bynajmniej nie spłyca to istoty problemu, wręcz przeciwnie – pokazuje, że warto sobie zdać sprawę z wagi niektórych procesów zachodzących w instytucji finansowej, dzięki którym wynik finansowy staje się dodatni.

Jeden z ważnych i aktualnych problemów w kontekście *Big Data* to poprawne określenie tego, kim jest naukowiec zajmujący się danymi czy inżynier danych (ang. *data scientist*) (Kincaid, 2013). Jedną z odpowiedzi może być: to ten, który dobrze poznał podstawy analizy danych i będzie w stanie szybko uzupełnić brakującą wiedzę, kiedy spotka się z prawdziwymi problemami w życiu biznesowym. Może być to też ten, kto umiejętnie opanował kilka dziedzin z odpowiednimi wagami: statystykę, by operować właściwym zestawem narzędzi zaawansowanej analizy; programowanie, by samodzielnie pisać algorytmy i tworzyć zaawansowane analizy i raporty. Trzeba także znać się na biznesie, by statystykę i programowanie umieć stosować przynajmniej w jednej dziedzinie. Owa umiejętność jest związana z rozumieniem procesów biznesowych, czyli tego, gdzie

się traci, inwestuje i zarabia pieniądze oraz jak „zgrać" wszystkie wymienione procesy, by razem przynosiły zyski. Istotą jest specyficzne wzmocnienie, interakcja tych cech czy kompetencji w jednej osobie. Powoduje to niewiarygodne przyspieszenie prac nad ulepszaniem procesów i sprawia, że tego typu fachowców jest niewielu na rynku pracy. Lecz jeśli się już pojawiają w naszym środowisku zawodowym, to patrzymy na nich z lekkim niedowierzaniem, bo wyłamują się z typowych wzorców. Nie organizują wielkich i kosztownych projektów, wykorzystują istniejące zasoby „bez szemrania" i wreszcie mają na wszystko czas, jednocześnie dotrzymując ustalonych terminów wdrożeń. Patrząc na nich z daleka, ma się wrażenie, że ich praca jest prosta, a nawet beztroska, że mają dużo wolnego czasu, gdy tymczasem inni tak ciężko pracują. Dzieje się tak dlatego, że inżynier danych pozwala sobie na myślenie wychodzące poza schematy (ang. *out of box*) i uważa to za najważniejszy element swojej pracy. Zobrazować to można w bardzo prosty sposób: dotychczasowe środowisko pracy było przyzwyczajone do wiosłowania, zatem od nowego pracownika oczekuje się, że efektem jego pracy będzie nowe i lepsze wiosło, tymczasem on proponuje silnik motorowy. Inny ważny element to dobrze dobrane i „szyte na miarę" procesy, które pozostawia za sobą. Wygląda to jak schody ruchome poruszające się do góry – czy się po nich idzie, czy stoi i tak jedzie się do góry. To właśnie czyni go spokojniejszym i sprawia, że dzięki dyskusjom i rozmowom z ludźmi lepiej poznaje problemy dotyczące procesów. Źle zaprojektowany proces to sytuacja odwrotna – schody poruszające się do dołu; aby przemieszczać się do góry, trzeba cały czas szybko wchodzić, a każdy przystanek czy odpoczynek sprowadza nas z powrotem na niższe poziomy. Jest to bardzo ciekawe, że z jednej strony żyjemy dziś w epoce *Big Data*, w której dane są dla nas tak istotne ze względu na swoją użyteczność, a z drugiej strony wiele procesów jest słabo zaprojektowanych. Bywa też paradoksalnie tak, że uświadamiając sobie liczne uchybienia procesów, decydujemy się zatrudnić inżyniera danych i od pierwszych dni jego pracy oczekujemy istotnych zmian. Tymczasem on pozornie nic nie zmienia, tylko pyta o miliony dziwnych szczegółów. Niestety procesu latami źle budowanego i zarządzanego nie można szybko ulepszyć

i nie jest to kwestia technologii, najczęściej wiąże się to ze zmianą mentalności wielu pracowników.

Ostatnią umiejętnością inżyniera danych jest komunikowanie się. Ta cecha jest nadal stanowczo zbyt rzadka w dzisiejszym biznesie i dlatego na naszych oczach biznes oddziela się od informatyki (działów IT). Dzieje się tak, ponieważ pracownicy obu tych obszarów nie mogą 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 stanowią jedno z najważniejszych źródeł podejmowania decyzji i pracownicy wszystkich departamentów zaczynają rozumieć swoją misję.

Rola inżyniera danych polega na prezentowaniu rzeczy trudnych, takich jak zaawansowane modelowanie statystyczne, w sposób prosty czy też zrozumiały przez osoby ze środowiska biznesu niekoniecznie znające się na statystyce. Trzeba zatem umieć formułować i opisywać większość problemów wielkościami finansowymi, miernikami, którymi posługuje się biznes. Nowa era *Big Data* stwarza poważne wyzwanie dla osób zajmujących się analizą danych. Z jednej strony liczba danych staje się tak duża, że zmusza firmy do tworzenia zespołów analitycznych i budowania nowych, wydajniejszych rozwiązań informatycznych. Z drugiej strony wymaga dowodów, że umiejętne wykorzystanie danych przynosi istotne korzyści finansowe. To zadanie wymaga wielu prób, wyrzeczeń i testów. Nie każda analiza danych przekłada się na szybki zarobek. Trzeba lat i możliwości popełniania wielu błędów, by wykształcił się dobry inżynier danych. Bardzo złudna jest nadzieja wielu firm konsultingowych, że gdy podpiszą kontrakt, gdy potencjalny klient zgłosi zapytanie ofertowe związane z projektem zaawansowanego modelowania, to znajdzie się wykonawców na rynku pracy. Jednak wykonawcy tacy muszą się gdzieś nauczyć, muszą mieć doświadczenie, a je zdobywa się tylko w pracy z danymi. Jeśli zatem nie można mieć danych rzeczywistych, trzeba sięgać po losowe, symulacyjne i dzięki nim kształcić przyszłych inżynierów danych.

Cel i struktura pracy

Sformułowanie celu

Podstawowym celem pracy jest wykazanie przydatności danych losowych w tworzeniu symulacji procesów biznesowych. Pomimo iż żyjemy w czasach, gdy dane odgrywają coraz to większą rolę, w zbyt małym stpniu przekonuje się środowiska biznesowe do wykorzystywania zaawansowanych modeli analitycznych optymalizujących procesy biznesowe.

Można zadać proste pytanie: jak moc predykcyjna modeli statystycznych wpływa na osiągane zyski w przedsiębiorstwach? Niestety zbyt mało uwagi poświęcamy temu problemowi i obecnie w literaturze praktycznie nie znajdziemy jednoznacznych kalkulacji. Próba odpowiedzi na to pytanie z podaniem sensownych wielkości finansowych jest podstawowym celem niniejszego opracowania.

Pytanie to można sformuować jeszcze inaczej: czy możliwe jest przygotowanie listy studiów przypadków, modeli finansowych powszechnie dziś znanych procesów biznesowych, by móc przeprowadzać prezentacje w różnych przedsiębiorstwach i przekonywać środowisko do szerszego otwarcia się na zaawansowane metody analizy danych w celu osiągania coraz to większych zysków?

Wreszcie, czy możemy na podstawie danych symulacyjnych przedstawiać i przybliżać główne problemy typowych procesów biznesowych, czy można dzięki temu, pomimo ochrony danych rzeczywistych, rozwijać badania naukowe i przygotowywać materiały edukacyjne? Czy możliwe jest szczegółowe dyskutowanie, krytykowanie i szukanie najlepszego rozwiązania w przypadku, gdy nie posiadamy rzeczywistych danych?

W książce przeanalizowano następujące procesy:

- akceptacji kredytowej kredytu ratalnego;
- akceptacji kredytowej połączonego biznesu: akwizycji kredytu ratalnego i sprzedaży krzyżowej kredytu gotówkowego;
- akceptacji kredytowej kredytu hipotecznego;

- zarządzania windykacją polubowną;
- zarządzania kampaniami reklamowymi;
- utrzymania odchodzących klientów.

Każdy z procesów, z wyjątkiem akceptacji kredytowej połączonego biznesu akwizycji i sprzedaży krzyżowej, jest prezentowany także w dołączonych do książki arkuszach kalkulacyjnych, dzięki którym studiowanie ich staje się znacznie ciekawsze i pozwala dostosować je do swoich potrzeb, zarówno zmieniając formuły, jak i w szczególności wprowadzając własne, rzeczywiste wartości parametrów.

Na sześć procesów tylko trzy odnoszą się do kredytowania klientów, czyli bezpośrednio do typowych procesów bankowych. Proces windykacji jest związany nie z kredytowaniem, ale z odzyskiwaniem długu. Może być on stosowany także wobec innych zobowiązań, nie tylko kredytowych. Pomimo umiejscowienia go w rozdziale 3, poświęconym procesom bankowym, jest to jednak proces ogólniejszy.

Proces biznesowy należy tu rozumieć jako: umiejętne biznesowe znalezienie różnicy pomiędzy przychodami i kosztami, która powoduje, że dany produkt czy usługa stają się rentowne. W dużych przedsiębiorstwach nie wystarczy analizować rachunku zysków i strat całej firmy. Wielokrotnie ze względu na strukturę firmy każdy pion, departament czy nawet wydział muszą same przed zarządem wykazać rentowność swoich procesów, stąd opisane procesy są przykładem raportowań finansowych, zarządczych, ujmujących pewien wycinek biznesu, który oddzielnie trzeba optymalizować. Pojęcia modelu biznesowego, owszem, lepiej używać w przypadku pełnych i rzeczywistych danych, ale niestety nikt takich danych nie pozwoli publikować. Pomysłem użytym w książce jest zatem stworzenie szablonu całego procesu biznesowego, łącznie z arkuszem kalkulacyjnym, ze wszystkimi regułami łączącymi wskaźniki i przykładowe parametry – możliwie najlepiej przybliżonymi do wartości rzeczywistych. Każdy potencjalny czytelnik książki może wprowadzić swoje dane do arkuszy i zbadać własne procesy z rzeczywistymi parametrami.

Problemy z tłumaczeniem

Termin „Credit Scoring" jest anglojęzyczny. Powinno się go pisać małymi literami, niestety zbyt często sięgamy dziś do literatury obcojęzycznej i musimy pogodzić się z pewnymi naleciałościami. Autor świadomie proponuje jako pierwsze słowa w tytule książki „Credit Scoring", bo one właśnie są najczęściej wpisywane do internetowych wyszukiwarek. Nie ma i nie będzie dobrego polskiego tłumaczenia tego terminu. Nikt nie zgodzi się na jego polski odpowiednik w postaci – metoda ocen punktowych.

Podstawowym modelem jest karta ocen punktowych, nikt nie mówi „karta punktowa". To tłumaczenie także budzi wątpliwości, dlatego wszyscy mówią „karta skoringowa". Można by tu zostawić literę „c", ale tak jak słowo „computer" już dawno w języku polskim zagościło jako „komputer", tak powoli możemy się pogodzić ze spolszczoną wersją karty skoringowej, modelu skoringowego, procesu skoringowego czy reguły skoringowej. Pewnie jeszcze długo będzie budzić zdziwienie słowo „skor" zamiast angielskiego *score*, czego raczej nie powinno się tłumaczyć, ale trudno będzie zrezygnować z pojęcia skorowania, czyli nadania klientom ocen punktowych. Mówimy o procesie skorowania i kodzie skoringowym oraz systemie skoringowym. Naprawdę niełatwo będzie zatrzymać skoringowe słowotwórstwo.

Na pytanie o zawód lub stanowisko osoby budującej modele najczęściej pada odpowiedź: Jestem skoringowcem. Buduję skoringi. W tym zlepku anglo-polskim kryje się więcej treści niż w jakimkolwiek innym tłumaczeniu. Niestety musimy się pogodzić z tym, że skoringowcy są wśród nas i skoringi są najlepszym narzędziem oceny zdolności kredytowej klienta.

Podobny problem wiąże się z tłumaczeniem pojęcia niewywiązania się ze zobowiązania kredytowego, nazywanego po angielsku *default*. Całe środowisko bankowców regularnie używa słowa *default*, mówi się także o statystyce *default rate* czy *bad rate*. Jeszcze większe problemy istnieją ze statystyką *lift* czy *gains*. Dlatego zdecydowano się na anglojęzyczne wersje dołączonych arkuszy kalkulacyjnych, gdyż stanowią one gotowe narzędzia do prezentacji w środowiskach międzynarodowych. Dodatkowo są pomocą dydaktyczną w precyzyjnym nazywaniu wskaźników w języku angielskim, co ma

także istotne znaczenie, gdyż ułatwia przeszukiwanie stron internetowych w celu studiowania materiału.

Struktura pracy

Praca podzielona jest na cztery rozdziały. W rozdziale 1 przybliżono tematykę modeli skoringowych w ujęciu historycznym oraz dokonano wprowadzenia w problematykę podstawowego procesu akceptacji kredytowej, w którym modele skoringowe znalazły swoje pierwsze zastosowania.

W następnym rozdziale wprowadza się wszystkie najważniejsze pojęcia i modele związane ze skoringiem. Przedstawione są typowe struktury danych i definicja zdarzenia niewywiązania się ze zobowiązania kredytowego. Omówione są wszelkie założenia i podstawy poprawnego wykorzystania modeli skoringowych w bankowości. W szczególności jest opisany model regresji logistycznej i zarysowana konstrukcja budowy karty skoringowej. Na końcu, w podrozdziale 2.4, jest omówiony sposób tworzenia danych symulacyjnych, na podstawie których w kolejnym rozdziale jest możliwe szczegółowe zaprezentowanie wszelkich istotnych problemów związanych ze stosowaniem modeli skoringowych.

W podrozdziale 3.1 jest po raz pierwszy przedstawiony model finansowy opłacalności procesu akceptacji kredytowej w kontekście stosowania modeli statystycznych. Tego typu opracowanie pozwala widzieć problemy bankowości w zupełnie innym świetle. Jednocześnie w ewidentny sposób można dostrzec potrzeby wdrażania coraz to większej liczby modeli predykcyjnych w celu optymalizacji procesów bankowych. Kolejny podrozdział pokazuje, że niektóre symulacje można przeprowadzić w bardzo uproszczony sposób, nie tracąc przy tym istoty rozumowania. Nie wchodząc w szczegóły dotyczące modelu statystycznego, można sprawdzić i przetestować wiele scenariuszy parametrów procesu, by wreszcie zdecydować, czy możliwy jest opłacalny proces w danym przypadku, czy można zarabiać na kredytach o niskim oprocentowaniu. Jaką moc predykcyjną powinny mieć modele predykcyjne, by proces był opłacalny? W podrozdziale 3.3 są omówione wszelkie sposoby liczenia statystyki mocy predykcyjnej, głównie statystyki Giniego, a także przedstawione popularne krzywe pomocne w ustalaniu punktu odcięcia czy

grupy docelowej, takie jak: ROC, CAP i Lorenza. W podrozdziałach 3.4 i 3.5 zostały omówione kolejne przykłady modeli biznesowych dla windykacji polubownej i procesu akceptacji kredytów hipotecznych. W ostatnim podrozdziale są przedstawione najczęstsze problemy występujące przy zarządzaniu złożonym procesem biznesowym – tanią akwizycją i drogą sprzedażą krzyżową. Opisane problemy uświadamiają Czytelnikowi, że sama budowa dobrych modeli nie wystarczy. Trzeba jeszcze umieć całym procesem zarządzać i rozumieć konsekwencje połączonego procesu dwóch produktów. Problem ten staje się bardziej widoczny i zmusza nas do wysiłku intelektualnego, by badać sposoby znajdowania złotego środka pomiędzy kosztem akwizycji a zarobkiem w sprzedaży krzyżowej, by umieć przewidywać zmiany rozkładów procesu przy zmianach strategii akceptacji.

Na zakończenie, w rozdziale 4, są pokazane zastosowania modeli skoringowych poza bankowością, głównie w procesach marketingowych czy w medycynie. Okazuje się, że modele te sprawdzają się równie dobrze także w innych procesach, pomagając przynosić milionowe zyski.

1. Rola i znaczenie modeli Credit Scoring w finansach

1.1. Historia i istota skoringu kredytowego

Pierwotnie Credit Scoring, tłumaczony często jako skoring kredytowy, był związany z procesem akceptacji wniosków kredytowych w bankach (Thonabauer i Nosslinger, 2004), używano tam prostych eksperckich kart skoringowych do wyznaczania oceny punktowej wniosku. Sposób naliczania punktów musiał być łatwy i umożliwiać nawet mniej wykwalifikowanym analitykom (których liczba wzrosła podczas II wojny światowej) 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 to pojęcie rozszerzyć na 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, co powoduje silną presję poszukiwania 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 dziś także w wielu innych procesach, w których klient podpisujący umowę, najczęściej zobowiązujący się do regularnych obciążeń 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. W niniejszej pracy zostaną też zaprezentowane zastosowania w telekomunikacji, marketingu i medycynie.

Dziś mówi się w kontekście *Big Data* o nowej erze, a analizy skoringowe są doskonałym tego przykładem, w szczególności stosowanym przy bardzo prostym procesie biznesowym. Ze względu na prostotę modeli skoringowych (głównie kart skoringowych) doskonale nadają się one dla początkujących, którzy chcą rozumieć, czym są analiza danych i jej zastosowania w biznesie, aby wyrobić sobie

ważne umiejętności i nie zgubić istoty problemu, 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 zdobyć. Właśnie w Credit Scoringu 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 i monitoring. Cały cykl życia modelu został właśnie tu poprawnie zdefiniowany i należy się tylko uczyć na podstawie Credit Scoringu oraz wcielać go w innych dziedzinach.

Początki Credit Scoringu sięgają lat 50. XX w., kiedy to firma konsultingowa o nazwie Fair Isaac & Company (dziś FICO) stworzyła pierwszy komercyjny system skoringowy (Poon, 2007). Pierwsze ważne argumenty dotyczące 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 3.1.

1.2. Rola i znaczenie modeli skoringowych

W pracy głównie koncertujemy się na statystycznych modelach oceny punktowej, zwanych także kartami skoringowymi (ang. *credit scorecard* lub ogólniej *Credit Scoring*) (Thomas et al., 2002; Anderson, 2007; Matuszyk, 2008). Najczęściej modele te są tworzone na bazie regresji logistycznej. Ich konstrukcja jest dość prosta oraz łatwa w interpretacji i dlatego stale są obecne 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 licze-

nia wymogów kapitałowych RWA (ang. *Risk Weighted Assets*) (BIS-BASEL, 2005).

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 przypadkach skrajnych, aby obserwować pełny cykl życia nawet zwykłych kredytów, takich jak kredyt ratalny, potrzeba przynajmniej 5, a może i 10 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, obecnie odnotowuje się wyjątkowo niskie wartości ryzyka kredytów konsumenckich. Nikt jednak nie jest w stanie zagwarantować tego, że kryzys nie powróci. Konsekwencje rekomendacji T, którą wydała Komisja Nadzoru Finansowego (KNF) i która spowodowała rozwinięcie się parabanków, ciągle nie są do końca zbadane. Pojawia się ciekawy problem niereprezentatywności danych rynku kredytowego w bazach Biura Informacji Kredytowej (BIK) i warto jemu poświęcić obszerniejsze badania. Obecny kryzys ekonomiczny skłania także wielu badaczy ku poszukiwaniu lepszych modeli predykcyjnych, bardziej stabilnych w czasie (Mays, 2009).

Model skoringowy stał się najlepszym narzędziem mierzącym i prognozującym ryzyko kredytowe klienta używanym masowo w zarządzaniu procesami biznesowymi. Sam tytuł książki *Credit-scoring: nowoczesna metoda oceny zdolności kredytowej* (Janc i Kraska, 2001) mówi sam za siebie. Co ciekawe, w dobie rekomendacji T mamy proste rozróżnienie na rynek regulowany i parabankowy. W tym pierwszym dużą część odmów kredytowych w procesie akceptacji stanowi kryterium zdolności kredytowej, czyli specyficzny wskaźnik finansowy związany ze stosunkiem zobowiązań klienta i kosztów do jego dochodu, a w przypadku parabanków odmowy następują głównie ze względu na modele skoringowe. W tej sytuacji właśnie parabanki

uczą nas, jak mierzyć zdolność kredytową i ryzyko kredytowe (dokładnie – pożyczkowe).

Jak na razie nie wymyślono lepszego narzędzia do oceny ryzyka, aczkolwiek modele wartości życiowej klienta (ang. *Customer Life Time Value* – CLTV) potrafią być jeszcze dokładniejsze, gdyż uwzględnia się tu zarówno ryzyko liczone jako wartość straty kredytowej, jak i przychody osiągnięte przez bank z danego klienta w całej jego przyszłej historii relacji (Ogden, 2009; DeBonis et al., 2002).

Niestety sama ocena punktowa klienta nie zawsze w pełni określi ryzyko. Wiele zmian powodują zewnętrzne czynniki, głównie cała koniunktura i powiązania rynków finansowych. Uwzględnienie tego w prognozowaniu ryzyka nie jest już możliwe w ramach prostych technik Credit Scoring i wymaga użycia bardziej zaawansowanych modeli, takich jak analizy historii zdarzeń (ang. *survival analysis*) ze zmiennymi zależnymi od czasu (Bellotti i Crook, 2009). Temat ten jednak znacząco wykracza poza obszar niniejszego opracowania.

1.3. Proces akceptacji kredytowej

Proces akceptacji wniosków kredytowych w bankach pełni jedną z kluczowych funkcji w zarządzaniu portfelem detalicznym (Thonabauer i Nosslinger, 2004), szczególnie dla *Consumer Finance*, czyli drobnych kredytów konsumenckich. Jeśli liczba wniosków kredytowych w miesiącu potrafi przekroczyć kilkadziesiąt tysięcy, to w procesie tym są potrzebne narzędzia statystyczne. Im więcej jest wniosków, tym większą rolę powinny odgrywać automatyczne decyzje podejmowane przez zaawansowane modele statystyczne.

W akceptacji kredytowej niezbędne są systemy informatyczne. Wnioski kredytowe wprowadza się do aplikacji Front-End, gdzie są przygotowane wszystkie pola potrzebne do identyfikacji wnioskowanego produktu i wnioskującego klienta. Wprowadzane są tu także różnego rodzaju dane, które są bardzo pomocne w podejmowaniu decyzji. Najczęściej zbiera się dane socjodemograficzne, takie jak: kod zawodu, rodzaj umowy o pracę, status małżeński, status mieszkaniowy, wynagrodzenie, liczba osób na utrzymaniu, dane teleadresowe itp.

Jeśli klient jest już znany na rynku bankowym, to istotną rolę odgrywają też dane pobierane z zewnętrznych baz bankowych, są nimi raport kredytowy BIK, a także różnego rodzaju bazy, w których zbiera się informacje o zastrzeżonych dokumentach i klientach nierzetelnych. Potrzebne są zatem systemy informatyczne umożliwiające pobieranie on-line (w czasie rzeczywistym) wszelkich danych z zewnętrznych źródeł. Podobnie dobrą praktyką jest weryfikowanie klienta w wewnętrznych bazach banku. W zależności od produktu udział w populacji wnioskujących klienta, który już raz aplikował o dany kredyt w historii, potrafi przekroczyć nawet 50%. Zasadne zatem staje się sprawdzenie, jak klient ten spłacał lub spłaca swoje poprzednie lub aktualne kredyty w naszym banku, a w przypadku dobrej historii spłacania analizowanie jego przypadku specjalnym przyspieszonym i uproszczonym procesem dla wybranych, znanych klientów.

Wszystkie dotychczasowe narzędzia były związane z wprowadzaniem wniosku i gromadzeniem dodatkowych danych. Jeśli wniosek jest już kompletny, to jest możliwa jego analiza i następuje podjęcie decyzji. Najczęściej następuje to poprzez narzędzie, zwane systemem decyzyjnym (ang. *decision engine* lub *scoring engine*). Decyzja jest podejmowana w wielu krokach. Każdy krok wiąże się z właściwie przygotowanym i przetestowanym zestawem reguł decyzyjnych.

Na początku są weryfikowane aspekty prawne, związane z bezpieczeństwem banku i klienta oraz z upewnieniem się, czy wnioskodawca podał poprawne informacje. Następnie najczęściej są sprawdzane reguły identyfikujące nierzetelnych lub podejrzanych klientów, co określa się zbiorczą nazwą – „weryfikacja na czarnych listach" (ang. *black lists*). Na tym etapie jest możliwe także przejście na tryb ręczny. Jeśli pojawi się podejrzenie, że dane z wniosku są niepoprawne, że źle identyfikujemy klienta, wtedy mogą pojawić się dodatkowe czynności sprawdzające, wykonywane przez specjalnie wyznaczonych w banku weryfikatorów. Mogą oni dzwonić do pracodawcy albo na wskazane numery telefonów do domu. Wszystkie tego typu czynności powinny zminimalizować ryzyko nadużyć. Trzeba pamiętać o tym, że w procesie bierze udział wiele stron. Nieuczciwość może pojawić się zarówno po stronie klienta, jak i po stronie wprowadzającego wniosek (w tym drugim przypadku ze względu na

system premiowy i presję działów sprzedaży oczekujących realizacji zamierzonych planów).

Kolejnym etapem w procesie może być weryfikowanie reguł związanych ze wszelkimi rekomendacjami nałożonymi przez nadzorcę, czyli KNF (Komisję Nadzoru Finansowego). Bankowy rynek kredytów jest rynkiem regulowanym i nadzorca ma prawo chronić konsumenta. Najczęściej rekomendacje odnoszą się do wytycznych ograniczających sprzedaż kredytów klientom bardzo przekredytowanym lub kredytów ze zbyt dużym kosztem, porównywanym do lichwy.

Najistotniejszym etapem jest wykorzystanie wszelkich narzędzi skoringowych do poprawnego określenia akceptowanego portfela, by cały proces był opłacalny, co w skrócie oznacza takie wybranie wniosków z całej populacji przychodzącej (wnioskującej), by większość z nich się spłacała, czyli przychody z ich udzielania pokryły z nawiązką stratę powstałą przez klientów niespłacających terminowo.

Możliwa jest też sytuacja, kiedy decyzja z automatycznego procesu jest przełamywana przez analityka kredytowego. Może to mieć miejsce szczególnie przy produktach takich, jak kredyt hipoteczny czy na zakup samochodu. Zdarza się, że analityk lub osoba wprowadzająca wniosek w oddziale banku zna już dość dobrze klienta lub że automatyczna odmowa następuje z powodu, który analityk może zweryfikować i uzasadnić jego niesłuszność.

Wszystkie kroki sprawdzanych reguł i interakcji z osobami biorącymi udział w procesie powinny być rejestrowane i zapisywane w bazie danych banku. Każdy stan wniosku, reguły i działania analityka powinny być odnotowane w systemie. Tylko wtedy możliwe jest uczenie się na błędach i nieustające poprawianie procesu. Można też zmieniać kolejność reguł lub nawet szukać właściwej albo też badać proces poprzez scenariusze, gdzie pewnych reguł nie ma lub są zmodyfikowane.

Ostatnimi elementami procesu są narzędzia umożliwiające wydrukowanie umowy, harmonogramu, założenie konta kredytowego i wreszcie uruchomienie środków dla klienta.

W dalszej części pracy omówienie systemu decyzyjnego ogranicza się do sprawdzania reguł skoringowych, gdyż one odgrywają

kluczową rolę w przymnażaniu kapitału banku, a zarazem są najtrudniejsze w zarządzaniu.

1.4. Przykładowa karta ocen punktowych

Szczegółowy sposób budowania kart ocen punktowych (kart skoringowych) metodą LOG jest oparty na regresji logistycznej, drzewach decyzyjnych i transformacji logit: każdej kategorii zmiennej jest przypisana jej wartość logit (Przanowski, 2014a). Jest ona prawie identyczna z metodą WoE stosowaną w SAS Credit Scoring Solution (Siddiqi, 2005).

Przykładowa postać karty ocen jest przedstawiona w tabeli 1. W prezentowanej karcie ocen są uwzględnione dwa predyktory – wiek i wynagrodzenie. Obie zmienne są podzielone na trzy kategorie rozłączne, dla uproszczenia podano tu tylko prawe granice przedziałów, środkowa kategoria wiekowa poprawnie powinna być określona przez dwie nierówności: $20 < \text{wiek} \leqslant 35$. Każdej kategorii są przypisane oceny cząstkowe. Finalna ocena punktowa danego klienta jest liczona jako suma ocen cząstkowych wynikających z właściwych kategorii.

Tabela 1. 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.

Forma karty ocen jest bardzo prosta i interpretowalna. Stąd jej powszechne stosowanie i możliwość wykorzystywania w wielu procesach także tam, gdzie wszystkie oceny punktowe obliczano wy-

łącznie ręcznie. Porównywanie ocen cząstkowych daje możliwość identyfikacji ważniejszych i mniej istotnych zmiennych w modelu. Pozwala też identyfikować kierunek zmian. Im wyższa jest ocena punktowa, tym lepszy jest klient, tym większa jest szansa spłacenia zobowiązania kredytowego w terminie bez żadnych opóźnień. Jeśli zatem oceny cząstkowe rosną wraz ze wzrostem wieku, oznacza to, że im starszy jest klient, tym jest mniej ryzykowny.

2. Metodologiczne podstawy wykorzystania modeli skoringowych

2.1. Podstawowe struktury danych i pojęcia

Omawiane w pracy metody statystyczne są przykładem typowych modeli predykcyjnych. Istota modelowania polega na odkryciu reguł, zależności pomiędzy zmiennymi niezależnymi, zmiennymi objaśniającymi (cechami lub charakterystykami), zwanymi także predyktorami (ang. *predictors*), a funkcją celu (ang. *target variable*), nazywaną zmienną objaśnianą. Liczba zmiennych może być dość duża i w niektórych firmach przekracza kilka tysięcy. Przygotowanie takiej struktury jest bardzo złożonym procesem i zajmuje średnio 80% całego czasu budowy modelu. Zbiór danych ze zmiennymi i funkcją celu zbiorczo nazywa się tabelą analityczną (ang. *analytical base table* – ABT), pojęcie wprowadzone przez firmę SAS Institute w ramach narzędzia SAS Credit Scoring Solution. Wierszem takiej tabeli jest jeden historyczny przypadek badanego zjawiska, zdarzenia. Funkcją celu jest kolumna zawierająca tylko dwie wartości: nastąpiło zdarzenie lub nie (w języku statystycznym mówi się, że zmienna odpowiedzi jest dwuwartościowa, binarna lub dychotomiczna).

Bardzo ważnym pojęciem w modelowaniu i strukturze danych jest zdarzenie modelowe. W przypadku bankowości i ryzyka kredytowego jest to zdarzenie niewywiązania się ze zobowiązania kredytowego (ang. *default*). Rozważany jest tu przypadek związany z definicją aplikacyjną, czyli z procesem akceptacji kredytowej. Wszystkie informacje o kliencie i jego wniosku kredytowym (aplikacji) zebrane przed i w trakcie aplikowania są danymi, na bazie których można wyliczać zmienne ABT. Należy podkreślić fakt, że klient aplikujący o kredyt może być już znany bankowi lub na rynku bankowym (co można sprawdzić w raporcie BIK), gdyż może składać wniosek o kolejny kredyt. Im jest więcej kredytów w jego historii, tym więcej można wyznaczyć zmiennych behawioralnych, czyli opartych na zachowaniu klienta. Jeśli jest to pierwszy wniosek klienta, mamy tylko informacje pochodzące z wniosku, których jest stosunkowo niewiele

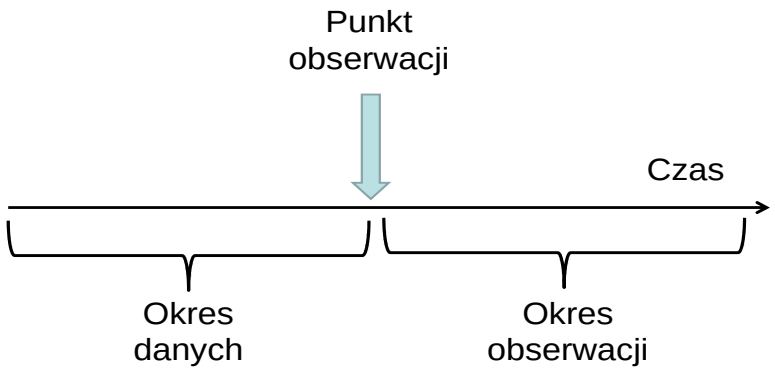
i które nie są często wiarygodne. Wiele informacji klient sam deklaruje, są to takie dane, jak: liczba osób na utrzymaniu, status małżeński, mieszkaniowy czy nawet wynagrodzenie.

Od momentu aplikacji, czyli „punktu obserwacji" (rysunek 1, str. 35), w „okresie obserwacji" (ang. *outcome period*), najczęściej w ciągu 12 miesięcy, badamy zajście zdarzenia, w naszym przypadku niewywiązania się ze zobowiązania, czyli posiadania więcej niż 90 dni opóźnienia w spłatach rat kredytowych. Innymi słowy, od momentu aplikacji badamy, czy w ciągu 12 miesięcy klient wpadł w opóźnienie większe niż 90-dniowe, czyli obejmujące trzy niespłacone raty. Takie zdarzenie modelowe nazwiemy w skrócie „aplikacyjną definicją *default*". W tym ujęciu dany wniosek kredytowy pojawia się tylko raz w jednym wierszu tabeli ABT.

Można także zdefiniować behawioralną definicję *default*. W tym wypadku w punkcie obserwacji rozważamy wszystkie „zdrowe", czyli jeszcze bez opóźnień, rachunki kredytowe. Punktem obserwacji jest najczęściej koniec miesiąca. Co miesiąc wszystkie „zdrowe" rachunki są badane pod kątem wywiązywania się ze zobowiązania. Rachunki, które cechuje zbyt duże prawdopodobieństwo wpadnięcia w opóźnienie, muszą być zidentyfikowane i powinna być dla nich wyliczona rezerwa finansowa. Dlatego cyklicznie model taki stosuje się wobec wszystkich „zdrowych" rachunków. Oznacza to, że dany rachunek może pojawić się wiele razy w ABT dla różnych miesięcy. Podobnie jak w definicji aplikacyjnej tu także obserwuje się rachunki w ciągu 12 miesięcy od punktu obserwacji i bada się wejście w opóźnienie powyżej 90 dni. W przypadku definicji aplikacyjnej będziemy interpretowali parametr PD zwrócony przez model skoringowy jako prawdopodobieństwo wejścia w opóźnienie większe od 90-dniowgo (90+) od daty aplikacji w ciągu 12 miesięcy pod warunkiem, że klient otrzyma od banku nowy kredyt, o który właśnie się starał. W przypadku definicji behawioralnej będzie to prawdopodobieństwo wejścia w *default* pod warunkiem, że posiada takie, a nie inne kredyty w punkcie obserwacji, czyli w danym miesiącu życia banku. Jest to zatem zupełnie inne warunkowanie, o którym niestety praktycy czasem zapominają.

Należy bardzo przestrzegać warunku, by wszystkie dane do wyliczeń ABT pochodziły z danych gromadzonych przed datą wnio-

Rysunek 1. Elementy definicji zdarzenia modelowego



Źródło: opracowanie własne.

skowania lub z samego wniosku, czyli z „okresu danych" (Provost i Fawcett, 2014), każda informacja pozyskana o kliencie później wprowadza istotny błąd w modelowaniu i może całkowicie przekreślić poprawność metody oraz wyników modelu. Co gorsze, wszelkie informacje istniejące po dacie wniosku, nazywane często informacjami wziętymi z przyszłości, powodują, że modele zyskują na swojej mocy predykcyjnej i potrafią prognozować z bardzo dużą, aż niewiarygodną dokładnością, ale tylko na danych historycznych. Przypuśćmy, że chcemy prognozować odchodzenie klientów, czyli zdarzenie rozwiązania umowy w ciągu 6 miesięcy od jej podpisania. Jeśli do danych ABT dodamy informację o wykonanym telefonie klienta do Call Center, podczas której zadeklarował chęć rozwiązania umowy, to zmienna identyfikująca to zdarzenie na pewno zostanie wybrana do modelu, bo raczej na 90% klient taki po pewnym czasie umowę rozwiąże. Model zatem absurdalnie będzie działał, testując dane historyczne zarówno o rozwiązanych umowach, jak i wykonanych wcześniej telefonach, ale niestety nigdy nie pomoże on nam przewidywać odejścia klientów podpisujących umowy dziś. Istota takiego modelowania powinna polegać na przewidzeniu odejścia, zanim nastąpi jakakolwiek akcja klienta zmierzającego do rozwiązania umowy. Klient wnioskujący, czyli podpisujący umowę dziś, nie ujaw-

nia chęci odejścia, bo właśnie deklaruje przystąpienie do umowy. Innym przykładem może być model prognozowania śmierci klienta. Najlepszy predyktor „wzięty z przyszłości" to sprawdzenie, czy data zgonu jest niepusta. W rzeczywistości prognozuje się zgon na zbiorze klientów, gdy wspomniana data jest tylko pusta. Choć przytoczone przykłady wydają się oczywiste, to jednak przy rzeczywistych problemach łatwo o pomyłkę. Trzeba bardzo dobrze rozumieć proces, którym zarządzamy, zanim zbudujemy ABT i zdarzenie modelowe. Mniej oczywistym przykładem może być sytuacja wykorzystania w modelowaniu informacji o nazwie banku z rachunku ROR (rachunku oszczędnościowo-rozliczeniowego) wnioskującego klienta. Przypuśćmy, że klient wnioskujący o kredyt dopiero po otrzymaniu akceptacji podaje numer rachunku, na który należy wykonać przelew. Informacja ta jest jednym słowem zgromadzona później niż wydawana decyzja kredytowa. Jeśli analityk budujący model nie pozna procesu, to analizując dane historyczne, może odnieść wrażenie, że numer rachunku istniał już przed decyzją i może go wykorzystać w identyfikacji banku. Może się okazać, że nawet nazwa banku będzie dobrym predyktorem. Niestety taki model potem nie będzie mógł być wdrożony, gdyż okaże się, że danej nie da się pozyskać w momencie podejmowania decyzji.

Dziś bardzo wiele firm w ramach swoich zespołów analitycznych utrzymuje i nieustająco rozwija ABT. Staje się ona jednym z istotnych aktywów firmy, choć niestety mało jeszcze docenianym przez jej zarząd. Budowa dobrej ABT gwarantuje szybkie i poprawne budowanie nowych modeli. Daje też możliwość weryfikacji poprawności danych, czyli ich jakości. Z roku na rok temat ten staje się coraz modniejszy. Wiele już napisano o jakości i pewnie jeszcze wiele informacji zostanie usystematyzowanych. Warto jednak pamiętać o kilku prostych przykładach. Jednym z najczęstszych błędów jest złe kodowanie wartości zero lub braku danych. Przypuśćmy, że liczymy średnią wartość limitów kart kredytowych klientów w naszym banku. Jeśli tylko połowa z nich posiada kartę, a druga ma wartość limitu zero zamiast braku danych, to średni limit będzie dwa razy mniejszy od spodziewanego. Tak prosty przypadek, a tak fałszywy wniosek.

Istotą bogatej ABT, zawierającej większość informacji o badanym zjawisku, jest możliwość weryfikacji i określenia, w jakim stopniu wartości funkcji celu są możliwe do przewidzenia. Jeśli zbada się wszystkie możliwe informacje zebrane w dostępnych bazach, wtedy ma się pewność, że nie da się zbudować lepszego modelu. Oczywiście problem dotyczy stwierdzenia „wszystkie możliwe", jeśli bowiem uwzględnia się tę samą informację, to i tak można zbudować różnego rodzaju zmienną, raz może to być średnie saldo klienta w ciągu ostatnich 12 miesięcy, a raz maksymalne. Niby ta sama informacja, a jednak może być lepszym lub gorszym predyktorem. Trzeba wielu lat doświadczeń i testów, by wyrobić w sobie cenną umiejętność budowania zmiennych ABT.

2.2. Statystyczne podstawy modelu skoringowego

Model wylicza się (w języku statystycznym: estymuje) na podstawie danych historycznych. Oznacza to, że dane są informacje zarówno sprzed daty wnioskowania, z daty wniosku, jak i z okresu po tej dacie. Mamy więc policzone wartości funkcji celu. Wiemy zatem, przy jakich danych klient spłacał kredyty, a przy jakich miał opóźnienia. Posiadając takie dane, możemy odkryć reguły uzależniające zdarzenie *default* od zmiennych ABT. Można odkryć np. reguły: że klienci młodsi gorzej spłacają kredyty od starszych, że emeryci są rzetelnymi klientami, a osoby z małym stażem pracy mogą mieć problemy w regularnych spłatach. Wykrycie reguł niestety nie gwarantuje poprawnego działania modelu stosowanego do teraźniejszych danych. Wychodzi się tu z założenia, że przeszłość ma wpływ na przyszłość i że prognozowane zjawisko jest związane z informacjami gromadzonymi w ABT. Niestety istnieją zdarzenia, których nie da się w pełni prognozować, w takiej sytuacji nawet nie wolno ulegać pokusie, że to tylko kwestia danych czy technik modelowych. Wygranej w LOTTO nie da się przewidzieć. Wielu zmian na rynkach finansowych czy na giełdzie nie powinno się prognozować, gdyż są to zjawiska nie w pełni deterministyczne. Nie da się poprawnie prognozować zachowania ludzi, gdyż czasem postępują irracjonalnie. Łatwiej jest jednak prognozować zachowanie dużej zbiorowości, gdyż tam większość postępuje racjonalnie. Ta właśnie

zasada jest gwarancją poprawności modelowania zdarzenia *default* w bankowości. Oznacza ona także, że modele skoringowe spełniają swoje zadanie tylko przy zjawiskach masowych, czyli wówczas, gdy liczba klientów jest duża, gdy zaczynają działać prawa statystyczne. Trudno jest podać jednoznaczną definicję dużej liczby klientów, ale przyjmuje się, że metody skoringowe są głównie stosowane wobec portfeli detalicznych, włączając w to także SME (małe i średnie przedsiębiorstwa). W przypadku klientów i portfeli korporacyjnych metody skoringowe najczęściej są korygowane dodatkowymi, eksperckimi metodami indywidualnie wobec każdego klienta lub zastępowane agencjami raitingowymi.

Jeśli zatem model buduje się na historycznych danych, to znaczy, że istnieje ryzyko błędu wynikające z różnicy populacji dzisiejszej i modelowej, czyli tej, na której model budowano. Bardzo istotnym problemem w budowie modelu jest dobranie populacji modelowej, tak by była ona najbardziej podobna do obecnej. Jeśli mamy zbudować model prognozujący *default* w ciągu 12 miesięcy od wnioskowania kredytu, to oznacza, że najświeższa informacja z dostępną daną funkcją celu pochodzi od populacji wnioskującej rok temu. Mamy zatem od razu różnicę jednego roku. Aby model był stabilny, trzeba mieć w populacji modelowej pewien odcinek czasowy, jest to czasem kilka miesięcy lub nawet kilka lat. Mamy więc wówczas jeszcze starszą populację. Bywa że tak starej historii kredytowej nie posiadamy w naszym banku. Wtedy wybiera się definicję funkcji celu, badając zajście zdarzenia *default* w krótszym horyzoncie czasowym, np. 6 miesięcy. Trzeba umiejętnie wybrać najlepszy scenariusz, manipulując albo odcinkiem czasowym populacji modelowej, albo horyzontem definicji *default* (*outcome period*). Tylko dogłębne poznanie procesu, zmian rynku, koniunktury i wielu jeszcze innych aspektów pozwoli poprawnie wybrać parametry danych i zbudować podstawowe struktury potrzebne do modelowania.

W efekcie model statystyczny każdemu klientowi wyznacza ocenę punktową (ang. *scorecard points* lub *score*). Ocena ta jest miernikiem zdolności kredytowej, czyli zdolności do spłacania zobowiązań w terminie. Budowa modelu polega zatem na znalezieniu wzoru wyliczania oceny. Finalnie wzór ten może być w miarę prosty i stanowić kombinację wag zmiennych. Dobranie tych wag i wybór zmiennych

stanowią właśnie największą trudność i są możliwe tylko dzięki wykorzystaniu zaawansowanych modeli statystycznych.

Posiadając już algorytm wyznaczania oceny punktowej, dość łatwo możemy wdrożyć model w systemie decyzyjnym. Nie wprowadza się tam zaawansowanych procedur statystycznych, ale tylko finalną postać wzoru, który często może być obliczony przy użyciu prostych narzędzi informatycznych, aczkolwiek niektóre modele skoringowe wymagają dość wydajnych systemów (jeśli zmienne są behawioralne i wyliczają różne średnie kroczące, to może się okazać, że modelu nie udaje się wdrożyć ze względu na wydajnościowe aspekty). Tego typu problem powinien być rozstrzygnięty przed procesem budowy modelu, przez zdefiniowanie jego kryteriów akceptacji (ang. *minimal requirements*).

Podstawowym modelem statystycznym do budowy modeli skoringowych jest naiwny klasyfikator Bayesa, rozwiązujący – jak sama nazwa wskazuje – problem klasyfikacji (Ćwik i Koronacki, 2005), czyli podjęcia decyzji, do której klasy (kategorii) ma należeć dany obiekt, charakteryzujący się danym zestawem zmiennych predyktorów. Wiąże się z nim założenie, że zmienne ABT są niezależne, co jest dość poważnym praktycznym problemem, gdyż w rzeczywistości niektóre zmienne w naturalny sposób od siebie zależą. W związku z tym najczęściej konstrukcję kart skoringowych (ang. *scorecard*) buduje się na podstawie modelu regresji logistycznej razem z prostymi algorytmami drzew decyzyjnych do kategoryzacji zmiennych ciągłych. W tym wypadku założenie o niezależności także odgrywa rolę w samej konstrukcji modelu, ale w efekcie powstaje dość odporny model, który najczęściej weryfikuje się pod kątem minimalizowania wyłącznie liniowej zależności predyktorów.

2.3. Binarna regresja logistyczna i drzewa decyzyjne

Modele regresyjne, w szczególności regresja liniowa, znamy już od setek lat, są związane z takimi twórcami, jak Adrien-Marie Legendre i Carl Friedrich Gauss, którzy napisali swoje największe dzieła w XVIII w. Wtedy powstała metoda najmniejszych kwadratów. Model ten, stosowany do dzisiaj, pozwala wyznaczyć zależność po-

między funkcją celu a predyktorami, przy założeniu że funkcja celu posiada rozkład ciągły i normalny. Niestety modelowanie zdarzenia *default* nie spełnia tych założeń. Przez wiele lat – pomimo łamania założeń – statystycy używali regresji liniowej, aż do czasu powstania lepszej metody – regresji logistycznej.

W celu jej zrozumienia na początku trzeba zdefiniować rozkład zero-jedynkowy. Rozważmy zdarzenie losowe polegające na zajściu zdarzenia *default* lub jego braku. Zmienna losowa Y przyjmuje zatem tylko dwie wartości $Y = 1$ lub $Y = 0$, gdzie wartość 1 utożsamiamy z zajściem zdarzenia *default*. Zajście zdarzenia posiada określone prawdopodobieństwo, które oznaczamy przez p, mamy zatem: $p = P(Y = 1)$. Prawdopodobieństwo zdarzenia przeciwnego, czyli braku *default*, można łatwo obliczyć: $P(Y = 0) = 1 - P(Y = 1) = 1 - p$. Przypuśćmy teraz, że zmienna losowa Y ma swoją realizację y, innymi słowy – została zaobserwowana jej wartość (wykonano pomiar). Obliczmy teraz prawdopodobieństwo zaobserwowania tej wartości. Możemy to zapisać w dwóch wariantach:

$$P(Y = y) = \begin{cases} p, & \text{gdy} \quad y = 1, \\ 1 - p, & \text{gdy} \quad y = 0, \end{cases}$$

albo w postaci jednego wzoru:

$$P(Y = y) = p^y (1 - p)^{(1-y)},$$

a po przekształceniach w finalnej wersji:

$$P(Y = y) = \exp\left(y \ln\left(\frac{p}{1-p} \right) + \ln(1 - p) \right).$$

Pojawia się tu po raz pierwszy człon definiujący funkcję logitową:

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p} \right),$$

która staje się ważnym elementem regresji logistycznej.

Rozważmy teraz sytuację bardziej ogólną. W naszej próbie losowej, zawierającej historyczne dane, zaobserwowaliśmy N obserwacji. Każda obserwacja funkcji losowej Y_n, związana ze statusem

zdarzenia *default*, ma wartość y_n, gdzie n jest numerem obserwacji. Interesującym nas modelem jest wyjaśnienie zależności pomiędzy prawdopodobieństwem zajścia zdarzenia *default*, co często matematycznie zapisuje się jako $p_n = P(Y_n = 1)$, a predyktorami oznaczanymi jako ciąg zmiennych $x_n^1, x_n^2, ..., x_n^m$, gdzie m jest liczbą zmiennych w ABT. Na początku definiuje się część regresyjną, czyli kombinację predyktorów:

$$X_n\beta = \sum_{i=0}^{m} \beta_i x_n^i = \beta_0 + \beta_1 x_n^1 + \beta_2 x_n^2 + ... + \beta_m x_n^m.$$

Kombinacja ta jest związana nieliniową zależnością z prawdopodobieństwem p_n. Funkcji wiążących można zdefiniować dość dużo, na podstawie praktyki najpopularniejszą stała się funkcja logitowa (a zależność nazwano „sigmoid"). Swoją popularność zawdzięcza możliwości interpretacji członu $\frac{p_n}{1-p_n}$, który nazywa się szansą zajścia zdarzenia *default* (ang. *odds*), jest to stosunek prawdopodobieństwa zajścia zdarzenia do prawdopodobieństwa zdarzenia przeciwnego. Mamy zatem model, który uzależnia logarytm naturalny z szansy albo logit z prawdopodobieństwa zajścia zdarzenia *default* od członu regresyjnego $X_n\beta$. Finalnie więc jest estymowane następujące równanie:

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

gdzie X_n są danymi wartościami predyktorów, p_n są teoretycznymi wartościami prawdopodobieństw zajścia zdarzenia *default* dla n-tej obserwacji, a wektor współczynników β jest szukany.

Od funkcji logit pochodzi też sama nazwa modelu regresji logistycznej (Hosmer i Lemenshow, 2000), czasem nazywanego modelem logitowym. Inne funkcje wiążące, jak i założenia co do rozkładów funkcji celu zostały uwzględnione w uogólnionych modelach liniowych (Dobson, 2002; Ptak-Chmielewska, 2013).

Współczynniki β_i są obliczane (estymowane) na podstawie metody największej wiarygodności. Różni się ona od wcześniej znanej metody najmniejszych kwadratów i jest niestety związana z bardziej złożonym algorytmem poszukiwania maksimum funkcji. Znajduje się je metodą iteracyjną, w każdym kroku przybliżając się do wyniku z coraz większą dokładnością. Jeśli kolejne kroki powodują, że

zmiana wyniku jest mniejsza od ustalonej dokładności, to algorytm się zatrzymuje i rozwiązanie jest znalezione. W przeciwnym przypadku algorytm jest rozbieżny i niestety trzeba wtedy zmienić lekko parametry wejściowe. Najpopularniejszym algorytmem jest metoda Newtona–Raphsona, która kolejne iteracje wyznacza, poruszając się po wektorze wyznaczonym przez gradient funkcji wiarygodności.

Metoda największej wiarygodności, opisana przez R.A. Fishera w XX w., jest oparta na bardzo prostym i uzasadnionym przesłaniu, że prawdopodobieństwo uzyskania takich, a nie innych wartości obserwacji w próbie musi być największe. Gdyby było inaczej, to otrzymalibyśmy inne wartości obserwacji. Mamy zatem, wykorzystując założenie o niezależności zaobserwowanych zdarzeń (czyli że prawdopodobieństwo zajścia kilku zdarzeń jednocześnie jest równe iloczynowi ich prawdopodobieństw):

$$P(Y_1 = y_1, Y_2 = y_2, ..., Y_N = y_N) =$$

$$P(Y_1 = y_1) \cdot P(Y_2 = y_2) \cdot ... \cdot P(Y_N = y_N) =$$

$$\prod_{n=1}^{N} P(Y_n = y_n) =$$

$$\prod_{n=1}^{N} \exp\left(y_n \ln\left(\frac{p_n}{1-p_n}\right) + \ln(1-p_n)\right) =$$

$$\exp\left(\sum_{n=1}^{N}\left(y_n \ln\left(\frac{p_n}{1-p_n}\right) + \ln(1-p_n)\right)\right).$$

Funkcją wiarygodności jest właśnie prawdopodobieństwo zaobserwowania wszystkich razem wartości y_n. Przykładając zatem dodatkowo funkcję logarytmu i wstawiając za logity odpowiednie człony regresyjne, otrzymamy finalną postać logarytmu z funkcji wiarygodności (ang. *likelihood* – L):

$$\ln(L(\beta)) = \sum_{n=1}^{N} \left(y_n X_n \beta - \ln\left(1 + \exp(X_n \beta)\right)\right).$$

Istotą metody maksimum wiarygodności jest zatem znalezienie takiego wektora współczynników β, by logarytm z funkcji wiarygodności był największy.

Wyrażenie regresyjne $X_n\beta$ jest oceną punktową. Najczęściej dokonuje się tu dodatkowych prostych przekształceń, by ocena ta była całkowita i miała lepszą interpretację (Przanowski, 2014a).

Dodatkowo ocena ta musi być rozbita na oceny cząstkowe związane z kategoriami predyktorów. Każdy predyktor, niezależnie od tego, czy jest zmienną ciągłą, czy nominalną (przykładem zmiennej ciągłej jest wiek, a nominalnej nazwa miasta), finalnie jest kategoryzowany, czyli zamieniany na zestaw od kilku do maksymalnie kilkunastu kategorii. W przypadku zmiennej nominalnej czasem potrzebne jest łączenie kilku wartości w jedną kategorię, a w przypadku zmiennej ciągłej trzeba znaleźć punkty podziałowe, np. aby podzielić na dwie grupy młodszych i starszych, trzeba określić granicę wieku. Łączenie wartości lub szukanie punktów podziałowych najczęściej wykonuje się algorytmami drzew decyzyjnych (klasyfikacyjnych) (Kamiński i Zawisza, 2012), wyliczając statystyki mierzące poziom jednorodności uzyskiwanych grup, takie jak entropia i indeks Giniego.

2.4. Dane symulacyjne. Opis algorytmu generowania danych

Podstawowe idee algorytmu zostały opublikowane przez autora niniejszej książki (Przanowski, 2013), a później przez niego rozwinięte (Przanowski, 2014a). Dane tworzone są miesiąc po miesiącu. W każdym miesięcznym etapie tworzenia danych są modyfikowane 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 przez określenie tych trzech nowych wartości zmiennych dla każdego rachunku. Na początku jest obliczany model skoringowy, 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, którzy klienci w następnym miesiącu spłacą raty, a którzy wpadną w więk-

sze zadłużenie. Mechanizm jest zatem związany z łańcuchem Markowa i modelem skoringowym. Zmiany cech klienta są także dokonywane przez odpowiednie macierze przejść, które powodują, że klientowi powiększa się lub zmniejsza wynagrodzenie, powiększa się lub zmniejsza liczba dzieci itp.

Zastosowania Credit Scoringu 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, którą określa we wniosku kredytowym. Nie można jednak każdemu historycznemu rachunkowi kredytowemu nadawać takiej samej wagi, inaczej w przypadku każdego klienta w dłuższym lub krótszym czasie pojawiłyby się opóźnienia i nie spłacałby kredytów. Muszą zatem istnieć priorytety, którymi kieruje się klient przy spłacaniu rat. Jest powszechnie znany fakt, 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 żelazka. Priorytety w dużej mierze są więc związane z samymi procesami kredytowymi i sposobami zabezpieczenia kredytów. Pojawiają się tu także nieracjonalne upodobania i przywiązanie klienta do marki, do zaufanej pani w okienku i wiele innych subtelności, których nie da się uwzględnić w modelowaniu. Odwołanie się do priorytetów jest jednocześnie jedynym słusznym rozwiązaniem, w przeciwnym wypadku sytuacja ta kończyłaby się jałowym rozważaniem – co było pierwsze: jajko czy kura? Spłacanie kredytu A nie może zależeć od spłacania kredytu B i jednocześnie odwrotnie – kredytu B od kredytu A. Wszystko od wszystkiego zależeć nie może.

Pojawia się jeszcze inny problem natury czysto algorytmicznej. Przyjmijmy, że klient miał dwa kredyty: pierwszy, a po jego spłaceniu – drugi. Przypuśćmy jednak, że testowany proces akceptacji kredytowej dla historii tego klienta odrzuci pierwszy z jego wniosków kredytowych, gdyż pojawiło się zbyt duże prawdopodobieństwo niespłacenia. Bank zatem nie ma informacji o historii pierwszego kre-

dytu 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 banku, to pójdzie do innego, jeśli tam także jego wniosek zostanie odrzucony, to pójdzie do parabanku, a jeśli i tam mu się nie uda, to pożyczy od znajomych lub rodziny. Można tu podążać za myślą klasyków ekonomii, że klient konsumuje niezależnie od swojego wynagrodzenia. Jego potrzeby konsumpcyjne, a zatem także kredytowe, są wynikiem czegoś więcej, co jest związane z 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 sytuacja finansowa wcale ich do tego nie zmusza. Mogliby zakupić dany towar bez wiązania się z bankiem. Kredyt gotówkowy jest wybierany przez pewien podzbiór klientów korzystających z kredytów ratalnych, jest czasem dla nich koniecznością i jego spłacalność jest zatem bardziej wrażliwa na sytuację finansową klienta.
- Ryzyko kredytów ratalnych jest znacząco mniejsze od ryzyka kredytów gotówkowych.
- Spłacalność kredytów gotówkowych zależy od historii obu rodzajów kredytów: ratalnego i gotówkowego.

- Jeśli klient ma wiele aktywnych kredytów, to najgorzej będzie spłacał kredyt zaciągnięty ostatnio. 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 klient w tym czasie ma rachunki aktywne, czyli niezamknięte. Związane jest to z procesem sprzedaży krzyżowej (ang. *cross-sell*), gdy 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 ma 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: spłacenie lub niespłacenie 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 nie spłaci siedmiu rat (180 dni opóźnienia), to rachunek kredytowy jest zamykany ze statusem B (ang. *bad status*), wszystkie dalsze etapy windykacyjne są pomijane.

- 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.

3. Model biznesowy w obszarze kredytowania przez instytucje finansowe

3.1. Opłacalność procesu akceptacji kredytowej. Podstawowe składniki zysku: prowizja, przychody odsetkowe i strata kredytowa

Generator danych losowych Consumer Finance w uproszczonej wersji został obszernie opisany w publikacji z 2013 roku (Przanowski, 2013). Jest to model tylko jednego produktu – kredytu ratalnego. Każdy klient posiada jeden kredyt. Zmienne są zatem budowane tylko na podstawie aktualnej historii kredytowej jednego kredytu.

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.

Podrozdział ten jest zmodyfikowaną wersją publikacji *Rola danych symulacyjnych w badaniach Credit Scoring* (Przanowski, 2014b) oraz uproszczoną rozdziału „Opłacalność procesu, wpływ mocy predykcyjnej na zysk" książki *Credit Scoring w erze Big-Data* (Przanowski, 2014a). W niniejszej pracy jest on potrzebny do zrozumienia kolejnego podrozdziału (3.2).

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 – żeby nie ujawnić w ten prosty sposób dość krytycznych dla funkcjonowania banku informacji. Wygodne jest zatem wykorzystanie danych losowych, gdyż w tym wypadku nie obowiązuje nas tajemnica. Jednocześnie nie jest istotne wykazanie przydatności bardzo szczegółowo z dokładnością 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 przygotowane, aby uzyskać wartość ryzyka populacji równą 47%. Zostały także zbudowane karty skoringowe z różną mocą predykcyjną, wyrażoną w statystyce Giniego (Siddiqi, 2005). Statystyka ta jest jedną z najbardziej popularnych miar badania mocy predykcyjnej modeli, czyli tego, jak trafnie model skoringowy potrafi odróżniać klientów spłacających kredyty od mających opóźnienia lub inaczej – na ile dokładnie potrafi przewidzieć zachowanie klienta (to, czy spłaci kredyt). Przyjmuje ona wartości z przedziału od 0% do 100% (więcej o statystyce można znaleźć w podrozdziale 3.3). Typowe modele skoringowe oparte na danych z wniosków aplikacyjnych osiągają wartości maksymalnie do 60%, natomiast modele behawioralne, oparte na danych długiej historii klienta w banku, potrafią osiągać nawet wartość 80%. Z reguły wartości powyżej 80% oznaczają przetrenowane modele lub źle przygotowane dane z informacjami wziętymi z przyszłości. Wartość 0% reprezentuje model losowy.

Dość ważnym parametrem całego procesu jest już wspomniane ryzyko populacji. Pojawia się tu ciekawa sprzeczność pomiędzy celem analitycznym i celem finansowym. Istotą poprawnego zarządzania procesem akceptacji jest przecież maksymalizacja zysku ze sprzedaży kredytów, co sprowadza się do umiejętnego manipulowania parametrami procesu, by z jednej strony straty nie były zbyt duże, a z drugiej by sprzedaż była na tyle duża, żeby przychody z poprawnie spłacanych kredytów pokryły – i to z nawiązką – straty powstałe z powodu kredytów niespłacanych. Czym zatem jest ryzyko populacji? W kontekście potrzeb operacyjnych jest ono nam zupełnie niepotrzebne. Można w dość prosty sposób zarządzać ryzykiem, obserwując jedynie ryzyko akceptowanych kredytów, i dzień po dniu starać się, by ryzyko to systematycznie spadało lub utrzymywało się na ustalonym poziomie. Wiele banków tak właśnie postrzega rolę swoich departamentów ryzyka. Pojawia się pierwsze zasadnicze pytanie: jaki poziom ryzyka jest oczekiwany? Z reguły dajemy prostą odpowiedź – taki, przy którym cały nasz proces jest opłacalny, czyli zysk jest dodatni. Łatwo jest uznać *status quo* – skoro przy obecnym poziomie ryzyka nasz bank miał zyski, to najlepiej utrzymać ten stan tak długo, jak się da. Być może jednak powinniśmy mieć większy apetyt na ryzyko? Być może, akceptując kilka procent wię-

cej, np. 3%, zwiększymy stratę o 5%, ale także rozpędzimy biznes sprzedażowy i w efekcie przychody wzrosną o 7%. Może się też okazać, że przy próbach akceptowania większej liczby wnioskujących ryzyko nie będzie wzrastało aż tak gwałtownie. Problem jest bardzo poważny, gdyż z finansowego punktu widzenia nie powinniśmy akceptować wniosków, które przerodzą się w niespłacane kredyty. Z analitycznego punktu widzenia, który jest kluczowy w optymalizacji procesów (Provost i Fawcett, 2014; Ostasiewicz, 2012), trzeba na początku stracić, by zyskać poprawne pomiary, a potem zarobić istotnie więcej od konkurencji. Jeśli chcemy skutecznie zarządzać procesem akceptacji i maksymalizować zyski, to musimy umieć wyznaczyć, najlepiej jak potrafimy, ryzyko populacji, czyli ryzyko scenariusza pełnej akceptacji. Oczywiście, w praktyce można to doprecyzować. Jeśli mamy stuprocentową pewność, że akceptacja pewnej grupy wniosków przyniesie tylko stratę, to nie trzeba na niej testować wartości ryzyka. Ta sytuacja może mieć sens np. dla klientów obecnych w międzybankowych rejestrach informacji o negatywnych zdarzeniach w spłacaniu kredytów w innych bankach. Trzeba jednak pamiętać o tym, że nawet pozornie jednoznaczne przypadki do odrzucania mogą dać szansę na zarobek. Skoro dziś firmy windykacyjne potrafią sprzedawać pożyczki gotówkowe, to znaczy, że najważniejszą tajemnicą (ang. *know how*) jest posiadanie poprawnie zmierzonego ryzyka populacji. Rozumiemy je trochę szerzej, nie tylko jako prawdopodobieństwo zdarzenia *default*, ale także jako procent odzyskanego zadłużenia. Być może klient łatwo wpada w zadłużenia, ale zawsze udaje się odzyskać dług w procesie windykacyjnym. Czy w takim przypadku powinniśmy odrzucać takich klientów? Oczywiście że nie, na nich też możemy zarobić. Możemy jedynie zastanawiać się nad aspektem moralnym, gdyż udzielanie kredytu klientowi, w którego przypadku spodziewamy się twardych procesów windykacyjnych, jest świadomym narażaniem go na poważne problemy. Jest to temat, który może rozważać Komisja Nadzoru Finansowego (KNF). Jednocześnie jednak nie jest moralne zaciągnięcie kredytu i nie spłacanie go. Klientów, którzy biorą kredyty „bez opamiętania", też powinno się pilnować i oni także powinni mieć świadomość odpowiedzialności. Wracając do głównej myśli związanej z ryzykiem populacji: mamy tu sytuację, gdy poprawne zarządzanie procesem

zmusza nas do estymowania wielkości ukrytej, której nie da się tak łatwo zaobserwować i której pomiar jest bardzo kosztowny. Tak naprawdę nie tylko pomiar ryzyka populacji jest potrzebny, musimy znać krzywe przychodów i strat przy różnych poziomach akceptacji. Zadanie to oczywiście jest bardzo trudne i ogólnie niewykonalne. Praktycznie jednak możliwe do przybliżania przez ciągle ponawiane testy. Z tego punktu widzenia zarządzanie ryzykiem jest związane ze świadomością istnienia niedoskonałości i presji lepszego mierzenia, by sprawdzić kolejny scenariusz. Ryzykiem zarządza się tylko, będąc w nieustannym ruchu, podejmując coraz to kolejne wyzwania. Dyrektor ryzyka nigdy nie może spocząć na laurach, nie może wypowiedzieć zdania: „Zrobiłem już wszystko".

Niestety bez posiadania informacji o rzeczywistych i szczegółowych kosztach prowadzenia przedsiębiorstwa nie da się przedstawić całego arkusza zysków i strat (ang. P&L). Ale wystarczającą informacją jest policzenie oczekiwanej straty, prowizji i przychodów odsetkowych. Wszystkie inne koszty będą tylko odejmowane od zysku, nie wpłyną 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 ponoszony przez bank przy udzielaniu kredytu), p – prowizja za udzielenie kredytu, płatna przy uruchomieniu kredytu, $x^l_{Amount}(i) = A_i$ – kwota kredytu, $x^l_{N_{inst}}(i) = N_i$ – 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 ze 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. W bardziej realistycznym podejściu za EAD powinno się przyjąć nieco mniejszą kwotę od pierwotnie zaciąganej, wielkość tę oblicza się na podstawie danych historycznych. Najczęściej sprowadza się ona do współczynnika, np. $\mathrm{EAD} = 80\%A_i$. Wtedy wartość LGD powinno się estymować dokładniej i może wynosić około 62%. Sumarycznie, mnożąc wszelkie człony podczas obliczania straty oczekiwanej, otrzymuje się te same

wyniki. Dlatego też najprostszym przybliżeniem jest wprowadzenie tylko jednego współczynnika dla wartości LGD, równego 50%.

W przypadku danych historycznych nie musimy bazować na prognozie, ale możemy przyjąć, że PD jest oparte na wartości zmiennej $default_{12}$ (dodatkowa liczba 12 oznacza, że chodzi o zajście zdarzenia *default* w ciągu 12 miesięcy od zaciągnięcia kredytu). Jeśli nastąpiło zdarzenie *default*, to $\mathrm{PD} = 100\%$, a jeśli nie, to $\mathrm{PD} = 0\%$, wtedy wyznacza się obserwowaną stratę L. Opieramy się tu na założeniu, że jeśli klient spłacający 36 miesięcy swoje zobowiązania kredytowe w ciągu pierwszego roku nie wpadł w zadłużenie większe niż 90-dniowe, to nie wpadnie już w nie do końca trwania kredytu i spłaci go terminowo. Nie jest to do końca poprawne. Być może dla niektórych portfeli należy okres obserwacji (patrz rysunek 1, str. 35) wydłużyć do 24 miesięcy lub do samego końca trwania kredytu. Można też wyznaczyć krzywą koncentracji zdarzeń *default* ze względu na okres obserwacji, wtedy możliwe jest wyznaczenie minimalnego okresu obserwacji, uwzględniającego np. 90% wszystkich zdarzeń *default*. Z formalnego punktu widzenia w kalkulacji pełnej straty powinno się uwzględniać cały okres obserwacji, czyli czas kredytu do jego spłacenia lub umorzenia. Jest to dość istotna różnica w stosunku do metody Basel II, w której rekomenduje się sposoby liczenia rezerw czy wymogów kapitałowych dla rocznego okresu rozliczeniowego. Tymczasem w naszym przypadku chodzi o całkowitą stratę, która powstaje w ciągu wielu lat.

Kwotę I przychodów odsetkowych łącznie z prowizją oblicza się na podstawie procentu składanego. Mamy zatem dla każdego i-tego kredytu:

$$L_i = \begin{cases} 50\% A_i, & \text{gdy nastąpiło zdarzenie } default_{12}, \\ 0, & \text{gdy nie nastąpiło zdarzenie } default_{12}; \end{cases}$$

$$I_i = \begin{cases} A_i p, & \text{gdy nastąpiło } default_{12}, \\ A_i (N_i r \frac{(1+r)^{N_i}}{(1+r)^{N_i}-1} + (p-1)), & \text{gdy nie nastąpiło } default_{12}. \end{cases}$$

Sumaryczny zysk (profit) P całego portfela obliczamy zatem zgodnie z następującym wzorem:

$$P = \sum_i I_i - L_i. \tag{3.1}$$

Dokładniejsze kalkulacje rentowności są związane z modelami kalkulacji skorygowanej o ryzyko RAPM (ang. *risk-adjusted performance measure*) i RAROC (ang. *risk-adjusted return on capital*), opisanymi w *The Essentials of Risk Management* (Crouhy et al., 2006). Polegają one na osiąganiu maksymalnej stopy zwrotu z kapitału własnego skorygowanej o ryzyko. Uwzględnia się tu znacznie więcej składowych kosztów niż w naszych modelach, nie tylko stratę oczekiwaną, ale także kapitał regulacyjny i kapitał ekonomiczny. Brak tak pełnego podejścia nie zmieni jednak znacząco naszych wyników i analiz, gdyż ich celem jest przede wszystkim wykazanie istotnego wpływu przyrostu mocy predykcyjnej modeli skoringowych na przyrost zysku banku.

Dla każdego modelu skoringowego z różnymi mocami predykcyjnymi możemy posortować wszystkie wnioski według wartości oceny punktowej od najmniej do najbardziej ryzykownego. Ustalając punkt odcięcia, wyznaczamy sumaryczną wartość zysku na zaakceptowanej części portfela i procent akceptacji. Otrzymujemy w ten sposób krzywe profit, prezentowane na rysunku 2. Niektóre z nich, np. dla Giniego z wartością 20%, nigdy nie przyniosą zysku bankowi, niezależnie od procentu akceptacji zawsze tracimy zainwestowane fundusze. 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 wynik banku jest ujemny i wynosi około $-44{,}5$ mln PLN. Najlepsze trzy krzywe pokazano dokładniej na rysunku 3. Modele z mocą większą od około 60% potrafią zidentyfikować opłacalne segmenty, przy czym im lepszy jest model, tym więcej możemy zaakceptować i więcej zarobić. W przypadku modelu o mocy 89%, dość dużej, by pojawiła się w praktyce, można zaakceptować prawie 44% wniosków i zyskać $10{,}5$ mln PLN. Dla tego modelu na rysunku 4 pokazano dodatkowo składowe zysku, czyli przychód i stratę, narastająco. Przy pełnej akceptacji strata sięga aż $72{,}2$ mln PLN. 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 straty (patrz

rysunek 5). Im większa jest moc modelu, tym krzywa straty jest bardziej zakrzywiona i tym dłuższy odcinek od zerowej akceptacji jest spłaszczony, a tym samym związany z małą stratą. Stopień zakrzywienia krzywej straty jest bardzo prostą interpretacją statystyki Giniego. W przypadku modelu o zerowej mocy strata będzie narastać liniowo. W przypadku mocy 100% będzie to łamana: do wartości dopełnienia ryzyka globalnego $1 - 47\% = 53\%$ będzie linią zerową, a potem gwałtownie liniowo będzie rosnąć do całkowitej straty $72{,}2$ mln PLN.

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 2 przedstawiono zebrane wskaźniki. Wystarczy zwiększyć moc modelu o 5%, a miesięcznie bank zarobi o prawie $1{,}5$ mln PLN więcej, zwiększając przy tym procent akceptacji o $3{,}5\%$. Można też, pozostawiając procent akceptacji (ang. *acceptance rate* – AR) na poziomie 20%, polepszać model i oszczędzać na stracie. W tym wypadku zaoszczędzimy miesięcznie prawie 900 tys. PLN. W przypadku akceptacji 40% oszczędność wyniesie aż $1{,}5$ mln 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.

3.2. Uproszczona symulacja w arkuszu kalkulacyjnym – przypadek kredytu ratalnego

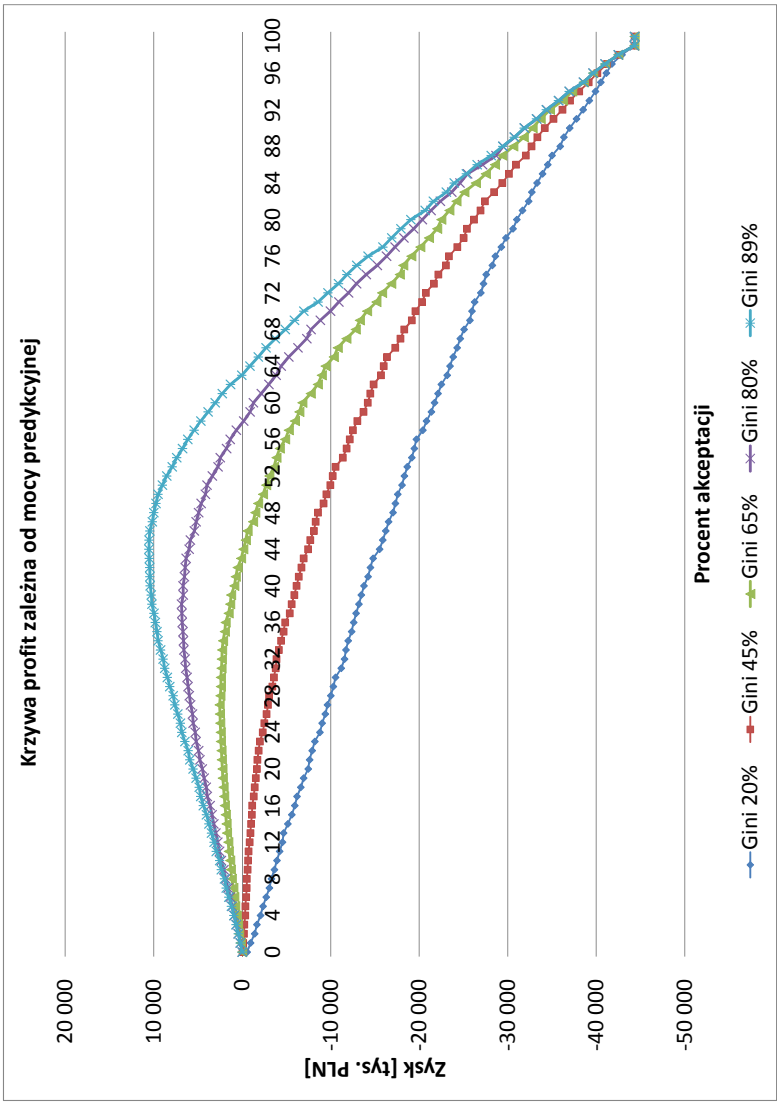
Symulacji procesu akceptacji jednego produktu można dokonać, używając bardzo uproszczonego modelu w arkuszu kalkulacyjnym o nazwie *acceptance_process_simulation.xlsx*. Obliczenia nie będą bardzo dokładne, nie uwzględnią specyfiki rozkładów parametrów ani w szczególności historii wniosków kredytowych z różnymi kwotami i charakterystykami klienta. Pomimo prostej konstrukcji można

Tabela 2. 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 tys. PLN
Zmiana straty oczekiwanej (AR = 20%)	872 tys. PLN
Zmiana straty oczekiwanej (AR = 40%)	1 529 tys. PLN

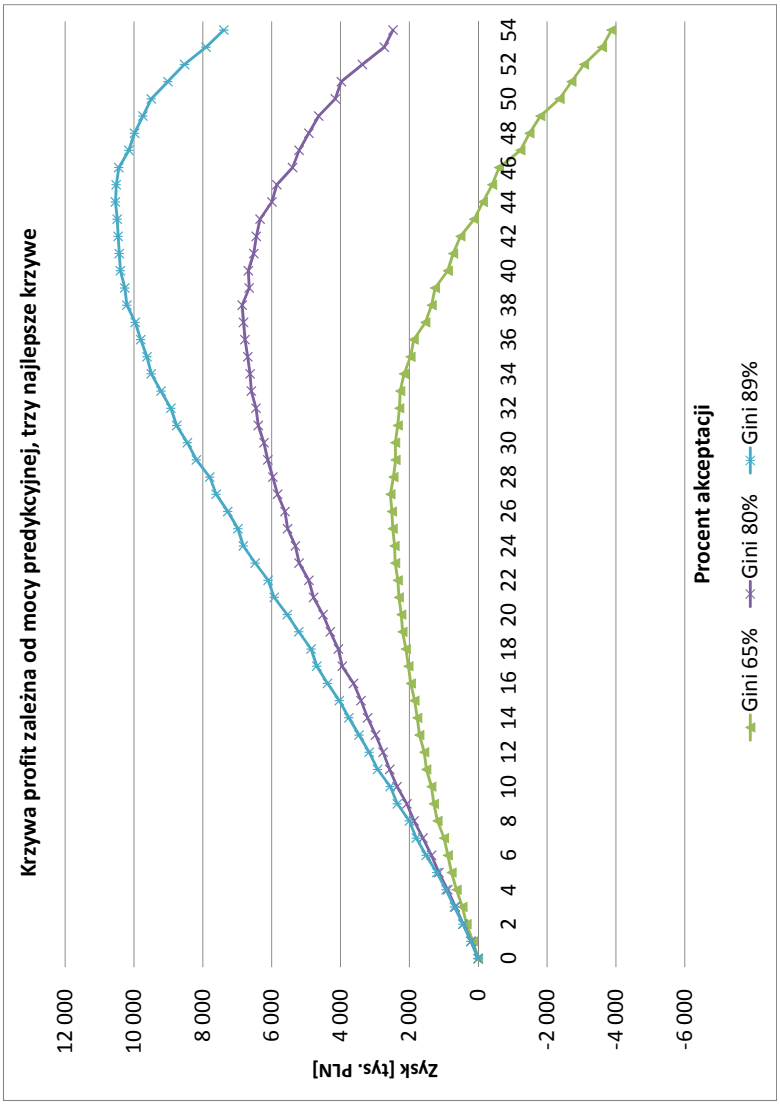
Źródło: Przanowski (2014b).

Rysunek 2. Krzywe profit



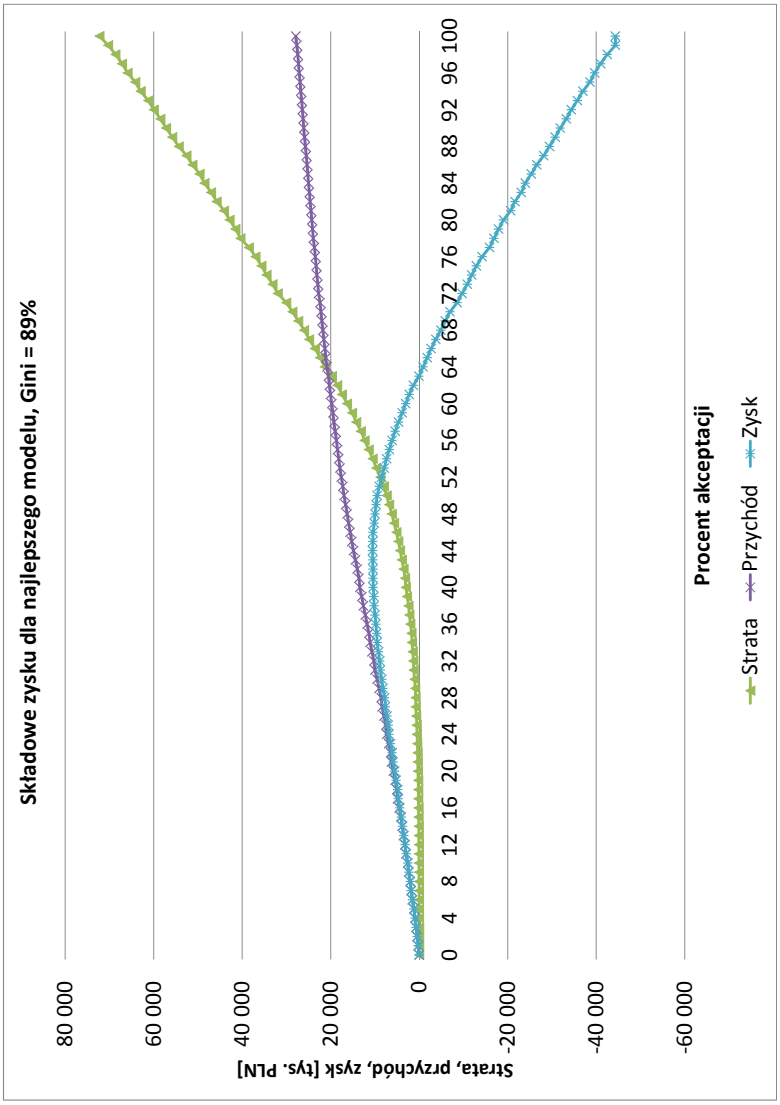
Źródło: Przanowski (2014b).

Rysunek 3. Najlepsze krzywe profit



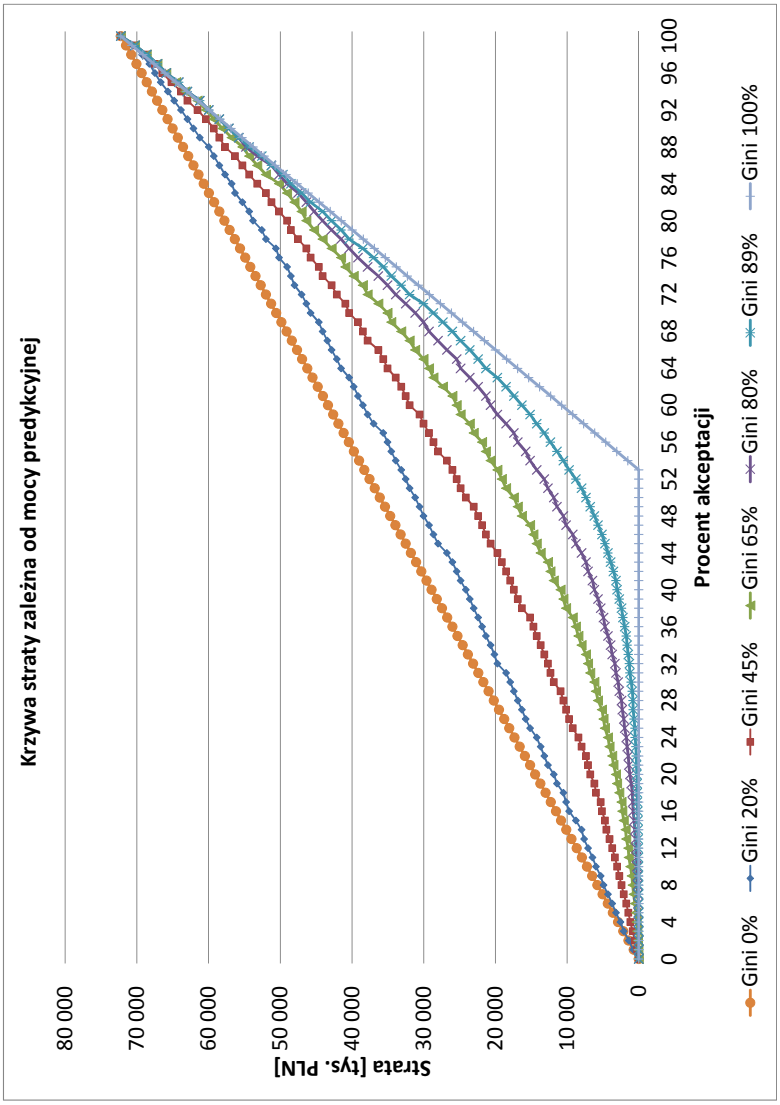
Źródło: Przanowski (2014b).

Rysunek 4. Składowe zysku dla najlepszego modelu



Źródło: opracowanie własne.

Rysunek 5. Krzywe straty



Źródło: opracowanie własne.

w ten sposób przybliżyć wartości podstawowych składników finansowych oraz przeprowadzić wiele studiów przypadków, starając się tak dobrać wszystkie parametry, by finalnie uzyskać pożądane wartości zysku.

Analizy takiego arkusza stają się bardzo przydatne dla właścicieli nowych firm (ang. *startup*) czy linii biznesowych, aby mogli sprawdzić poprawność swojego nowego procesu, zanim rozpoczną inwestycję. Badanie zależności pomiędzy parametrami jest szczególnie przydatne w wyznaczaniu minimalnej mocy predykcyjnej modelu skoringowego, by nadal utrzymać istotny zysk. Przy niektórych parametrach, np. zbyt dużym ryzyku populacji, może się okazać, że opłacalność procesu staje się możliwa dopiero przy wartości Giniego 90%, co niestety nie jest możliwe w rzeczywistości. Taka analiza stanowi bardzo dobrą naukę i pozwala uniknąć wielu późniejszych rozczarowań.

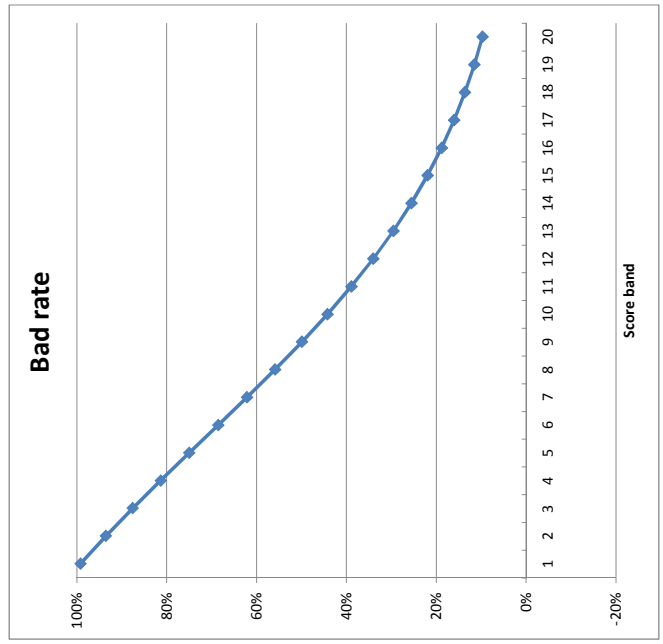
Na rysunku 6 zostały przedstawione podstawowe parametry wprowadzane do arkusza. Najważniejszymi z nich (w kolorze żółtym) są:

- miesięczna liczba wszystkich wniosków (w arkuszu – *number of applications per month*);
- kwota kredytu (*average loan amount*);
- liczba rat (*average number of installments*);
- oprocentowanie (*annual percentage rate – or net margin*);
- współczynnik straty LGD (*LGD – Loss Given Default*);
- prowizja (*provision charged on disbursement day*);
- ryzyko populacji (*global risk in market – default12*).

W arkuszu wszystkich opisów dokonano w języku angielskim, aby można było używać symulacji podczas różnych prezentacji, także w środowisku międzynarodowym.

Zbiór wniosków kredytowych jest dzielony na 20 równolicznych grup – *score band* (patrz tabela 3). Istotą każdego modelu skoringowego jest uporządkowanie wszystkich wniosków w kolejności prawdopodobieństwa zajścia zdarzenia *default*. W arkuszu sprowadza się

Rysunek 6. Fragment arkusza kalkulacyjnego. Podstawowe parametry kredytu ratalnego



Number of applications per month	50 000
Average loan amount	5 000
Average number of installments	36
Annual percentage rate (or net margin)	12%
LGD (Loss Given Default)	50%
Provision charged on disbursement day	6%

Gini global	65,54%
Gini on accepted	24,50%

Global risk in market (default12)	47%
Accepted risk	7,40%
Acceptance rate	40,00%

Global loss	58 750 000
Global income	33 882 258
Global profit	-24 867 742

Accepted loss	9 246 726
Accepted income	20 842 459
Accepted profit	11 595 733

Charge provision	300,00
Income from interest rates	978,58

Źródło: opracowanie własne.

Tabela 3. Fragment arkusza kalkulacyjnego. Grupy skoringowe (*score band*) w przypadku kredytu ratalnego

Score bands	Number of applications	Observed default12 ratio (bad rate)	Accepted flag
1	2 500	99,34%	0
2	2 500	93,69%	0
3	2 500	87,72%	0
4	2 500	81,50%	0
5	2 500	75,12%	0
6	2 500	68,69%	0
7	2 500	62,30%	0
8	2 500	56,06%	0
9	2 500	50,05%	0
10	2 500	44,36%	0
11	2 500	39,05%	0
12	2 500	34,15%	0
13	2 500	29,69%	1
14	2 500	25,68%	1
15	2 500	22,10%	1
16	2 500	18,94%	1
17	2 500	16,18%	1
18	2 500	13,77%	1
19	2 500	11,69%	1
20	2 500	9,89%	1
	50000	47,00%	

a	-0,18
b	1

Źródło: opracowanie własne.

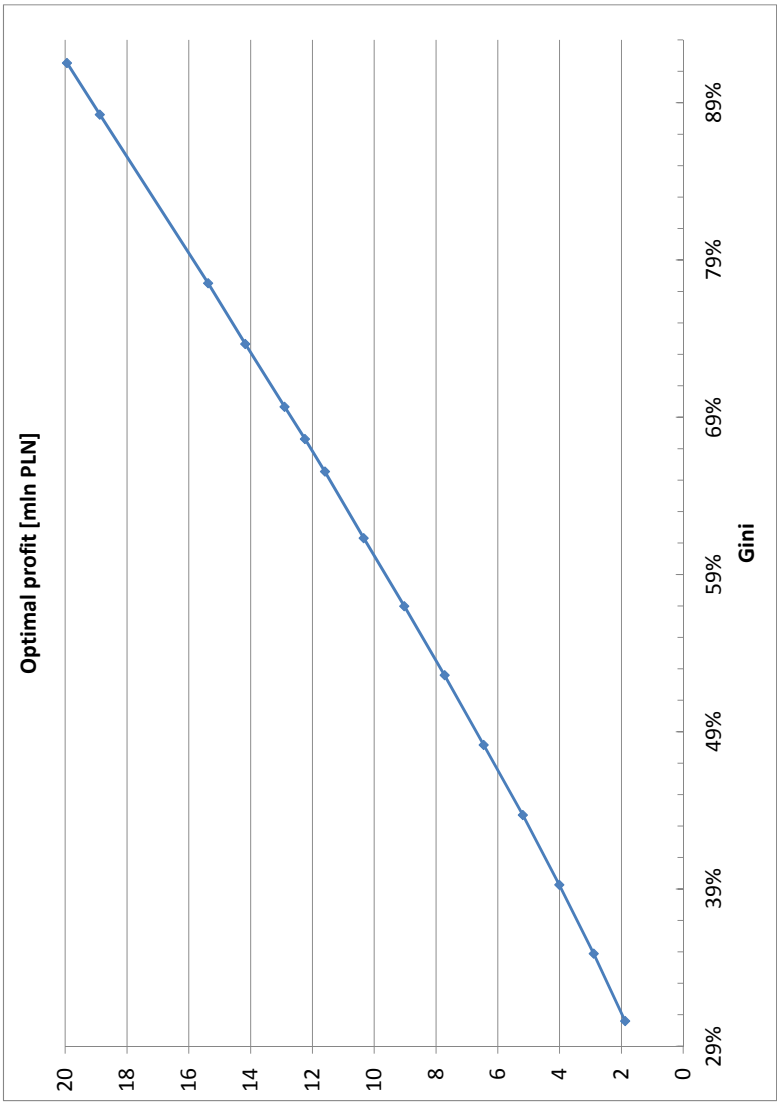
Tabela 4. Fragment arkusza kalkulacyjnego. Lista parametrów w zależności od zmieniającej się mocy predykcyjnej modelu w przypadku kredytu ratalnego

Accepted loss	b	a	Gini	Optimal Profit	Optimal Acc rate
9 246 726	1	-0,18	65,54%	11 595 733	40,00%
7 953 661	1	-0,25	91,53%	19 942 990	50,00%
8 657 135	1	-0,24	88,25%	18 879 738	50,00%
8 862 058	1	-0,21	77,52%	15 373 572	45,00%
9 657 328	1	-0,2	73,66%	14 171 577	45,00%
10 497 918	1	-0,19	69,66%	12 901 084	45,00%
10 934 456	1	-0,185	67,62%	12 241 286	45,00%
9 246 726	1	-0,18	65,54%	11 595 733	40,00%
10 077 636	1	-0,17	61,31%	10 339 871	40,00%
10 947 157	1	-0,16	56,99%	9 025 651	40,00%
9 695 338	1	-0,15	52,59%	7 721 248	35,00%
10 532 486	1	-0,14	48,15%	6 455 957	35,00%
9 256 670	1	-0,13	43,70%	5 187 825	30,00%
10 035 560	1	-0,12	39,26%	4 010 588	30,00%
8 657 903	1	-0,11	34,88%	2 896 381	25,00%
7 215 725	1	-0,1	30,60%	1 879 693	20,00%

1% of Gini	296 451	PLN
5% of Gini	1 482 254	PLN
10% of Gini	2 964 507	PLN

Źródło: opracowanie własne.

Rysunek 7. Fragment arkusza kalkulacyjnego. Wykres obrazujący współzależność pomiędzy optymalnym zyskiem i mocą predykcyjną modelu w przypadku kredytu ratalnego



Źródło: opracowanie własne.

to do wyznaczenia statystyki *bad rate* udziału złych klientów w każdej z grup *score band*. Im lepszy jest model, tym istotniejsze jest zróżnicowanie statystyki *bad rate* w grupach. Można to zaobserwować na rysunku 6, na którym jest przedstawiony kształt linii *bad rate* w zależności od grupy *score band*. W przypadku pierwszej grupy *bad rate* jest największy, dla dwudziestej najmniejszy. Kształt tej krzywej jest sterowany funkcją odwrotności do logitu określoną wzorem:

$$bad\ rate = \frac{1}{1 + \exp(-(a \cdot score\ band + b))},$$

gdzie parametry a i b są dobierane ręcznie. Na podstawie ich wartości jest liczona statystyka Giniego, dzięki której można tak dobrać parametry, by uzyskiwać obserwowane w rzeczywistości wyniki.

W zależności od ryzyka (*bad rate*) w każdej z grup automatycznie jest dobierany poziom akceptacji tak, aby uzyskać największą wartość zysku. Metody liczenia zysku zostały już szczegółowo omówione w podrozdziale 3.1. Możemy zaobserwować, że statystyka Giniego liczona na całym zbiorze wniosków różni się od statystyki policzonej na podzbiorze tylko akceptowanych wniosków. W rzeczywistości tylko ta druga jest obserwowana, co powoduje, że czasem mylnie można to interpretować, że model skoringowy nie ma wystarczającej mocy dyskryminacyjnej.

W podrozdziale 3.1 przedstawiono także rozumowanie doprowadzające do bardzo ciekawego wniosku, że poprawa mocy predykcyjnej modelu powoduje istotny przyrost zysku. W przypadku uproszczonej symulacji w arkuszu można dość prosto udowodnić postawioną tezę. Mianowicie, wykonując wielokrotnie obliczenia, wstawiając ręcznie różne wartości parametru a (komórka D47), uzyskujemy listę zmieniających się wartości: statystyki Giniego, optymalnego procentu akceptacji i optymalnego zysku ukazanego w formie zielonej linii arkusza, przedstawionej w tabeli 4. Każda taka linia jest następnie wstawiana oddzielnie i tworzy historię scenariuszy. Po wykonaniu wielu takich obliczeń można przekonać się, że zależność pomiędzy optymalnym zyskiem a mocą predykcyjną jest prawie idealnie liniowa (patrz rysunek 7) i w związku z tym można policzyć wielkości przyrostów. Mamy następujące wnioski: przyrost Giniego

o 1% zwiększa zysk o 296 tys. PLN. Przyrost o 5% Giniego zwiększa zysk aż o 1482 tys. PLN, co jest bardzo bliskie liczbie prezentowanej w tabeli 2. Pomimo przybliżenia i prostych obliczeń uzyskano bardzo podobne liczby otrzymane dzięki zaawansowanemu oprogramowaniu SAS. Nie oznacza to bynajmniej, że wszystko daje się obliczyć w arkuszu kalkulacyjnym. Prawdziwych symulacji i zmiany strategii procesu akceptacji zawsze dokonuje się na podstawie danych historycznych, które procesuje się jeszcze raz w systemie decyzyjnym, co z reguły jest zbyt skomplikowane dla arkusza kalkulacyjnego. Warto tu przedstawić podstawowe funkcje systemu decyzyjnego. Wiele firm popełnia dość przykre w skutkach błędy, gdy wdrażając nowy system decyzyjny, koncentruje się jedynie na pierwszej istotnej funkcji systemu, czyli procesowaniu wniosków w środowisku produkcyjnym. Jeśli uważa się, że jest to jedyna funkcja systemu decyzyjnego, to skazuje się bank na bardzo poważne straty finansowe związane z błędami operacyjnymi. Każdy doświadczony użytkownik systemu decyzyjnego przyzna, że nigdy nie udało się zdefiniować poprawnie działających procesów przy pierwszym wdrożeniu. Z reguły jest to proces ciągłego doskonalenia. System decyzyjny musi zatem umieć eksportować do baz danych lub logów wiele pośrednich wartości obliczanych zmiennych. Bez tego utrudnione staje się identyfikowanie błędów i poprawne wykonywanie analiz jakości procesu. Musi być druga funkcjonalność, czyli zrzucanie pośrednich wartości. Trzecią jest wykonywanie testów masowych, czyli procesowanie dużej liczby historycznych wniosków z pewnego przedziału czasowego przy zmienionej liście reguł decyzyjnych. Tego typu testy dają gwarancję poprawności nowo wprowadzanych reguł oraz podstawowych wskaźników na nich liczonych, takich jak procent akceptacji (ang. *approval rate*). Dobry system decyzyjny musi zatem posiadać trzy funkcje: procesować wnioski, zrzucać wartości pośrednie i wykonywać testy masowe.

Jest jeszcze czwarta funkcja systemu decyzyjnego, jak na razie przez wiele firm pomijana. Jest to tworzenie w sposób automatyczny dokumentacji z procesu decyzyjnego. Temat wydawać się może nieistotny, ale w praktyce z reguły zawsze brakuje czasu na tworzenie i uaktualnienie dokumentacji. Oznacza to, że w instytucji szybko reagującej na zmiany po prostu dokumentacji nie ma. To w ko-

lejnych krokach prowadzi do niezastępowalności pracowników oraz braku możliwości bezpiecznego prowadzenia biznesu, gdyż po pewnym czasie nikt w organizacji już nie zna się na procesie.

3.3. Statystyki mierzenia mocy predykcyjnej modeli

Podstawowym źródłem informacji, na bazie których przygotowano niniejszy podrozdział, jest artykuł *How to Measure the Quality of Credit Scoring Models* (Řezáč i Řezáč, 2011), w którym autorzy bardzo starannie opisali znane metody liczenia statystyk mocy predykcyjnej. Niestety pomimo znajomości pojęć istnieje odnośnie do tego tematu dość duża rozbieżność dotycząca proponowanych nazw liczonych statystyk i krzywych. Prawie w każdym podręczniku czy artykule znajdziemy różne opisy na osiach współrzędnych krzywych ROC, CAP i Lorenza. Naszym podstawowym celem będzie przedstawienie sposobu liczenia statystyki Giniego oraz próba wyjaśnienia jej interpretacji. Statystykę tę można obliczyć, posługując się różnymi wzorami, wszystkie one doprowadzają do tej samej liczby. Wyjaśnione własności i udowodnione wzory tożsamości można też znaleźć – poza wspomnianym artykułem – w wybranych publikacjach (Anderson, 2007; Krzyśko et al., 2008; Engelmann et al., 2003; BIS-WP14, 2005).

Wszystkich obliczeń dokonano w arkuszu kalkulacyjnym o nazwie *gini_curves.xlsx*. Podstawowymi parametrami są liczba wszystkich wierszy (klientów lub analizowanych przypadków, *number of cases*), równa 20 tys., oraz ryzyko populacji (*global bad rate*), równe 14%. Dodatkowo parametrami *a* i *b* steruje się kształtem krzywej *bad rate* na podstawie funkcji odwrotnej do logitowej, co z kolei decyduje finalnie o wartości statystyki Giniego (*Gini global*), w tym wypadku wynoszącej $77{,}6\%$ (patrz tabela 5).

Dla każdego *score bands* o numerach od $s = 1$ do $s = 20$ mamy wyliczone liczby dobrych i złych klientów (*goods* i *bads*). Możemy teraz wyznaczyć liczby skumulowanych dobrych i złych, a potem ich udziały (*cum goods%* i *cum bads%*), oznaczając odpowiednio CDF_s^G i CDF_s^B, postępując w kolejności od $s = 1$ do $s = 20$. Powstały zatem skumulowane udziały nazywane, dystrybuantami do-

Tabela 5. Fragment arkusza kalkulacyjnego. Podstawowe parametry i wyliczone wielkości do liczenia statystyk predykcyjności

Number of cases	20 000
Gini global	77,59%
Global bad rate	14,00%

CAP	
y	
%Bad captured	

Score bands	Number of cases		Observed bad rate	Bads	Goods	%Good captured Cumgoods%	Gains Cum bads%	Cumgoods+	Cumbads-	Z	
0						0,00%	0,00%				
1	1 000		61,86%	619	381	2,22%	22,09%	2,22%	22,09%	22,41%	0,49%
2	1 000		52,68%	527	473	4,97%	40,91%	7,19%	18,81%	23,10%	1,35%
3	1 000		43,13%	431	569	8,27%	56,31%	13,24%	15,40%	25,44%	2,04%
4	1 000		33,95%	340	660	12,11%	68,44%	20,39%	12,13%	27,91%	2,47%
5	1 000		25,77%	258	742	16,43%	77,64%	28,55%	9,20%	30,83%	2,63%
6	1 000		18,95%	190	810	21,14%	84,41%	37,57%	6,77%	32,97%	2,54%
7	1 000		13,59%	136	864	26,17%	89,26%	47,31%	4,85%	35,27%	2,30%
8	1 000		9,56%	96	904	31,42%	92,68%	57,59%	3,41%	37,24%	1,97%
9	1 000		6,62%	66	934	36,85%	95,04%	68,28%	2,37%	38,86%	1,62%
10	1 000		4,54%	45	955	42,40%	96,67%	79,26%	1,62%	40,15%	1,29%
11	1 000		3,09%	31	969	48,04%	97,77%	90,44%	1,11%	41,15%	1,00%
12	1 000		2,10%	21	979	53,73%	98,52%	101,77%	0,75%	41,91%	0,76%
13	1 000		1,42%	14	986	59,46%	99,03%	113,19%	0,51%	42,48%	0,57%
14	1 000		0,95%	10	990	65,22%	99,37%	124,68%	0,34%	42,90%	0,42%
15	1 000		0,64%	6	994	71,00%	99,60%	136,22%	0,23%	43,21%	0,31%
16	1 000		0,43%	4	996	76,78%	99,75%	147,78%	0,15%	43,44%	0,23%
17	1 000		0,29%	3	997	82,58%	99,85%	159,37%	0,10%	43,60%	0,16%
18	1 000		0,19%	2	998	88,38%	99,92%	170,97%	0,07%	43,72%	0,12%
19	1 000		0,13%	1	999	94,19%	99,97%	182,58%	0,05%	43,80%	0,08%
20	1 000		0,09%	1	999	100,00%	100,00%	194,19%	0,03%	43,86%	0,06%
	20000		14,00%	2 800	17 200	KS	0,63			Sum Z	22,41%
			14,00%	2 800						Gini	77,59%
		a		-0,4						Formal AUC	-77,59%
		b		1							

Źródło: opracowanie własne.

brych i złych (ang. *cumulative distribution function*). Na ich podstawie możemy teraz wyznaczyć krzywą Lorenza (patrz rysunek 12), gdzie osią poziomą jest CDF_s^B, oznaczane także jako *% bads captured*, czyli ile procent złych objęto, a pionową CDF_s^G, oznaczane jako *% goods captured*, czyli ile procent dobrych dosięgnięto. Możemy także dla ułatwienia kolejnych zapisów przyjąć, że dla $s = 0$ mamy: $\mathrm{CDF}_0^B = 0$ i $\mathrm{CDF}_0^G = 0$. Podwojone pole powierzchni pomiędzy krzywymi niebieską i czerwoną wyznacza statystykę Giniego. Można ją zapisać wzorem, licząc pole pod krzywą jako sumę pól kolejnych trapezów wzdłuż krzywych. Mamy zatem:

$$\mathrm{Gini} = 1 - \sum_{s=1}^{20} (\mathrm{CDF}_s^B - \mathrm{CDF}_{s-1}^B)(\mathrm{CDF}_s^G + \mathrm{CDF}_{s-1}^G).$$

W taki sposób właśnie jest liczona statystyka Giniego w każdym rozważanym w książce arkuszu kalkulacyjnym (patrz statystyki: *cumgoods+* i *cumbads–*). W tym wypadku jej wartość jest podana w komórce J30 i wynosi $77{,}59\%$.

W wielu instytucjach przyjęło się używać innej statystyki mocy predykcyjnej, nazywanej statystyką Kołmogorowa–Smirnowa i oznaczanej jako KS. Jest to maksymalna odległość pomiędzy wspomnianymi dystrybuantami. Wykresy tych dystrybuant tworzą dość charakterystyczny rysunek, zwany rybim okiem (ang. *fish eye*), patrz rysunek 13. Statystykę liczymy zatem wzorem:

$$\mathrm{KS} = \mathrm{MAX}_{s=1}^{20}(\mathrm{CDF}_s^B - \mathrm{CDF}_s^G).$$

W arkuszu statystyka ta jest policzona w komórce G29 i wynosi $0{,}63$.

Na podstawie dopełnień dystrybuant (inaczej odwrotnych), czyli statystyk (*TPrate* i *FPrate*):

$$\mathrm{ICDF}_s^B = 1 - \mathrm{CDF}_{s-1}^B,$$

$$\mathrm{ICDF}_s^G = 1 - \mathrm{CDF}_{s-1}^G,$$

dla których dodatkowo $\mathrm{ICDF}_{21}^G = 1$ i $\mathrm{ICDF}_{21}^B = 1$, możemy wyznaczyć krzywą ROC (ang. *Receiver Operating Characteristic*), patrz rysunek 10. Na osi poziomej znajduje się statystyka ICDF_s^B, oznaczana różnie: *false alarm rate*, *1-specificity* – specyficzność lub *% bads*

remain, czyli ile procent pozostałych złych. Oś pionową reprezentuje ICDF_s^G, oznaczana jako: *hit rate*, *sensitivity* – czułość lub *% goods remain*, czyli ile procent pozostałych dobrych. Pole pod tą krzywą jest nazywane AUC (ang. *area under curve*). Policzono je w komórce Y29 i jest ściśle związane z wartością Giniego wzorem:

$$\mathrm{Gini} = 2\mathrm{AUC} - 1.$$

Kolejną krzywą, którą często rozważa się przy modelach predykcyjnych, jest krzywa CAP (ang. *Cumulative Accuracy Profile*), patrz rysunek 9. Na osi poziomej jest przedstawiony udział w populacji, czyli w naszym wypadku może to być albo numer *score band*, albo statystyka procentu akceptacji wyznaczona wzorem:

$$acc\ rate_s = s \cdot 5\%.$$

Oś tę często oznacza się jako *depth*, czyli jak głęboko wybiera się podzbiór populacji. Na osi pionowej jest już znana statystyka CDF_s^B, oznaczana jako *sensitivity* (jest to inna czułość niż w krzywej ROC) lub *%bad captured*, czyli ile procent złych. Statystyka ta jest też oznaczana jako *gains*, szczególnie w modelach marketingowych, w których porównuje się nią mierniki kampanii reklamowych, gdyż oznacza ona udział respondentów w wyznaczanej grupie docelowej w stosunku do całej populacji respondentów. Dodatkowo na wykresie rysuje się specjalną krzywą, łamaną, reprezentującą idealny model, który potrafi w pierwszych percentylach wybierać wyłącznie złych klientów. Okazuje się, że statystykę Giniego można też wyznaczyć na bazie krzywych CAP i idealnego modelu. Mianowicie stosunek pola powierzchni wyznaczonej przez krzywe czerwoną i niebieską do pola wyznaczonego przez czerwoną i zieloną jest właśnie statystyką Giniego (patrz wartość komórki AE31).

Do omówienia pozostały jeszcze dwie krzywe, które są do siebie dość podobne – *lift* (dokładnie skumulowany *lift*) i *bad rate* (patrz odpowiednio rysunki 11 i 8). Pierwsza reprezentuje statystykę interpretowaną następująco: ile razy na danym skumulowanym percentylu populacji model wybiera złych klientów lepiej od modelu losowego. Liczymy ją wzorem:

$$lift_s = \frac{\mathrm{CDF}_s^B}{acc\ rate_s}.$$

Warto zauważyć, że tak wprowadzona definicja statystyki *lift* jest tak naprawdę stosunkiem dwóch dystrybuant – złych klientów do wszystkich klientów; albo jeszcze inaczej: stosunek dystrybuanty złych klientów związanej z modelem skoringowym do dystrybuanty złych klientów modelu losowego.

Statystyka *bad rate* jest typową miarą ryzyka, czyli udziałem złych klientów w danym *score band*. Krzywe *lift* i *bad rate* odgrywają istotną rolę w dobieraniu grupy docelowej kampanii lub punktu odcięcia. Dość często zdarza się, że budujemy kilka modeli predykcyjnych. Każdy z nich ma tę samą wartość statystyki Giniego. Nie oznacza to jednak, że punkt odcięcia będziemy mieli taki sam dla wszystkich modeli. Może się okazać, że tylko jeden ma na 5% percentylu największą wartość *lift* albo *bad rate*. Trzeba mieć świadomość tego, że pole pod krzywą może być takie samo dla rożnych modeli, ale kształty krzywych ROC, CAP i *lift* mogą być rożne. Najczęściej w praktyce w doborze modelu i punktu odcięcia stosuje się krzywe kwotowe, np. krzywe profit (patrz rysunek 2, str. 57). Bywa jednak, że stosuje się miary ilościowe, wtedy do wyboru punktu odcięcia zamiast Giniego używa się raczej *lift* lub *gains*, gdyż interesuje nas wybór modelu, który w danym percentylu wyselekcjonował najwięcej złych klientów. W przypadku wyliczania wymogów kapitałowych tam, gdzie chodzi o pokrycie całego portfela zróżnicowaną wartością *bad rate* (dokładnie estymacją PD), możemy się oprzeć na statystyce Giniego. Jeśli jednak będziemy używać zawsze podobnych technik budowy modeli, np. zawsze regresję logistyczną, to kształt krzywych będzie zagwarantowany przez tę technikę i statystykę Giniego da się wtedy wykorzystywać do obu typów omówionych zastosowań.

Jednym z najlepszych sposobów interpretacji wartości statystyki Giniego jest odwołanie się do wskaźnika D Somersa dla przypadku zmiennej binarnej. Rozważmy dwa różne wiersze – lub inaczej – dwóch klientów, z których pierwszy jest oznaczony jako zły, a drugi jako dobry. Nasz model predykcyjny każdemu klientowi przyporządkowuje wartość teoretyczną PD (ang. *probability of default*) w arkuszu oznaczaną jako *bad rate*, jest to oczekiwana wartość ryzyka, jakiego możemy się spodziewać w przypadku danego klienta (formalnie jest to wartość oczekiwana, a w arkuszu obserwowana).

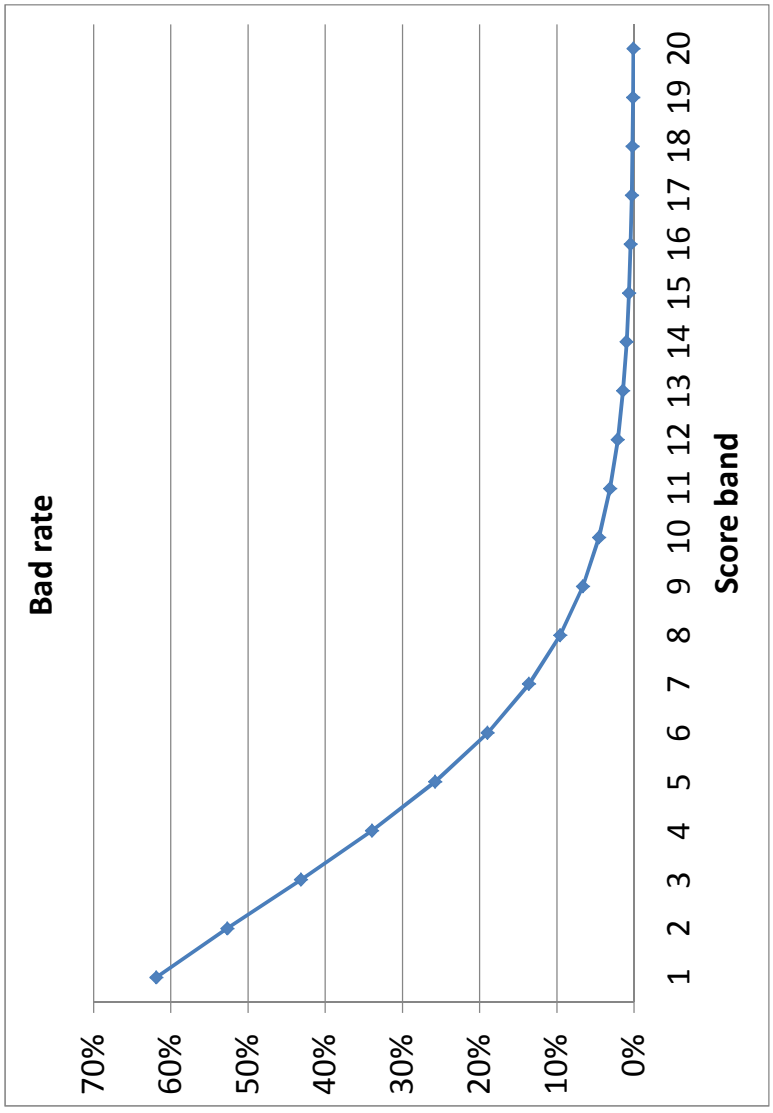
Ponieważ pierwszy klient jest zły, a drugi dobry, to model ten powinien zwracać *bad rate* dla pierwszego większy niż dla drugiego. Jest to naturalne i logiczne oczekiwanie właściwości modelu. Udział takich przypadków wśród wszystkich możliwych par klientów (zły, dobry) nazywamy „procentem zgodnych". Odwrotna sytuacja oznacza „procent niezgodnych", a równość *bad rate* oznacza „procent równych". Okazuje się, że statystykę Giniego wyraża się w różnicy procentu zgodnych i niezgodnych (patrz komórka AD39). Właśnie ten wzór pomaga przedstawić najlepszą interpretację statystyki Giniego. Na początku rozważmy sytuację, gdy procent zgodnych jest równy procentowi niezgodnych, czyli każdy jest równy 50%. Bierzemy dowolną parę (zły, dobry), mamy wtedy 50% szansy, że ich statystyki *bad rate* będą zgodne albo niezgodne. Oznacza to, że model nic nie rozróżnia, czyli Gini jest równy zeru. Jeśli założymy, że procent równych jest zawsze zerowy, to w łatwy sposób możemy interpretować naszą statystykę Giniego. Jeśli np. wynosi ona 60%, to szansa na trafienie zgodnych wynosi: $\frac{100\%+60\%}{2} = 80\%$. Innymi słowy, szansa, że nasz model ustawi naszych klientów we właściwej kolejności, wynosi 80%. Jest to także wartość statystyki AUC, co oznacza, że interpretacja AUC jest jeszcze prostsza. Tak naprawdę łatwiejsza i najpoprawniejsza jest interpretacja procentu zgodnych.

Przy okazji statystyk Giniego, *lift*, *gains* oraz krzywych ROC, CAP i Lorenza wprowadza się także mierniki oparte na macierzy klasyfikacji (ang. *confusion matrix*). Pojawiają się tu takie pojęcia z języka angielskiego, jak *true positive* (TP), *true negative* (TN), *false positive* (FP) czy *false negative* (FN) i wiele pochodnych, które na nich bazują. W naszym przypadku okazały się niepotrzebne, gdyż najważniejsze było zrozumienie statystyki Giniego.

3.4. Optymalizacja procesu windykacji polubownej

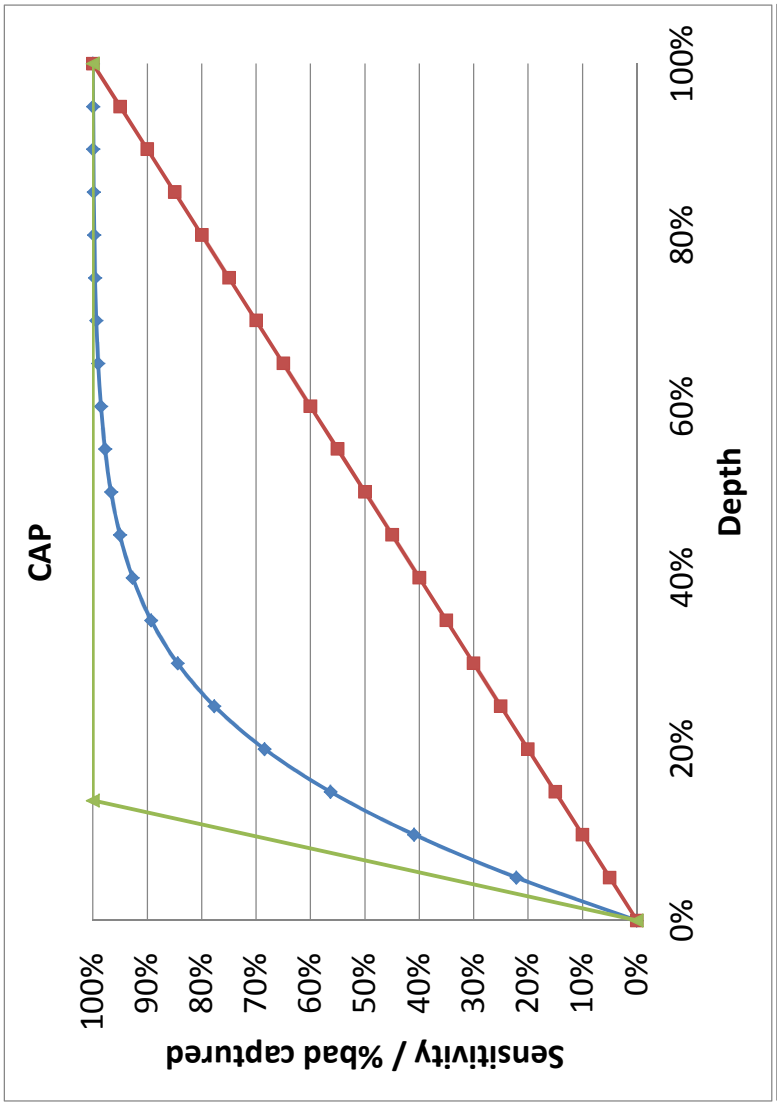
Bardzo ciekawym i znacznie trudniejszym przypadkiem jest zarządzanie procesem windykacji polubownej (ang. *amicable*). Jest to proces zarządzania kontaktami z klientami, informowania ich o powstałej zaległości i motywowania do jej uregulowania. Jest to etap wczesnej windykacji, gdy klient opóźnia się ze spłatą maksymalnie do 60 lub 90 dni. Na początku trzeba sobie odpowiedzieć na podstawo-

Rysunek 8. Fragment arkusza kalkulacyjnego. Krzywa *bad rate*



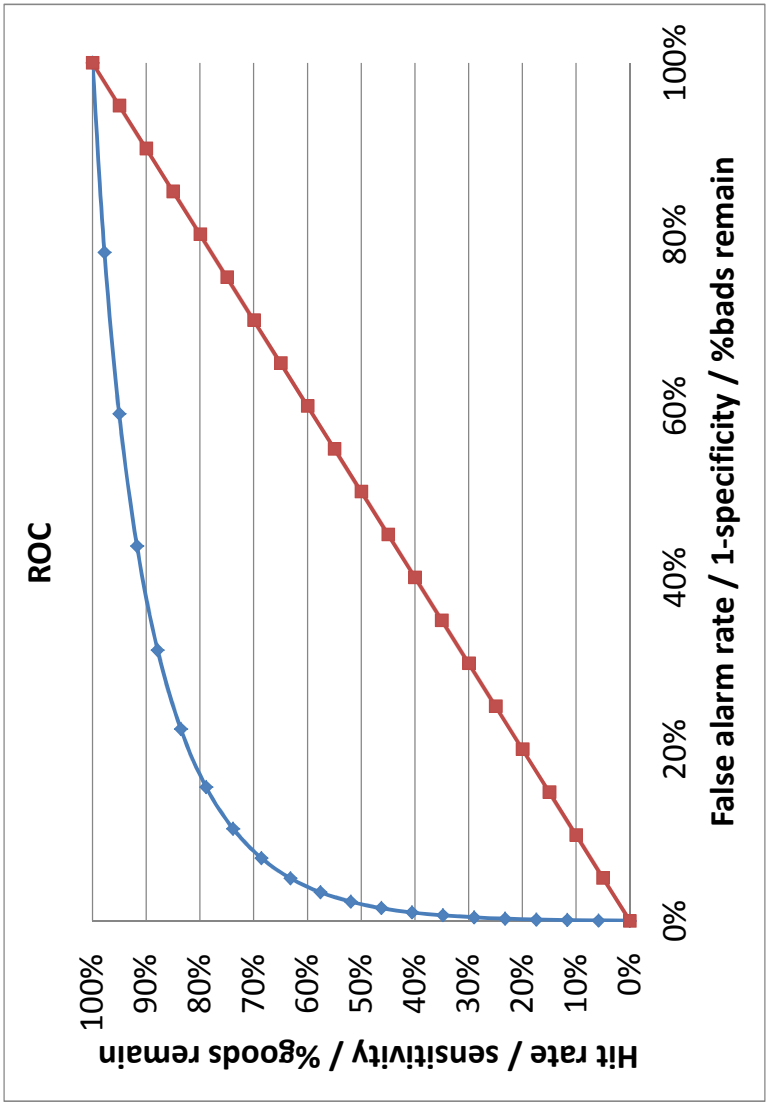
Źródło: opracowanie własne.

Rysunek 9. Fragment arkusza kalkulacyjnego. Krzywa CAP (*Cumulative Accuracy Profile*)



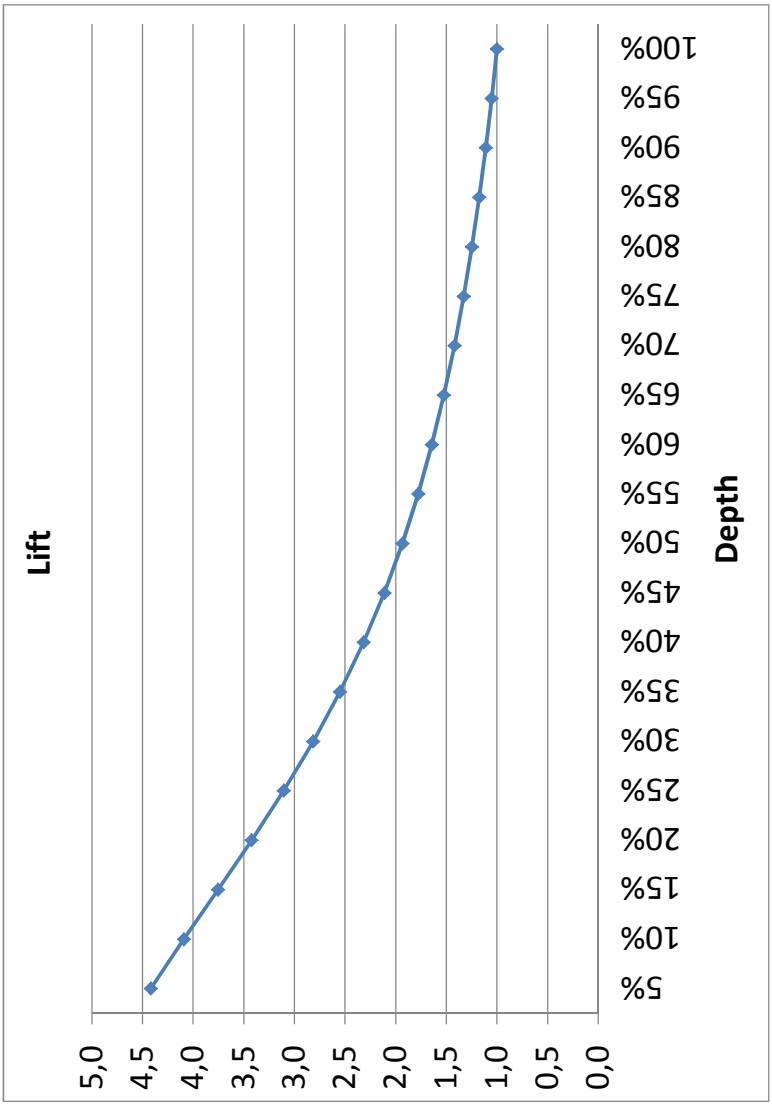
Źródło: opracowanie własne.

Rysunek 10. Fragment arkusza kalkulacyjnego. Krzywa ROC (*Receiver Operating Characteristic*)



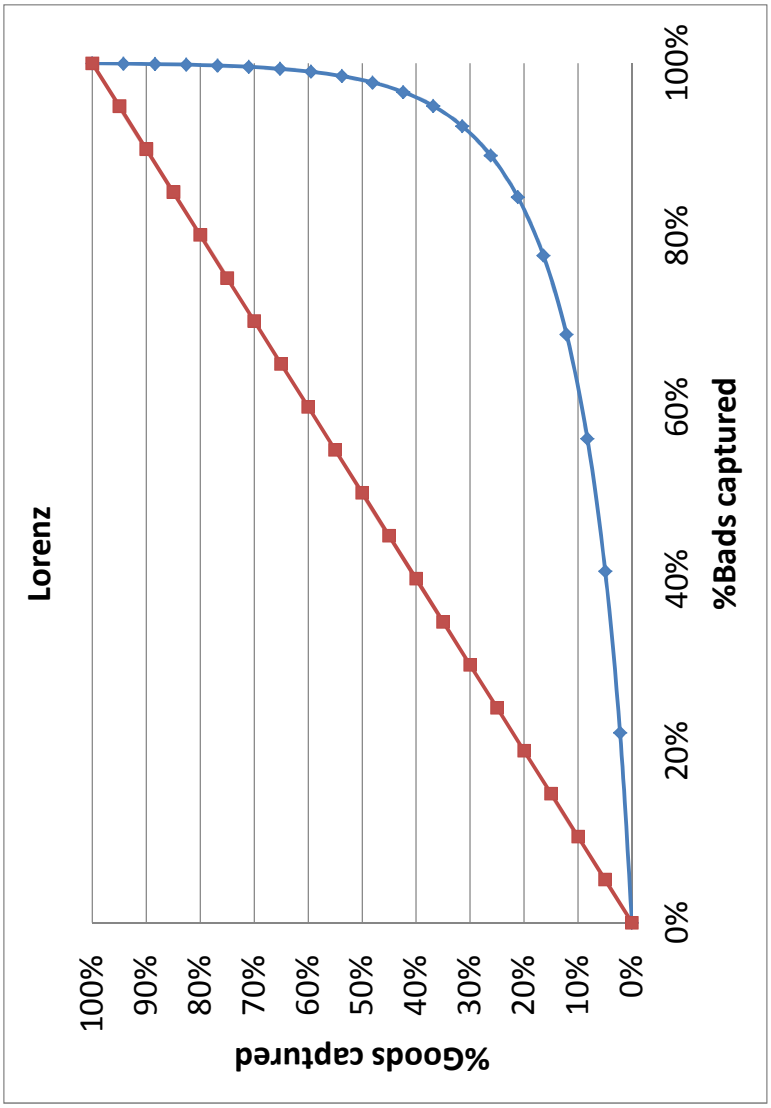
Źródło: opracowanie własne.

Rysunek 11. Fragment arkusza kalkulacyjnego. Krzywa *lift*



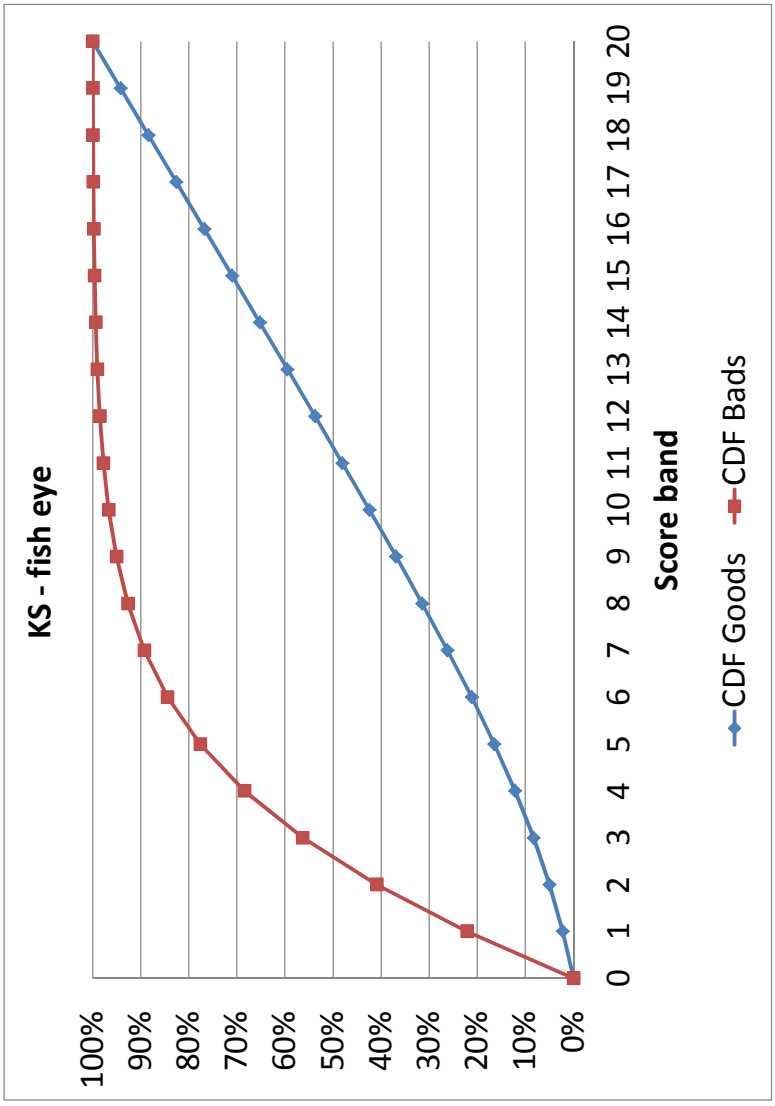
Źródło: opracowanie własne.

Rysunek 12. Fragment arkusza kalkulacyjnego. Krzywa Lorenza



Źródło: opracowanie własne.

Rysunek 13. Fragment arkusza kalkulacyjnego. Wykres rybie oko (*fish eye*) – wyznaczanie statystyki KS (Kołmogorowa–Smirnowa)



Źródło: opracowanie własne.

we pytanie: co daje nam ten proces? Czy windykacja powinna się utrzymywać z dodatkowych opłat windykacyjnych, które nalicza się klientowi z racji poniesienia kosztu kontaktu z nim? Jeśli odpowiedzielibyśmy twierdząco, to moglibyśmy zaburzyć poprawne funkcjonowanie banku. Moglibyśmy zniechęcić klientów spłacających kredyty, którzy czasem mają tendencję do małych opóźnień. Tego typu klient jest dla banku najcenniejszy, gdyż spłaca kredyty i dodatkowo jeszcze czasem płaci za przypominanie o mijających terminach wymagalności. Jeśli taki klient będzie obciążony zbyt dużymi opłatami windykacyjnymi, to „pomożemy" mu wpaść w jeszcze większe zadłużenie. Naliczenie opłat klientom niespłacającym odbije się tylko na większym mrożeniu kapitału ze względu na większe rezerwy. Nie jest zatem dobrym pomysłem regulowanie opłat do takich stawek, by windykacja miała się z tego utrzymać. Co zatem jest celem tego procesu? Bank ponosi koszt windykacji, by finalnie zmniejszyć stratę. Czy więc w przypadku windykacji można kwotę zmniejszonych strat uznać za jej przychód? To rozumowanie także nie jest poprawne. Windykacja ma sprawić, by straty spadły do mniejszego poziomu, co finalnie spowoduje ustalenie lepszego poziomu procentu akceptacji kredytowej w głównym procesie, w którym wnioski się akceptuje. Jeśli w podstawowym procesie będzie się akceptowć wnioski, które przynoszą dużą stratę, to potem windykacja może już tego nie naprawić. Nie jest ona bynajmniej lekarstwem na wszystkie błędy głównego procesu. Stanowi jednak narzędzie wspierające, które powoduje sprzężenie zwrotne. Na początku procesu akceptacji trzeba zaakceptować wnioski, potem część z nich trafi do windykacji i zostanie poprawiona wartość straty, która pierwotnie mogła być większa. Wreszcie finalna wartość straty spowoduje powstanie nowych reguł akceptacji, które będą już dostosowane (skalibrowane) do zmniejszonej straty na skutek działań windykacyjnych. Oznacza to, że zmniejszona strata nie jest zyskiem windykacji, tylko okazją lepszej kalibracji procesu akceptacji. Bank zatem ponosi koszt windykacji, który jest wstawiany w rachunek zysków i strat procesu akceptacji (głównego procesu).

Skoro ustaliliśmy już podstawowy cel procesu windykacji polubownej, czyli zmniejszanie straty, to możemy teraz przystąpić do omówienia szczegółowo wpływu modeli predykcyjnych, które także

w tym procesie są stosowane. Zanim jednak to nastąpi, musimy określić parametry pozwalające poprawnie mierzyć efektywność windykacji. Otóż najważniejsze pytanie brzmi: jaka byłaby strata, gdyby windykacji nie było? Pytanie to, podobne do pytania towarzyszącego pomiarowi ryzyka populacji, jak zwykle jest wyzwaniem analitycznym. Niemniej bez odpowiedzi na nie poniesionych kosztów pracy windykacji nie da się uzasadnić. Najgorszą formą marnotrawienia pieniędzy jest lęk przed testem. Jeśli nie stracimy, to nigdy nie nauczymy się oszczędzać. Mierzenie wpływu działań windykacji na zmianę straty banku jest jednym z jej kluczowych zadań. Z operacyjnego punktu widzenia istnieje wiele wskaźników, które doskonale pomogą usprawniać procesy i maksymalizować wykorzystanie czasu pracy operatorów telefonicznych. Są to takie miary, jak: udział efektywnych połączeń telefonicznych, liczba obietnic zapłaty itp. Jednak wszelkie tego typu statystyki nie pomogą nam w ustaleniu poziomu kosztu windykacji, który bank powinien ponieść.

Przeprowadźmy zatem symulację tego procesu (patrz arkusz kalkulacyjny o nazwie *collection_amicable_simulation.xlsx*). Przypuśćmy, że miesięcznie proces windykacji obejmuje 100 tys. kredytów (*number of collection entrances / accounts*) – patrz rysunek 6. Objęcie windykacją może być różnie definiowane, tu proponujemy proste podejście: klient ma jeden dzień opóźnienia, nie zapłacił w terminie raty kredytowej. Opóźnienie klikudniowe nie jest dużym problemem. Wielu klientów ma swój własny harmonogram rachunków budżetu rodzinnego. Z reguły jest on w dużym stopniu zależny od dnia w miesiącu, kiedy przychodzi wynagrodzenie. Nie jest zatem możliwe, by klient dotrzymywał terminu spłaty narzuconego przez bank, gdyż np. dzień po dacie wymagalności klient dostaje wypłatę. W takim wypadku będzie zrozumiałe to, że klient regularnie dokonuje wpłat kilka dni po dacie wyznaczonej przez bank. Dzwonienie do takiego klienta wydaje się stratą czasu. Ten przypadek jest przykładem klienta samospłacającego (ang. *selfpayment*). Takich klientów w portfelu windykacyjnym może być dość dużo. Jedno z ważnych zadań dla windykacji to właśnie ich identyfikacja.

Powinniśmy umieć oddzielać klientów, z którymi musimy się skontaktować, bo inaczej nie zapłacą, od takich, którzy zrobią to nawet bez naszej interwencji. Zadanie wcale nie jest takie proste,

bo jak zwykle wymaga podejścia analitycznego, a nie operacyjnego. Trzeba pozwolić na testy, podczas których część klientów kwalifikujących się do windykacji nie jest objęta jakimkolwiek działaniem windykacyjnym. Taka metoda jest dobrze znana w marketingu. Tu przy wszystkich kampaniach reklamowych tworzy się najczęściej dodatkowo grupy kontrolne, czyli losowo wybrane małe podzbiory klientów, wobec których nie wykonuje się żadnych akcji. W ten sposób mamy możliwość zmierzenia poziomu referencyjnego, aktywności klientów bez ponoszenia dodatkowych kosztów marketingowych. Niestety w przypadku windykacji brak akcji może oznaczać straty większe niż koszty kontaktu z klientem. Mamy tu zatem sytuację odwrotną, na którą trudniej się zdecydować. Dzięki grupom kontrolnym potrafimy policzyć, czy opłaca nam się kontaktowanie z klientami. Mierzymy mianowicie różnicę w *response rate* (udział respondentów) lub w przypadku windykacji *bad rate* dla segmentu *score band* pomiędzy grupą, w której wykonywano akcje, i grupą kontrolną. Metoda ta ma swoją angielską nazwę – *uplift* (albo: *incremental modelling*, *true lift modelling* czy *net modelling*) i jest popularna już od 1999 roku (Radcliffe i Surry, 1999; Lo, 2002). Pozwala ona nie tylko identyfikować klientów, którzy zareagują pozytywnie na naszą akcję (staną się respondentami), ale także takich, którzy będą mniej chętni do reakcji ze względu na jej wykonanie. Możemy wtedy mówić o wypaleniu klienta. Ten przypadek odnosi się przede wszystkim do działań marketingowych, ale także może być obecny w przypadku windykacji, tu również klienta możemy zniechęcić ze względu na przypominanie mu o czymś, o czym dobrze wie i co właśnie zamierzał zrobić. Efekt zniechęcenia być może nie objawi się brakiem spłaty zaległego zobowiązania, ale np. rezygnacją z odbierania przyszłych telefonów z działów zarówno windykacji, jak i marketingu. Finalnie możemy takiego klienta stracić.

Innym sposobem zarządzania windykacją jest koncentracja na zaobserwowanym ryzyku kredytowym, czyli liczonej statystyce *bad rate*. Jeśli klient należy do grupy z bardzo małym ryzykiem kredytowym, to nie musimy do niego dzwonić. Jest to całkiem uzasadniona koncepcja stosowana w praktyce, aczkolwiek nie do końca odpowiednia. Windykacja jest zjawiskiem, które zaburza poprawne wnioskowanie statystyczne. Właściwe pytanie brzmi: jakie jest ry-

zyko przy podjętych działaniach windykacyjnych, a jakie bez nich? Otóż możemy niechcący popełnić błąd podobny do błędu spotykanego przy zagadnieniu wpływu wniosków odrzuconych, omawianym w kolejnych podrozdziałach. Jeśli nasza dotychczasowa strategia windykacyjna polegała na kontaktowaniu się z daną grupą klientów i ryzyko ich jest stosunkowo małe, to nie należy wnioskować z tego, że nie warto do takich klientów dzwonić. Może się niestety okazać, że owo małe ryzyko jest właśnie efektem pracy windykacji. Trudno jest poprawnie odseparować wpływ windykacji od działania klienta, który sam z siebie może chcieć uregulować zobowiązanie. Podsumowując, należy stwierdzić, że musimy wykonać wiele testów, by rozpoznać wpływ windykacji na zmianę ryzyka.

Scenariusz pierwszy – mały wpływ na ryzyko

Przypuśćmy, że działania windykacyjne nie są bardzo skuteczne. Załóżmy dla uproszczenia, że mamy tylko dwa sposoby upominania klientów. Jeden nazwany strategią LOW – związany z niskim kosztem i niską skutecznością, przypuśćmy na poziomie $0{,}998$ (*low strategy risk improvement*), patrz tabela 6. Oznacza to, że ryzyko na skutek działania strategii LOW zmniejsza się o 2 promile. Druga strategia, o nazwie HIGH, ma skuteczność minimalnie większą, na poziomie $0{,}995$ (*high strategy risk improvement*); oczywiście jest ona także kosztowniejsza. Uwzględniamy tu głównie koszty związane z: zatrudnieniem operatorów w Call Center, wykonywaniem połączeń telefonicznych i wysyłaniem SMS-ów. Koszt przypadający na jeden kredyt obsługiwany w windykacji możemy obliczyć, zakładając także, że jeden operator w miesiącu może wykonać około 2000 rozmów. Dodatkowo ustalamy koszt połączenia telefonicznego na 8 PLN (w koszt ten włączone jest także utrzymanie centrali). Koszt tańszego serwisu, wysłania SMS-a lub automatycznego dzwonienia średnio wynosi 2 PLN (*automatic dialler or sms cost / per account*). Uwzględniając koszt jednego operatora – 8 tys. PLN (*FTE cost*), możemy już policzyć koszty dla dwóch strategii LOW i HIGH przypadające na jeden kredyt, mamy odpowiednio: 3 PLN (*low strategy cost per account*) i 12 PLN (*high strategy cost per account*). Za usługi windykacyjne klient ponosi dodatkowe opłaty, w naszym przypadku są to kwoty: 6 PLN (*high strategy collection fee / per account*)

Tabela 6. Fragment arkusza kalkulacyjnego. Parametry windykacji polubownej

Number of collection entrances / accounts	100 000
Average loan amount	7 000
High strategy cost per account	12
Low strategy cost per account	3
High strategy risk improvement	0,995
Low strategy risk improvement	0,998
High strategy collection fee / per account	6
Low strategy collection fee / per account	1
Number of accounts connected by operator / per months	2 000
FTE cost	8 000
Full staff cost	400 000
Operator cost / per account	4
Telephone connection cost / per account	8
Automatic dialler or sms cost / per account	2
LGD (Loss Given Default)	60%
Provision charged on disbursement day	0,5%
Global risk in collection (default24)	65%
Gini global	54,72%
Global loss	273 000 000
Final loss	271 814 040
Loss profit	1 185 960
Fees	107 600
Costs	885 000
Collection profit	-777 400
Total profit	408 560

Źródło: opracowanie własne.

i 1 PLN (*low strategy collection fee / per account*). Nie jest jednak dobrym pomysłem uwzględnianie tych opłat w przypadku klientów złych, niespłacających, gdyż tylko powiększa to poziom oczekiwanej straty. Innymi słowy, opłaty te nie zostaną pobrane od razu, tylko ewentualnie w dłuższym okresie odzyskiwania długu, czyli będą miały mniejszą wartość w momencie naliczania, po zdyskontowaniu. To, czy takie opłaty doliczać złym klientom, czy też nie, jest kwestią testów i sprawą do rozstrzygnięcia. Wzory zawsze można lekko zmienić. Jednocześnie, jak zaznaczyliśmy wcześniej, opłaty te nie powinny być duże, a zatem nie powinny też znacząco wpłynąć na pełną kalkulację opłacalności procesu.

Przypuśćmy, że portfel windykacyjny został podzielony na 20 grup związanych z modelem predykcyjnym, który ma moc około 55%. Pojawia się teraz pytanie: wobec której z grup stosować strategię LOW, a wobec której HIGH? Odpowiedzi na to pytanie można udzielić, przyjmując proste założenie: koszt związany ze zmniejszeniem straty nie może być większy od zaoszczędzonej straty. Jeśli zmniejszamy stratę o 10 PLN, to nie możemy tego robić zgodnie ze strategią, która będzie kosztować 20 PLN. W tym przypadku tylko tracimy pieniądze. Kryterium to jest łatwe do zasymulowania. Przy parametrach naszego procesu okazuje się, że strategię LOW zastosujemy tylko wobec pierwszych siedmiu grup *score band*. Trzeba tu wspomnieć o dodatkowych przychodach w windykacji związanych z naliczaniem opłat: za LOW – 1 PLN, a za HIGH – 6 PLN. Możemy zatem wyznaczyć trzy wielkości: ujemny wynik windykacji (*collection profit*), równy przychodom z opłat minus koszty windykacyjne, zysk ze zmniejszonej straty (*loss profit*) i całkowity zysk (*total profit*). Te trzy wskaźniki mogą być teraz policzone dla różnych modeli z różnymi mocami predykcyjnymi (patrz tabela 7).

Przy małych parametrach skuteczności windykacji dość trudno jest wykazać jej przydatność. Analizując rysunek 14, możemy zauważyć, że, owszem, całkowity zysk rośnie wraz ze wzrostem mocy predykcyjnej modelu, ale nie jest to takie istotne – każdy 1% wzrostu mocy predykcyjnej daje około 1 tys. PLN poprawy całkowitego zysku. Sytuacja jest inna w przypadku zysku z windykacji i zysku ze zmniejszenia straty – mamy tu nawet zależności odwrotne. Przypadek ten jest o tyle ważny, że pokazuje sytuację, w której istnie-

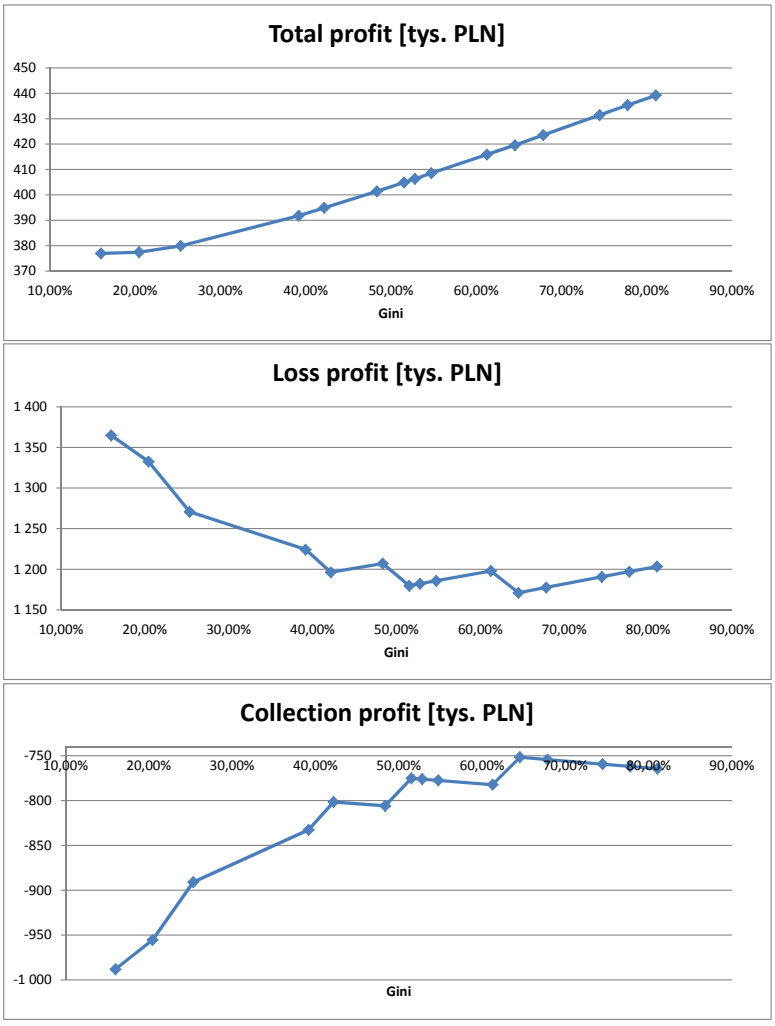
Tabela 7. Fragment arkusza kalkulacyjnego. Wskaźniki windykacji zależne od mocy predykcyjnej modelu w przypadku małej skuteczności

b	a	Gini	Collection profit	Total profit	Loss profit
1,2	-0,12	54,72%	-777 400	408 560	1 185 960
1,2	-0,16	81,05%	-764 334	439 197	1 203 531
1,2	-0,155	77,76%	-761 832	435 359	1 197 191
1,2	-0,15	74,46%	-759 298	431 471	1 190 768
1,2	-0,14	67,84%	-754 155	423 581	1 177 736
1,2	-0,135	64,53%	-751 559	419 598	1 171 157
1,2	-0,13	61,24%	-782 192	415 911	1 198 102
1,2	-0,12	54,72%	-777 400	408 560	1 185 960
1,2	-0,117	52,79%	-775 963	406 355	1 182 319
1,2	-0,115	51,51%	-775 007	404 889	1 179 896
1,2	-0,11	48,34%	-805 795	401 439	1 207 234
1,2	-0,1	42,16%	-801 536	394 905	1 196 440
1,2	-0,095	39,15%	-832 555	391 810	1 224 366
1,2	-0,07	25,32%	-890 891	379 940	1 270 831
1,2	-0,06	20,43%	-955 327	377 431	1 332 758
1,2	-0,05	15,96%	-988 050	376 950	1 365 000

	Total profit	Loss profit	Collection profit
1% of Gini	956	2 481	3 437 PLN
5% of Gini	4 782	12 404	17 186 PLN
10% of Gini	9 564	24 809	34 373 PLN

Źródło: opracowanie własne.

Rysunek 14. Fragment arkusza kalkulacyjnego. Wykresy liniowe wskaźników windykacji zależne od mocy predykcyjnej modelu w przypadku małej skuteczności



Źródło: opracowanie własne.

nie modelu predykcyjnego nie przynosi dodatkowych zysków czy oszczędności. Musimy zdawać sobie sprawę z tego, że niektóre układy danych czy wartości parametrów nie pozwolą nam dobrze wykorzystać modeli predykcyjnych. To bardzo ważna nauka dla inżyniera danych. Są takie procesy i takie sytuacje, w których jego praca nie przyniesie korzyści i ważne jest to, aby po prostu nie poświęcał na nie swojego czasu. Niestety wiele procesów bankowych może być obarczonych tego typu wadą. Czasem wynika to ze zbyt dużej liczby błędów operacyjnych. Jeśli np. model jest poprawnie wdrożony i ma dużą moc predykcyjną, ale systemy do obsługi wysyłania SMS-ów będą niepoprawnie konfigurowane i będą wysyłały je do złej grupy klientów, to może się okazać, że cała subtelność „pomiaru mikrometrem" związana ze stosowaniem modelu predykcyjnego zostanie zniszczona „cięciami siekiery" błędnie wysyłanych SMS-ów. Trzeba stwierdzić jednoznacznie, że modele optymalizujące procesy mają sens tylko wtedy, gdy procesy te są w pełni kontrolowane, sterowalne i w miarę stabilne w czasie.

Scenariusz drugi – istotny wpływ na ryzyko

Wystarczy zmienić dwa parametry skuteczności strategii LOW i HIGH na wartości odpowiednio – $0{,}95$ i $0{,}9$, by nagle uzyskać zupełnie inne wartości wskaźników (patrz arkusz o nazwie *collection_amicable_simulation2.xlsx*). W tym przypadku zawsze dla każdej grupy *score band* opłaca się stosować strategię HIGH. Zmniejszenie straty jest na tyle istotne, że całkowicie pokrywa wszelkie koszty windykacyjne. Oczywiście nie chodzi tu o wydanie dużych sum w związku z windykacją. W tym przypadku decydujemy się zatem na użycie modelu i przykładowo wyznaczamy proporcje pomiędzy LOW i HIGH – 50%:50%. Dzięki temu możemy przekonać się, jak bardzo zmniejszamy straty, mając większą moc predykcyjną modelu (patrz tabela 8). W tym wypadku także nie uzyskujemy dużych zmian w koszcie windykacji, jest on prawie stały, natomiast zależności pomiędzy mocą predykcyjną a całkowitym zyskiem lub zyskiem ze zmniejszenia straty są liniowe (patrz rysunek 15). Oznacza to, że przyrost mocy predykcyjnej o 1% przynosi około 36 tys. PLN całkowitego zysku.

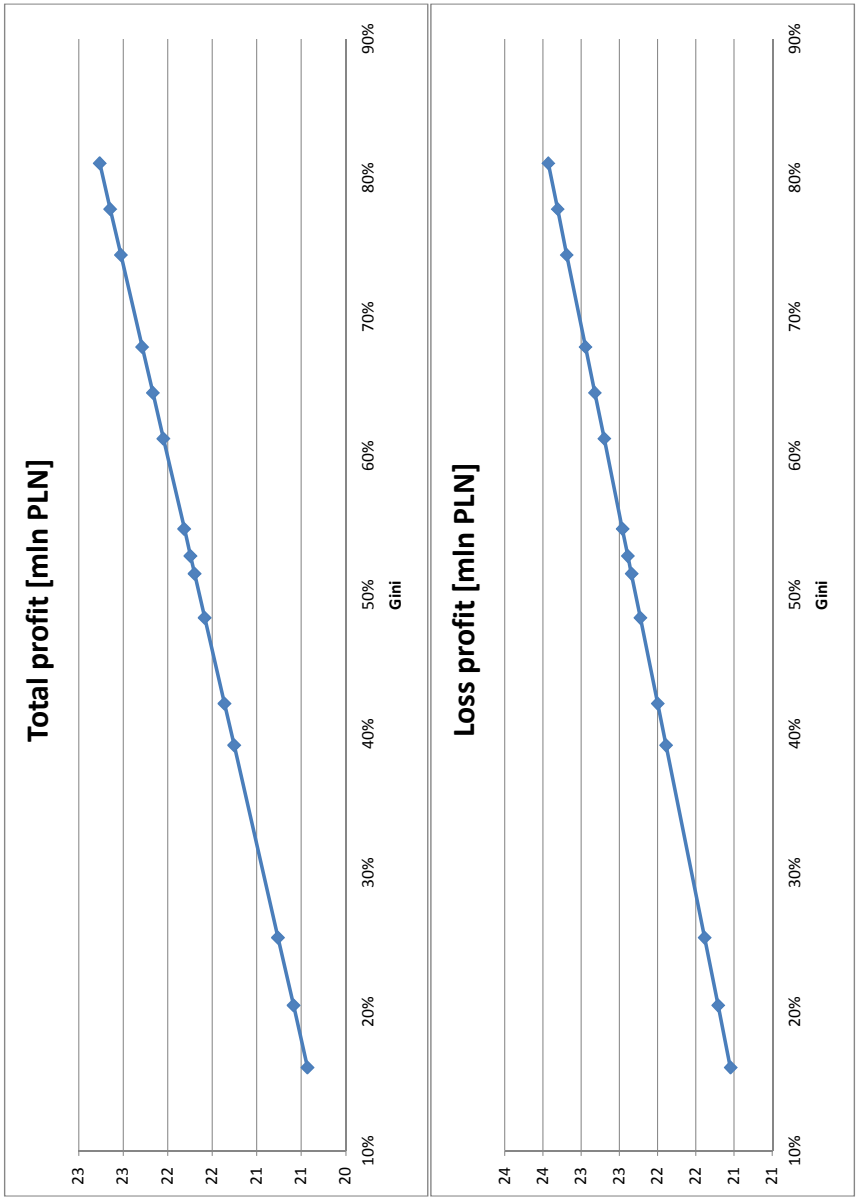
Tabela 8. Fragment arkusza kalkulacyjnego. Wskaźniki windykacji zależne od mocy predykcyjnej modelu w przypadku istotnej skuteczności

b	a	Gini	Collection profit	Total profit	Loss profit
1,2	-0,16	81,05%	-669 080	22 765 042	23 434 122
1,2	-0,16	81,05%	-669 080	22 765 042	23 434 122
1,2	-0,155	77,76%	-666 509	22 646 270	23 312 779
1,2	-0,15	74,46%	-663 926	22 526 954	23 190 879
1,2	-0,14	67,84%	-658 746	22 287 690	22 946 436
1,2	-0,135	64,53%	-656 161	22 168 271	22 824 431
1,2	-0,13	61,24%	-653 586	22 049 359	22 702 945
1,2	-0,12	54,72%	-648 495	21 814 184	22 462 678
1,2	-0,117	52,79%	-646 988	21 744 586	22 391 574
1,2	-0,115	51,51%	-645 990	21 698 489	22 344 479
1,2	-0,11	48,34%	-643 521	21 584 432	22 227 953
1,2	-0,1	42,16%	-638 713	21 362 333	22 001 045
1,2	-0,095	39,15%	-636 385	21 254 815	21 891 200
1,2	-0,07	25,32%	-625 710	20 761 721	21 387 431
1,2	-0,06	20,43%	-621 962	20 588 582	21 210 544
1,2	-0,05	15,96%	-618 542	20 430 630	21 049 172

	Total profit	Loss profit	Collection profit	
1% of Gini	35 867	-36 643	-776	PLN
5% of Gini	179 334	-183 216	-3 882	PLN
10% of Gini	358 667	-366 432	-7 765	PLN

Źródło: opracowanie własne.

Rysunek 15. Fragment arkusza kalkulacyjnego. Wykresy liniowe wskaźników windykacji zależne od mocy predykcyjnej modelu w przypadku istotnej skuteczności



Źródło: opracowanie własne.

Można teraz porównać cztery scenariusze (patrz arkusz o nazwie *collection_amicable_simulation3.xlsx*), w pierwszym najkosztowniejszym, w którym wszędzie stosujemy strategię HIGH (patrz tabela 9), mamy zysk windykacji -950 tys. PLN i zmniejszenie straty o 27 mln PLN. Strategia LOW stosowana wszędzie zmniejszy zysk windykacji o 700 tys. PLN, ale stratę tylko o $13{,}5$ mln PLN. Kiedy zaczniemy stosować modele predykcyjne, to możemy znaleźć rozwiązania pośrednie. Dla modelu o mocy około 16% możemy, stosując proporcje $50\%{:}50\%$ strategii LOW i HIGH, uzyskać -600 tys. PLN zysku windykacyjnego i 21 mln PLN zmniejszenia straty. Stosując model o mocy 60%, mamy porównywalny zysk windykacyjny, ale istotnie – o $1{,}6$ mln PLN – mniejsze straty w stosunku do modelu z wartością Giniego 16%. To pokazuje, że istotą stosowania modeli predykcyjnych (skoringowych) w windykacji jest lepsza alokacja zasobów, by przy porównywalnym koszcie operacyjnym uzyskać istotnie mniejsze straty albo przy mniejszym koszcie windykacyjnym te same straty.

Na zakończenie zwróćmy jeszcze uwagę na dwa szczegóły. Wartości parametrów w obliczaniu opłacalności procesu windykacji polubownej są przykładowe i nie do końca zgodne z rzeczywistością. Należy się spodziewać, że w prawdziwym procesie parametry skuteczności strategii powinny być znacznie lepsze, liczba wykonywanych rozmów przez operatora też może być większa. Układ parametrów jest tu bardzo ważny. Finalne wyniki są bardzo wrażliwe na zmianę jakiegokolwiek z nich. Oznacza to, że w niektórych układach użyteczność modelu predykcyjnego nie będzie istotna, co więcej – może powodować, że nie warto ponosić dużych kosztów windykacji polubownej. Drugi szczegół dotyczy trudności związanych z umiejętnym odseparowaniem wpływu windykacji polubownej od kolejnych etapów windykacji, w szczególności windykacji prawnej, w trakcie której długi odzyskuje się przez pracę komornika. Jeśli do budowy modelu i do ustalania strategii używa się definicji *default* z okresem obserwacji 12 czy 24 miesięce[1], to wskaźnik ten zahacza o czas windykacji prawnej i efekt spłacenia zobowiązania może

[1] W przypadku windykacji stosuje się inną definicję zdarzenia *default*, zmienia się punkt obserwacji z daty wniosku na datę objęcia windykacją (patrz rysunek 1, str. 35).

być związany z pracą albo komornika, albo operatora telefonicznego, albo obu naraz. Błędne będzie tu wnioskowanie statystyczne, które wskaże przewagę komornika nad operatorem lub na odwrót. Dobrze jest w takim wypadku używać definicji *default* z krótszym okresem obserwacji, nawet do 3 miesięcy, dlatego że wtedy uwzględni się tylko windykację polubowną. Wszystkie tego typu szczegóły dopracowuje się przez nieustające testy i analizy. Nie da się całego procesu zdefiniować i zoptymalizować od razu, trzeba próbować i ciągle wprowadzać poprawki.

Tabela 9. Fragment arkusza kalkulacyjnego. Finalne porównanie wskaźników windykacji w przypadku istotnej skuteczności

Strategy	Collection profit	Loss profit	Difference
Only high	-951 000	27 300 000	13 650 000
Only low	-261 750	13 650 000	
Gini 60% 50/50	-653 586	22 702 945	1 653 772
Gini 16% 50/50	-618 542	21 049 172	

Źródło: opracowanie własne.

3.5. Przypadek procesu akceptacji kredytów hipotecznych

Proces akceptacji kredytów hipotecznych znacząco różni się od procesu akceptacji drobnych kredytów portfela *Consumer Finance*. Każda różnica wymaga tu dokładniejszego rozważenia.

Czas kredytowania – zamiast być kilkuletni – jest kilkudziesięcioletni, w naszym przypadku wynosi aż 240 miesięcy (patrz arkusz o nazwie *mortgage_simulation.xlsx*). To całkowicie zmienia podejście do zarządzania ryzykiem. Kto z nas jest w stanie przewidzieć

sytuację rynków finansowych za 20 lat? Nie możemy zatem wierzyć w dużą moc modeli predykcyjnych rozróżniających klientów w momencie aplikowania. Wszelkie wskaźniki liczone w tym czasie nie są w stanie prognozować poprawnie zachowania klienta w ciągu najbliższych kilkudziesięciu lat. Przyjęta tu definicja *default* z okresem 240 miesięcy jest więc tylko hipotetyczna. Mało banków ma tak długą historię danych. Budowa modeli oparta na definicji *default* z tak długim okresem obserwacji (patrz rysunek 1, str. 35) powoduje cofnięcie się o więcej niż 20 lat, czyli do sytuacji na rynku znacznie odbiegającej od dzisiejszej. Nikt nie zbuduje dobrego modelu w takim podejściu. Modele w praktyce muszą być zatem budowane na bazie krótszego okresu obserwacji. Dobór tego okresu musi być oczywiście umiejętnie ustalony, by z jednej strony był odpowiednio krótki, a z drugiej zawierał już dość duży udział wszystkich zdarzeń *default*, do których może dojść w całym okresie kredytowania. Tego typu analiza i tak nie da poprawnych wyników, gdyż w ciągu 20 lat może pojawić się wiele poważnych zmian rynkowych, co powoduje, że w przypadku kredytów hipotecznych bardzo ważną rolę odgrywają modele testów warunków skrajnych (ang. *stress testing* (BIS, 2009)).

Jeśli pogodzimy się już z faktem słabości modeli predykcyjnych stosowanych do prognozowania zjawisk w ciągu następnych 20 lat, to zaczynamy uświadamiać sobie znaczenie innych narzędzi bankowych w utrzymaniu portfela hipotecznego. Muszą tu zatem odgrywać ważną rolę windykacja i wszelkie metody monitoringu aktualnej kondycji finansowej klientów mających kredyty w naszym banku. Powinniśmy być otwarci na różnego rodzaju restrukturyzację lub inne rodzaje zmian warunków kredytowania w przypadkach, gdy bank pozyska informacje o zmianie warunków finansowych klienta.

W procesie tym ważną rolę odgrywa też sposób zabezpieczenia kredytu. Co więcej, może się on zmieniać w czasie. Przykładowo, w okresie osłabienia prosperity można prosić klienta o dodatkowe ubezpieczenie. Niestety ważne są tu także wszelkie zmiany rynku nieruchomości. Słynna statystyka LtV (ang. *loan to value*) wpływa istotnie na sposoby segmentacji portfela.

Wreszcie, jak w żadnym innym procesie, tak w przypadku hipoteki proces zbierania dokumentów składanych razem z wnioskiem

kredytowym i ich analizy staje się bardzo kosztowny i trudny w zarządzaniu. Potrzebne są tu różnego rodzaju ręczne procesowania. Nie da się wszystkich etapów akceptacji przeprowadzić w sposób czysto automatyczny. Odgrywa tu zatem rolę indywidualna decyzja analityka kredytowego, która może lekko zaburzyć wnioskowanie statystyczne. Nie można jednak tego zlikwidować, gdyż duża liczba dokumentów jest potrzebna, by jak najlepiej się zabezpieczyć na okres najbliższych 20 lat.

Przechodząc już do kalkulacji rentowności procesu akceptacji, należy stwierdzić, że nie możemy zapomnieć o wartości pieniądza w czasie. W tym wypadku do liczenia przychodu banku nie wolno brać całkowitego oprocentowania, musimy je pomniejszyć o koszty pozyskania kapitału. Będziemy tu zatem wykorzystywać marżę banku, czyli część tych odsetek, które są już czystym zyskiem, po uwzględnieniu kosztu kapitału (dyskontowania).

Nawet jeśli $default_{240}$ jest pojęciem niepraktycznym, to i tak pozwoli nam na zbadanie wpływu przyrostu mocy predykcyjnej na przyrost zysku banku. Ryzyko obserwowane portfeli hipotecznych z reguły nie jest bardzo duże, czasem zdarzają się nawet portfele, w których liczba zdarzeń *default* jest tak mała, że powoduje dość duże trudności w budowie modeli, stosuje się wtedy specjalne metody związane z pojęciem rzadkich zdarzeń (ang. *rare events*) lub grubych ogonów (ang. *fat tails*). Małe ryzyko jest związane z metodą zabezpieczenia przez wpis hipoteki przymusowej, które powoduje, że klient może zostać pozbawiony możliwości mieszkania. Powoduje to też małą wartość parametru LGD (na poziomie 30%). Edward Altman, twórca słynnego modelu Z-score (Altman, 1968), na konferencjach powtarza często zdanie, że budujący modele skoringowe kochają bankrutów – musi być ich wystarczająco dużo, by móc budować modele. Jest to swego rodzaju dylemat inżyniera danych, gdyż analizuje on przypadki często związane z ludzkimi dramatami. Jednocześnie jednak możemy nadać temu wygodną interpretację: trudne przypadki analizuje się po to, by było ich w przyszłości mniej, by ochronić przed poniesieniem straty nie tylko bank, ale także klienta, by zlikwidować jego życiowe problemy.

Mamy zatem ryzyko populacji o wartości 40%, być może za duże, ale trzeba pamiętać o tym, że ostatni kryzys był spowodowa-

ny między innymi udzielaniem kredytów mieszkaniowych bezrobotnym. Przy mocy predykcyjnej o wartości 80%, dość dużej jak na okres obserwacji zdarzenia *default*, ryzyko akceptowane jest na poziomie $2{,}5\%$, co mniej więcej zgadza się z rzeczywistością. Prowizja i marża odsetkowa przyjmują wartość $0{,}5\%$ (patrz tabela 10).

Akceptujemy średnio 35% z 10 tys. wniosków miesięcznie. Średnia kwota kredytu wynosi 300 tys. PLN. Przy tak określonych parametrach mamy 33 mln PLN zysku miesięcznie. W przypadku pełnej akceptacji otrzymalibyśmy wynik ujemny -260 mln PLN miesięcznie. Podobnie jak w przypadku kredytu ratalnego przeprowadzamy nasze kalkulacje dla różnych wartości mocy predykcyjnej modeli (tabela 11). Okazuje się, że 5-procentowa zmiana mocy powoduje zwiększenie zysku aż o 4 mln PLN. Liczba ta daje do myślenia. Zauważmy też, że już przy mocy 44-procentowej zarabiamy około $1{,}6$ mln PLN miesięcznie, co w stosunku do -260 mln PLN jest już pewnym osiągnięciem. Niestety zależność wartości mocy od zysku w tym przypadku nie jest już w pełni liniowa, co pokazano na rysunku 16, ale jest wystarczająco zbliżona do liniowej, by możliwe było traktowanie naszych wyników jako wiarygodnych.

Podjęta próba liczenia rentowności procesu akceptacji kredytów hipotecznych może być przeanalizowana jeszcze inaczej. Wielu czytelników z pewnością skrytykuje analizowane liczenie oczekiwanej straty w okresie 20 lat w formie iloczynu prostych stałych wskaźników EAD i LGD. Każdego klienta zdarzenie *default* może dotyczyć w zupełnie innym czasie spłacania, który jest tu dość długi. Wartość iloczynu EAD i LGD nie powinna być stałą równą kwocie kredytu pomnożonej przez 30% dla wszystkich klientów. Jest to dość duże przybliżenie. Można zatem w przypadku akceptacji kredytów hipotecznych zagadnienie to sprowadzić do pytania: o ile zmieni się wartość straty oczekiwanej, innymi słowy – rezerw odkładanych przez bank dla pierwszego roku historii? Uwzględniamy wtedy $default_{12}$ i liczymy tylko zmianę straty względem zmiany mocy predykcyjnej. W kolejnych latach historii kredytów w rzeczywistym procesie są wyliczane modele IRB, takie jak PD, LGD i EAD, które pozwalają w każdym roku rozliczeniowym wyliczać odpowiednie wymogi kapitałowe. Oczywiście także w tym wypadku możemy sobie zadać pytanie o to, w jakim stopniu moce predykcyjne wszystkich tych mo-

Tabela 10. Fragment arkusza kalkulacyjnego. Parametry procesu akceptacji kredytów hipotecznych

Average loan amount	300 000
Average number of installments	240
Net margin	0,5%
LGD (Loss Given Default)	30%
Provision charged on disbursement day	0,5%

Gini global	80,85%
Gini on accepted	28,05%

Global risk in market (default240)	40%
Accepted risk	2,41%
Acceptance rate	35,00%

Global loss	360 000 000
Global income	100 874 412
Global profit	-259 125 588

Accepted loss	21 698 060
Accepted income	54 790 112
Accepted profit	33 092 052

Źródło: opracowanie własne.

deli wpływają na wyliczane wymogi. Nawet jeśli nasz uproszczony sposób liczenia rentowności nie jest dokładny i opiera się na dość prostych założeniach, to i tak wniosek z naszej analizy jest uniwersalny.

Podsumowując przypadek procesu akceptacji kredytów hipotecznych, możemy śmiało stwierdzić, że – chociaż za pomocą modelu używanego w procesie nie jesteśmy w stanie przewidzieć wszystkich zdarzeń *default* w okresie 20 lat – to i tak zwiększenie jego mocy predykcyjnej przynosi nam milionowe zyski miesięcznie.

3.6. Strategie akceptacji (jak zarządzać procesem?)

Podrozdział ten stanowi zmodyfikowaną wersję publikacji *Rola danych symulacyjnych w badaniach Credit Scoring* (Przanowski, 2014b) oraz rozdziału „Model biznesowy: akwizycja i sprzedaż krzyżowa" książki *Credit Scoring w erze Big-Data* (Przanowski, 2014a). Jest on potrzebny do spójnego przedstawienia problemów Credit Scoringu, w szczególności by lepiej rozumieć rolę strategii i procesu wdrożenia modeli oraz by nie pomijać problemu wniosków odrzuconych, co stanowi największe wyzwanie związane z Credit Scoringiem.

Większość właścicieli firm finansowych, mając nadzieję, że kryzys już się skończył, rozpoczęło walkę o klienta na dużą skalę. Czas kryzysu (2008–2009) był okresem, gdy banki bardzo ostrożnie zarządzały procesem akceptacji i znacząco zmniejszały populację akceptowaną. Dziś coraz częściej 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, których cechuje małe ryzyko, nie są aktywni kredytowo i 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 klientów zachęcić czymś wygodnym, atrakcyjnym i dość tanim albo zupełnie bezpłatnym, a jak się do nas przyzwyczają, to zaproponować produkty drogie, na których będzie się zarabiać. Ogólnie tego typu modele mają strukturę: tania akwizycja, droga sprzedaż krzy-

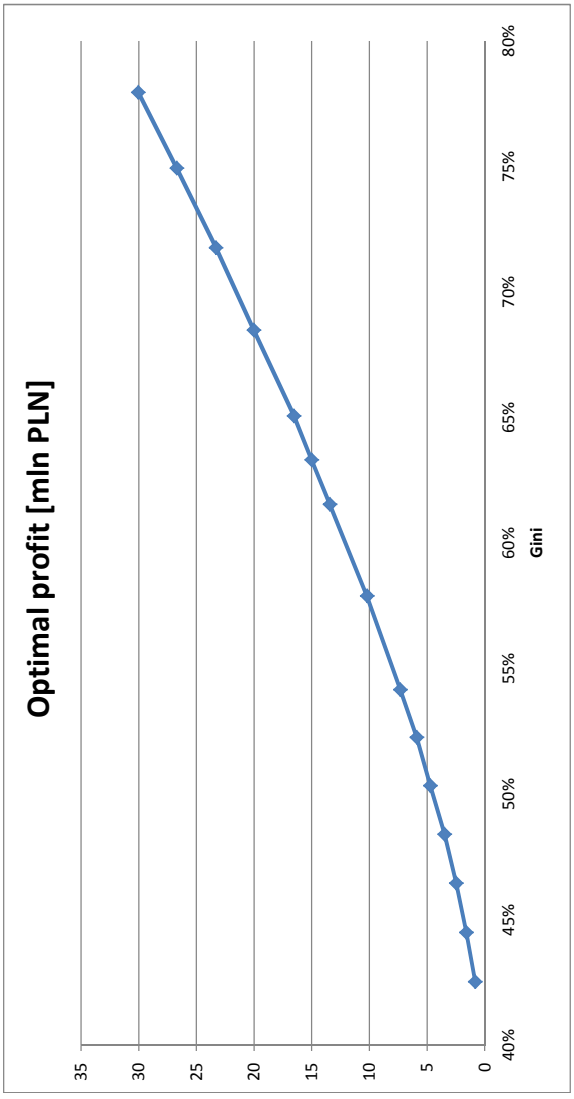
Tabela 11. Fragment arkusza kalkulacyjnego. Lista parametrów w zależności od zmieniającej się mocy predykcyjnej modelu dla procesu akceptacji kredytów hipotecznych

Accepted loss	b	a	Gini	Optimal profit	Opt acc rate
21 698 060	1	-0,25	80,85%	33 092 052	35,00%
24 267 340	1	-0,24	77,96%	30 042 819	35,00%
27 088 410	1	-0,23	74,93%	26 694 761	35,00%
22 869 549	1	-0,22	71,77%	23 295 523	30,00%
25 624 496	1	-0,21	68,48%	20 025 940	30,00%
21 465 655	1	-0,2	65,07%	16 555 470	25,00%
22 756 675	1	-0,195	63,32%	15 023 281	25,00%
24 107 104	1	-0,19	61,54%	13 420 586	25,00%
19 732 632	1	-0,18	57,90%	10 206 028	20,00%
22 176 134	1	-0,17	54,16%	7 306 069	20,00%
16 275 286	1	-0,165	52,26%	5 903 020	15,00%
17 271 822	1	-0,16	50,34%	4 720 328	15,00%
18 309 906	1	-0,155	48,40%	3 488 325	15,00%
12 076 010	1	-0,15	46,46%	2 480 539	10,00%
12 815 105	1	-0,145	44,50%	1 603 378	10,00%
6 388 078	1	-0,14	42,53%	824 801	5,00%

1%	of Gini	824 885	PLN
5%	of Gini	4 124 426	PLN
10%	of Gini	8 248 852	PLN

Źródło: opracowanie własne.

Rysunek 16. Fragment arkusza kalkulacyjnego. Wykres obrazujący współzależność pomiędzy optymalnym zyskiem i mocą predykcyjną modelu dla procesu akceptacji kredytów hipotecznych



Źródło: opracowanie własne.

żowa (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 branże FMCG czy AGD pełne są 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. Jest to fragment ogólnego modelu zwanego ACURA od angielskich słów: *acquisition*, *cross-sell*, *up-sell*, *retention* i *advocacy* (Payne, 2005, 2002). Klienci biorący na raty lodówkę czy telewizor plazmowy są bardzo zadowoleni 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 opłacalnym. Jest to najlepszy przykład zastosowania Credit Scoringu i wykazania jego użyteczności w osiąganiu niemal milionowych korzyści finansowych.

Omówione metody uzyskania przykładowych danych symulacyjnych służą właśnie pogłębieniu analiz i optymalizacji procesu biznesowego, który można prosto nazwać: najpierw kredyt ratalny, potem gotówkowy. Jeśli chcemy wykazać przydatność danych losowych w badaniach nad Credit Scoringiem 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 zagadnienia. Modelu 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.

Parametry modelu

Dużym atutem materiału prezentowanego w pracy jest właśnie oparcie się na studiach przypadków. Wszystkie przedstawione wyniki (w tym rozdziale) są rezultatem konkretnych obliczeń na sy-

mulacyjnych 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.

Wyniki symulacji, podstawowe raporty

Wszystkich obliczeń dokonano na laptopie Dell Latitude. Podstawowymi strukturami danych symulacyjnych są tabele: Produkcja i Transakcje (Przanowski, 2014a). 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 17.

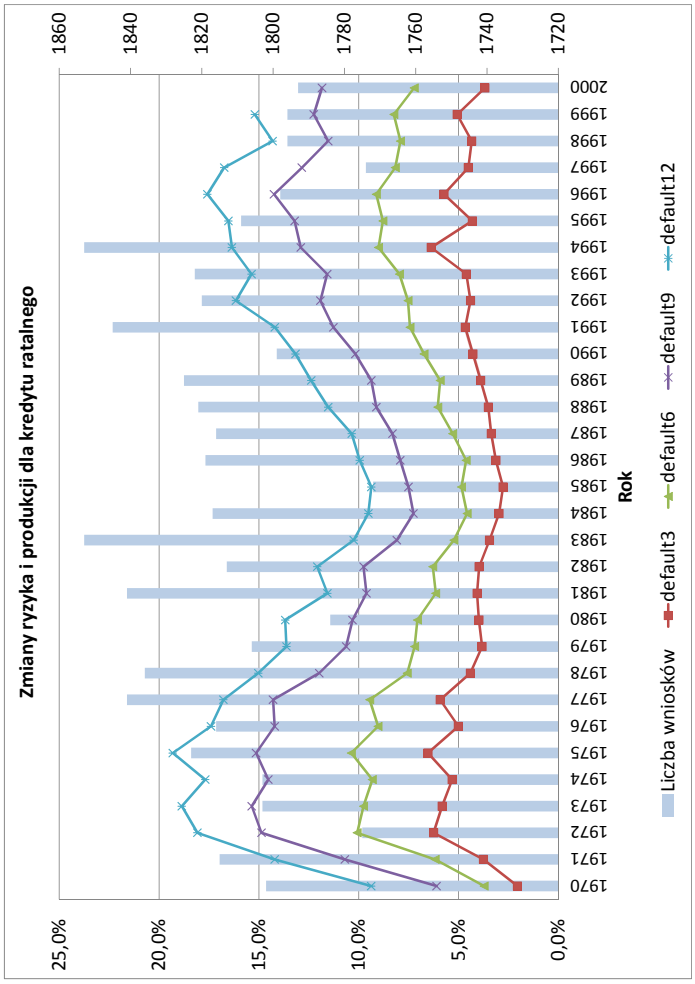
W przypadku kredytu gotówkowego czas obliczeń wynosi 11 godzin i 57 minut. Zbiór Produkcja_cross zawiera 60 222 wiersze i 19 kolumn, a Transakcje_cross 1 023 716 wierszy i 8 kolumn. Raport z wartości ryzyka przedstawiono na rysunku 18.

Miesięczne wielkości portfeli obu produktów razem oraz współczynnik konwersji (ang. *response rate*), z okresem obserwacji 1 miesiąca, przedstawiono na rysunku 19.

Implementacja modeli, system decyzyjny

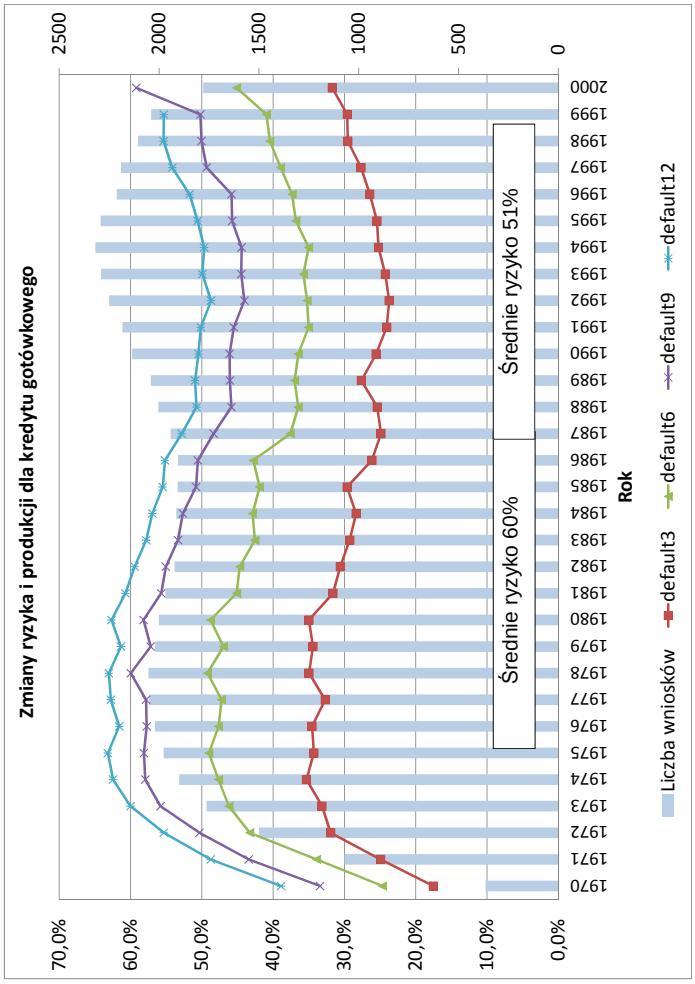
Wszystkie wygenerowane kredyty ratalne i gotówkowe są traktowane jako potencjalny portfel banku. Co szczegółowo opisano w podrozdziale 2.4, wszystkie kredyty są tak czy inaczej brane przez klientów. Ze względu na priorytety klienta, być może także ze względu na politykę banku, nie wszystkie kredyty nawet ten sam klient spłaca tak samo. Bank może obniżyć koszty strat kredytowych, ograniczając posiadanie kredytów w swoim portfelu. Umiejętny wybór kredytów jest zadaniem dla modeli skoringowych, które są zaimplementowane w systemie decyzyjnym (ang. *decision engine* lub *scoring engine*). Decyzja o akceptacji powoduje, że dany kredyt jest brany 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 stała, zmienia się tylko kredytodawca. System decyzyjny oczywiście nie zna przyszłej historii klienta, zna tylko jego histo-

Rysunek 17. Kredyt ratalny



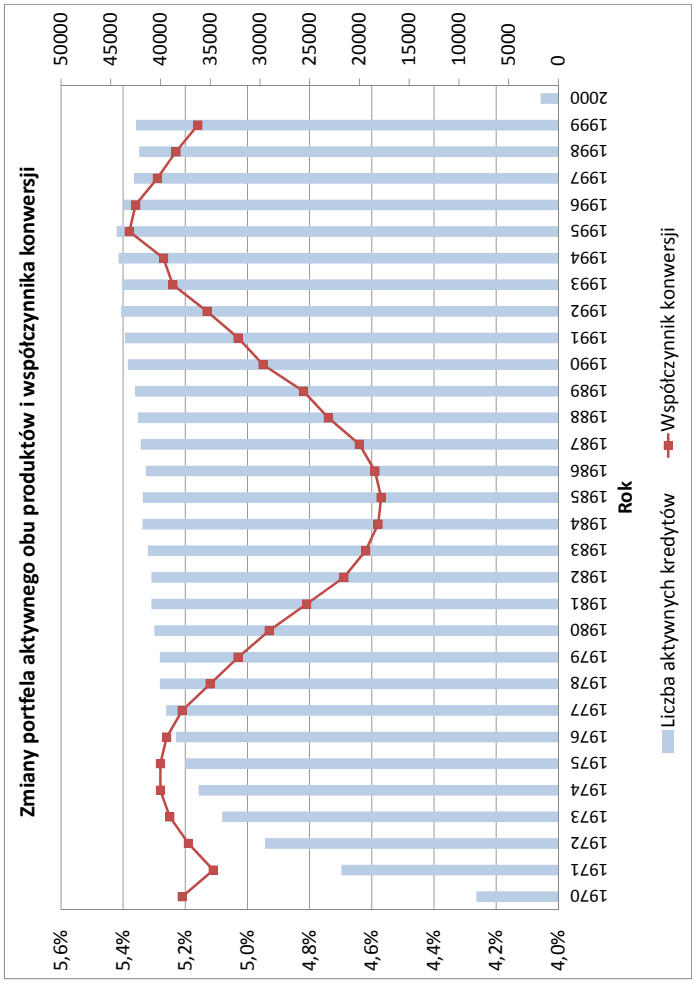
Źródło: opracowanie własne.

Rysunek 18. Kredyt gotówkowy



Źródło: opracowanie własne.

Rysunek 19. Portfele miesięczne



Źródło: opracowanie własne.

rię do momentu złożenia 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 ma albo większą, albo mniejszą wiedzę o kliencie. Pojawia się tu bardzo ciekawy problem, czy lepiej zaakceptować klienta z jego ryzykownym kredytem 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 w Polsce jest właśnie rozwijane i utrzymywane przez Biuro Informacji Kredytowej (BIK). Im większą wiedzę mamy 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 sytuacja: część klientów spłaca kredyty dobrze w jednym banku, jednocześnie w drugim mając opóźnienia, możemy to obserwować właśnie na podstawie danych BIK. Nawet jeśli opisany model zachowania nie jest doskonały, to warto 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, gdy bank, dysponując kapitałem początkowym, akceptuje początkowe straty, aby nauczyć się w przyszłości optymalizować procesy. Niestety obecnie banki zabezpieczają się 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 okres, kiedy procesy już w miarę się ustabilizowały, ale ryzyko w tym czasie było większe

od ryzyka w latach 1988–1998, gdy korzystaliśmy już z modeli, mając nadzieję otrzymania istotnych zysków. Zostały zbudowane cztery przykładowe modele akceptacyjne (obliczane na czas wniosku kredytowego). Trzy modele prognozujące zdarzenie $default_{12}$: 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 6 miesięcy obserwacji). 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 na poziomie 45% dla kredytu ratalnego i 55% dla kredytu gotówkowego. Dodatkowo wszystkie kredyty gotówkowe zostały wygenerowane ze stałą kwotą kredytu – 5000 PLN oraz okresem kredytowania – 24 miesiące. Wskaźniki finansowe procesu dla okresu modelowego 1975–1987 przedstawiono w tabeli 12 (metoda wyliczania oparta na wzorze 3.1, str. 53). W tabeli 13 przedstawiono moce modeli predykcyjnych. Są one nawet za duże jak na możliwości rzeczywistych procesów, szczególnie dla modelu prognozującego zdarzenie *response* (PR Css), gdyż w rzeczywistości uzyskuje się modele z mocą mniejszą niż 80%. Jednocześnie jednak 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 (w tym wypadku ujemny wynik) to około -40 mln PLN. Nie lada wyzwaniem staje się doprowadzenie do opłacalności tego procesu. Problem jest na tyle poważny, że niniejsze opracowanie daje tylko pewne propozycje, wskazując istotną rolę danych losowych w dalszych badaniach naukowych. Obecnie coraz bardziej

rozwijają się zagadnienia opłacalności, nazywane modelami wyceny wartości życiowej klientów, ang. *Customer Life Time Value* – CLTV lub CLV (Ogden, 2009; DeBonis et al., 2002). W przypadku naszego procesu uproszczoną wersją modelu CLTV jest model Cross PR Css.

Tabela 12. 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 13. 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.

Przejdźmy zatem przez etapy wyznaczania optymalnych 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. W kontekscie analiz CLTV powinno się do zagadnienia podejść bardziej dokładnie i rozważyć szereg 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 to, 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 na gotówkowy. 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 proces akceptacji 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 ocen punktowych 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 14). Analizując segmenty, tworzymy kolejne reguły: dla kredytu ratalnego, jeżeli $PD\_Ins > 8{,}19\%$, to odrzucamy, a jeżeli $8{,}19\% \geqslant PD\_Ins > 2{,}18\%$ i ($PR\_Css < 2{,}8\%$ lub $Cross\_PD\_Css > 27{,}24\%$), to też odrzucamy.

Zwróćmy uwagę na fakt, że została wprowadzona dodatkowa reguła nie tylko oparta na mierniku ryzyka, ale także mówiąca: jeśli ryzyko kredytu ratalnego jest w górnej półce, to, aby się opłacało, musimy mieć gwarancję wzięcia przyszłego kredytu gotówkowego na ustalonym poziomie prawdopodobieństwa oraz tego, że będzie on na odpowiednio niskim poziomie ryzyka.

Tak ustalone 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 akceptowalibyśmy wszystkie kredyty ratalne spełniające tylko jedną regułę – $PD\_Ins \leqslant 8{,}19\%$ – i proces przyniósłby zysk $1\,212\,261$ PLN. Stracilibyśmy zatem około 470 tys. PLN, co dawałoby o prawie 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.

Tabela 14. 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.

Testowanie strategii akceptacji

W celu głębszego zrozumienia problemów związanych z wdrażaniem modeli skoringowych rozważmy strategię decyzyjną przynoszącą zysk (tabela 15), która jest związana z metodą wyznaczenia optymalnych reguł, opisanych w niniejszym podrozdziale. 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 1 mln PLN, 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 braku aktywności klienta w momencie wnioskowania o kredyt gotówkowy (50% udziału). Ale i tak możemy być zadowoleni: zamiast wyniku ujemnego -40 mln

PLN mamy zysk około 700 tys. PLN (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, poprawnie wykonane jego wszystkie składowe kroki zagwarantują nam osiągnięcie spodziewanych korzyści. Aczkolwiek do końca nie da się przewidzieć skutków gwałtownej, niemal rewolucyjnej zmiany strategii. Zauważmy, że zastosowaliśmy strategię przynoszącą zysk po zrealizowaniu strategii, w przypadku której akceptowaliśmy wszystkich klientów. Ze 100% akceptacji przeszliśmy na 26% akceptacji kredytu ratalnego i na 16% – gotówkowego. Tak duża zmiana całkowicie zaburzyła rozkłady naszych ocen punktowych, a tym samym i wartości prawdopodobieństw.

Zauważmy dodatkowo, że procent akceptacji kredytu gotówkowego 16,23% w rzeczywistości będzie miał inną wartość. Związane jest to z niemożliwymi do zaobserwowania liczbami, które nazwiemy niewidzialnymi. Liczbę klientów nieznanych możemy poznać 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: w jakim stopniu procent akceptacji kredytu ratalnego wpływa na procent akceptacji kredytów gotówkowych? Jest to już dość trudne do wykazania na bazie danych symulacyjnych. Z reguły całkiem dużą skłonność do zaciągania kredytów można zaobserwować wśród klientów, którzy przy kredycie ratalnym znajdują się 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 dojść jeszcze do innej sytuacji. Sam procent akceptacji kredytów gotówkowych nadal będzie podobny, ale liczba akceptowanych kredytów gotówkowych może znacząco się zmniejszyć. Trzeba zawsze mieć świadomość faktu, 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 zachowanie klienta wobec potencjalnych produktów sprzedaży krzyżowej.

W nowym procesie ten sam model w tym samym okresie pokazu-

je 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 ona z brakiem informacji w danych banku o wszystkich kredytach klientów. Ponieważ przy nowej strategii akceptuje się lepsze kredyty, bank ma informację tylko o lepszych kredytach 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 przynoszącej zysk próbować więcej akceptować), to musimy liczyć się z dość dużym błędem niedoszacowanego ryzyka. Zauważmy, że nie tylko rozkłady parametrów PD się zmieniły, ale także moce predykcyjne modeli. Statystyka Giniego modelu Cross PD Css z 74% spadła 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, oczekując nagrody. Zwróćmy uwagę na to, że w realnym świecie obserwujemy tylko liczbę $21{,}34\%$, a obserwacja tej drugiej $(40{,}72\%)$ jest możliwa tylko w badaniach na danych symulacyjnych. Można by pytać: 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 zbudowano by wówczas nowy model i pewnie wcale nie lepszy. Obserwowane pomiary po wdrożeniu zostają 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, 2014): świadomy menadżer podejmuje decyzje na podstawie danych oraz liczb zarówno obserwowanych, jak i ukrytych (niewidzialnych), nieobserwowanych. 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.

Wpływ 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), jest

tak bardzo nieprzewidywalny, jak duża jest zmiana przeprowadzana w procesie. Niestety nie ma idealnych metod pozwalających radzić sobie ze wspomnianym problemem. W książce *Credit Scoring w erze Big-Data* (Przanowski, 2014a) omawia się ten problem na wiele sposobów, pokazując nie tylko różne rozwiązania, lecz także niedoskonałości każdego z nich. Dodatkowo zaprezentowano wiele różnych strategii akceptacji. Okazuje się, że niektóre z nich przynoszą większe zyski w okresie kryzysu, a inne w czasie prosperity. Z prac prowadzonych ze studentami wynika, że temat doboru strategii jest bardzo istotny i nie jest możliwe podanie łatwego i jednoznacznego sposobu doboru najlepszej z nich. Co więcej, modele wartości życiowej klienta CLTV prognozujące faktyczną kwotę wszystkich przyszłych skumulowanych zysków okazują się najlepszymi narzędziami w definiowaniu strategii akceptacji. Opisanie metod budowy takich modeli i stworzenie przykładowej strategii wykraczają poza obszar tematyczny niniejszej pracy.

Tabela 15. Strategia przynosząca zysk

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.

4. Inne zastosowania Credit Scoringu

4.1. Optymalizacja kampanii reklamowych

Wykorzystanie modeli predykcyjnych, także modeli kart skoringowych, w zarządzaniu kampaniami reklamowymi stało się dziś standardem. Można śmiało twierdzić, że w instytucjach finansowych rozwijanie modeli predykcyjnych dotyczy głównie dwóch zespołów: jednego odpowiedzialnego za zarządzanie szeroko rozumianym ryzykiem; drugiego odpowiedzialnego za działania marketingowe czy sprzedażowe przede wszystkim związane ze spersonalizowanymi kampaniami reklamowymi (marketing bezpośredni, ang. *below the line* – BTL), czyli przygotowanymi dla konkretnego klienta, przez wysłanie listu zwykłą drogą pocztową, emaila, SMS-a lub wykonanie telefonu.

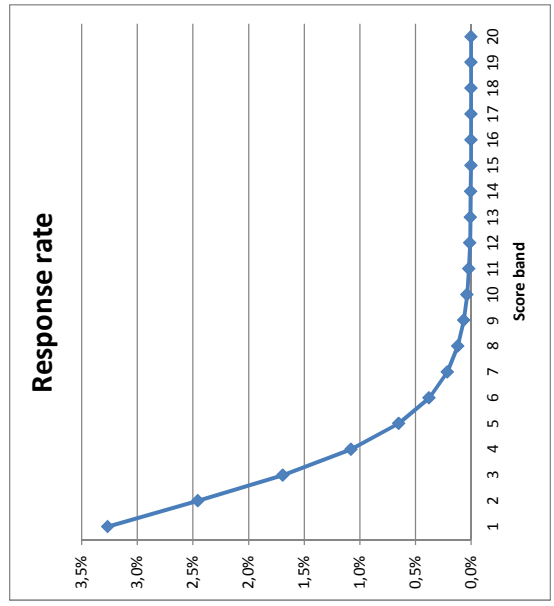
Koszty spersonalizowanych kampanii są zdecydowanie mniejsze od kosztów kampanii powszechnych (ang. *above the line* – ATL), głównie wykorzystujących kanały telewizyjne, radiowe i plakaty. Co więcej, pozwalają one na stworzenie specyficznej relacji z klientem, która daje długoterminowe korzyści i przede wszystkim powoduje, że klient ma wrażenie, że proponowany produkt jest właśnie dla niego. Klient nabiera do instytucji większego zaufania, powoli zaczyna utożsamiać propozycję z własnymi potrzebami, aż wreszcie staje się coraz bardziej podatny na tak dobrane reklamy. Proces nawiązywania tego typu relacji ogólnie jest zwany zarządzaniem relacjami z klientem (ang. *customer relationship management* – CRM) i jest już opisany w literaturze od 10 lat (Payne, 2005). Ważnym elementem tworzenia relacji jest ciągłe doskonalenie procesów docierania do klientów, przez rozwijanie zarówno kanałów dystrybucyjnych, jak i samych modeli predykcyjnych optymalizujących koszty kampanii. Im więcej danych o klientach ma dana instytucja, tym lepsze modele potrafi budować, a tym samym mniejsze koszty ponosi przy kolejnych kampaniach. Rozpoczyna się zatem wyścig dużych korporacji. Czas odgrywa tu ważną rolę: im szybciej zdobędzie się więk-

szy zasób precyzyjnych informacji o klientach, tym większą zyska się przewagę nad konkurencją.

Do lepszego zrozumienia przydatności modeli skoringowych w zarządzaniu kampaniami reklamowymi może posłużyć prosty przykład w arkuszu kalkulacyjnym. Rozważmy kampanię, którą potencjalnie można by wykonać aż dla 7 mln klientów (*number of cases*), patrz rysunek 20, oraz arkusz o nazwie *campaign_management.xlsx*. Przypuśćmy, że średni koszt dotarcia do jednego klienta kampanii BTL to 5 PLN (*average cost of contact, offer, campaign*). Dodatkowo na podstawie kampanii testowych najczęściej mamy informację, że średni procent odpowiedzi dotyczących pozytywnych reakcji na kampanię (ang. *response rate*) jest na poziomie $0{,}5\%$ (*global response rate*). Klient, który pozytywnie reaguje na naszą kampanię, kupując nasz produkt, przynosi średnio 800 PLN przychodu (*average income on responded case*). Jest to zatem przychód z pojedynczego respondenta, czyli osoby reagującej na kampanię. Jeśli zdecydujemy się wykonać kampanię dla wszystkich klientów (7 mln), to poniesiemy koszty w wysokości 35 mln PLN, nasi respondenci przyczynią się do przychodu 28 mln PLN, co sumarycznie przyniesie nam ujemny wynik w wysokości -7 mln PLN. Tak zdefiniowany model biznesowy nie przyniesie nam nigdy korzyści i narazimy instytucję finansową na duże straty.

Jeśli jednak posiadane dane o naszych klientach są wystarczająco długo gromadzone i reprezentują pożądaną jakość, to jest możliwe zbudowanie modelu predykcyjnego, który znacząco zmodyfikuje nasz proces biznesowy, powodując istotne zyski. Mianowicie, możliwe jest ustawienie naszych klientów w kolejności od tych, którzy najmniej reagują na kampanię, do tych, którzy reagują najbardziej. Ustala się wtedy właściwy punkt odcięcia i realizuje kampanię skierowaną do wybranej grupy docelowej. W naszym przypadku wysyłamy tylko 1750 tys. ofert, co powoduje znaczące zmniejszenie kosztów z 35 mln PLN do niecałych 9 mln PLN. Używany jest tu model z bardzo dużą mocą predykcyjną liczoną statystyką Giniego na poziomie 78%. Na jego podstawie potrafimy wyselekcjonować w ramach wysyłanych ofert aż 91% potencjalnych respondentów – patrz statystyka *captured percent (gains)*. Inną statystyką używaną w tego typu procesach jest *lift*, równa 3,6, pozwala ona na oblicze-

Rysunek 20. Fragment arkusza kalkulacyjnego. Kalkulacja kampanii reklamowej



Number of cases	7 000 000
Average income on responded case	800
Average cost of contact, offer, campaign	5

Gini global
78,36%

Global response rate	0,5%
Accepted response rate	1,83%
Acceptance rate	25,00%
Cummulative lift on accepted	3,66
Captured percent (Gains)	91,43%
Global cost	35 000 000
Global income	28 000 000
Global profit	-7 000 000

Accepted cost	8 750 000
Accepted income	25 599 340
Accepted profit	16 849 340
Number of offers	1 750 000
Number of expected responders	31 999

Źródło: opracowanie własne.

Tabela 16. Fragment arkusza kalkulacyjnego. Zależności zysku z kampanii reklamowej od mocy predykcyjnej modelu

Accepted cost	b	a	Gini	Optimal profit	Optimal acc rate	Captured percent (Gains)
8 750 000	1	-0,6	78,36%	16 849 340	25,00%	91,43%
8 750 000	1	-0,6	78,36%	16 849 340	25,00%	91,43%
8 750 000	1	-0,5	73,81%	15 330 310	25,00%	86,00%
10 500 000	1	-0,4	67,06%	13 134 993	30,00%	84,41%
10 500 000	1	-0,3	56,38%	10 018 792	30,00%	73,28%
12 250 000	1	-0,28	53,55%	9 281 163	35,00%	76,90%
12 250 000	1	-0,26	50,43%	8 476 738	35,00%	74,02%
12 250 000	1	-0,24	47,01%	7 606 998	35,00%	70,92%
12 250 000	1	-0,23	45,18%	7 149 195	35,00%	69,28%
12 250 000	1	-0,22	43,28%	6 677 393	35,00%	67,60%
12 250 000	1	-0,2	39,24%	5 697 230	35,00%	64,10%
12 250 000	1	-0,18	34,91%	4 680 287	35,00%	60,47%
12 250 000	1	-0,172	33,11%	4 267 211	35,00%	58,99%
12 250 000	1	-0,17	32,66%	4 163 641	35,00%	58,62%
12 250 000	1	-0,16	30,35%	3 645 139	35,00%	56,77%
12 250 000	1	-0,15	28,01%	3 127 820	35,00%	54,92%

1%	of Gini	272 569	PLN
5%	of Gini	1 362 844	PLN
10%	of Gini	2 725 687	PLN

Źródło: opracowanie własne.

nie, ile razy lepiej nasz model od zwykłego modelu losowego selekcjonuje respondentów w grupie docelowej, innymi słowy, dzięki modelowi potrafimy prawie cztery razy skuteczniej trafiać do pożądanych klientów niż w sytuacji wysyłania ofert losowo. Pozytywna reakcja na kampanię w grupie docelowej wynosi prawie 2%, co właśnie oznacza, że prawie cztery razy lepiej identyfikujemy klientów, bo w całej populacji reakacja ta jest na poziomie $0{,}5\%$. Finalnie przychód z kampanii będzie tylko o $2{,}5$ mln PLN mniejszy od maksymalnego, ale całkowity wynik staje się dodatni i wynosi prawie 17 mln PLN.

Zamiast zatem tracić 7 mln PLN na każdej kampanii zarabiamy 17 mln PLN. Jeśli tych kampanii prowadzimy wiele w roku, to wnioski nasuwają się same – warto jest wykorzystywać modele predykcyjne i nasza instytucja jest w stanie wygrywać z konkurencją.

Może się jednak okazać, że przy obecnej wiedzy o procesach i danych o klientach nie jesteśmy w stanie zbudować modelu o mocy 78%. W dość łatwy sposób w symulacji można przeprowadzić wiele scenariuszy, analizując kolejno wzrastające wartości mocy predykcyjnej (patrz tabela 16). Widać wyraźnie, że nawet przy stosunkowo małej wartości statystyki Giniego, równej 28%, uzyskujemy już dodatni wynik na poziomie 3 mln PLN. Docieramy wtedy do prawie 55% respondentów. Dodatkowo można wyliczyć korzyść z lepszego modelu, poprawiając jego predykcyjność o 1%. Powiększa się wtedy zysk o dodatkowe prawie 300 tys. PLN.

Zwróćmy uwagę na to, że parametry modelu biznesowego muszą być odpowiednie, nie każda ich kombinacja finalnie przynosi dodatnie zyski. Problemy mogą być zarówno po stronie zbyt małej liczby klientów, niewłaściwych proporcji kwotowych pomiędzy kosztem dotarcia do klienta a przychodem z respondenta, jak i po stronie samego modelu czy jakości danych, na podstawie których model był budowany. Wreszcie może się okazać, że sam proces jest bardzo niestabilny. Ze względu na zmienny rynek lub zmienne upodobania klientów może stać się oczywiste, że danego produktu nie da się sprzedawać przez kampanie bezpośrednie. Opisane modele biznesowe najczęściej są stosowane w typowych powtarzalnych procesach sprzedaży krzyżowej (ang. *cross-sell – up-sell*).

Kolejna metoda jeszcze lepszego zmniejszania kosztów to modelowanie *uplift*, opisane w podrozdziale 3.4. Chodzi tu o oddzielenie takich klientów, którzy dobrowolnie będą korzystać z kolejnych produktów, od takich, do których musimy skierować kampanię, bo inaczej nie skorzystają z produktu. Mamy też sytuację odwrotną, gdy klienta możemy zniechęcić, jeśli dotrzemy do niego z daną kampanią reklamową. W obu tych przypadkach możemy dzięki takiemu modelowaniu zaoszczędzić niepotrzebne koszty. Tego typu działania powinny być elementem codziennego organizowania kampanii. Jednocześnie każda z nich powinna mieć swoją grupę kontrolną. Trzeba też starannie zadbać o to, by grupa kontrolna nie była przez pomyłkę grupą docelową innej kampanii. Co więcej, trzeba też umieć wyznaczyć limity kontaktów z klientem. To zagadnienie jest bardzo ciekawe i trudne. Możemy mówić o limicie indywidualnym dla klienta i kanału dystrybucyjnego. W praktyce nie prowadzimy tylko jednej kampanii i nie jest łatwo organizować proces zarządzania kampaniami w przypadku przedsiębiorstwa mającego miliony klientów. W takim układzie procesy te muszą stać się automatyczne i wykonywane przez jeden dedykowany system. Narzędzie tak organizowanych kampanii jest jednak bardziej sztuką intelektualną niż zagadnieniem związanym z jakimś oprogramowaniem. Możemy je nazwać angielskim pojęciem *Multichannel Campaign Management* (MCCM), wprowadzonym przez Gartnera w 2013 roku (Sarner i Hopkins, 2013), które można przetłumaczyć jako wielokanałowe zarządzanie kampaniami reklamowymi. Od pewnego czasu nazywa się je także *Omnichannel*, by podkreślić spójność ofert w środowisku wielokanałowym. Istotą takiego podejścia jest tworzenie i gromadzenie wszelkich pasywnych kampanii reklamowych (inaczej ofert) w jednym systemie na poziomie klienta. Następnie przez w pełni kontrolowane i automatyczne procesy generuje się kampanie aktywne dzięki zaawansowanym regułom analitycznym i statystycznym, wybierając najlepsze kampanie dla danego klienta, dedykowane do właściwych kanałów dystrybucji i regulowane także ustalonymi limitami kontaktów. W procesie tym są podejmowane wciąż decyzje wyboru z wielu możliwych ofert najlepszej kampanii aktywnej dla danego klienta. Przy podejmowaniu decyzji mogą znacząco pomóc nie tylko modele *response*, ale także LTV czy CLTV (*Customer Life*

Time Value), wspomniane w podrozdziale 1.2. Bardzo ważnym tematem w zarządzaniu kampaniami jest podejście klienckie, a nie produktowe. Dziś często w przedsiębiorstwach spotyka się typową hierarchię stanowisk, opartą na oferowanych produktach. Powoduje to, że osoba odpowiedzialna za produkt w naturalny sposób, ze względu na swój produktowy cel, będzie dążyła do tego, by kampanie dotyczące jej produktu były najliczniejsze. Z punktu widzenia klienta, a także w kontekście globalnego zysku przedsiębiorstwa może to być jednak zupełnie niewłaściwe podejście. Po pierwsze może się okazać, że daje się zbudować segmentację klientów określającą kombinacje produktowe najlepsze dla danego segmentu. Po drugie niektóre produkty mogą być mniej opłacalne w krótkim horyzoncie czasowym, a inne w długim. Należy zatem używać zaawansowanych narzędzi analitycznych, mierząc najlepiej jak się da potencjał klienta i dobierając sekwencje produktowe w taki sposób, by maksymalizować zyski. Dzieje się to zazwyczaj, gdy budujemy długą relację z klientem i odpowiednio obciążamy go opłatami w czasie, niekoniecznie od razu dużymi na starcie. Fakt budowania długookresowych relacji staje się powodem powstawania zespołów CRM (ang. *customer relationship management*), które powinny być odseparowane od hierarchii produktowej w firmie. Wtedy mają szansę sprawować finalną kontrolę nad właściwym sposobem zarządzania kampaniami i nad metodami maksymalizacji zysku.

4.2. Utrzymanie odchodzących klientów

Jak wspomniano w podrozdziale 3.6, model biznesowy ACURA (ang. *acquisition, cross-sell, up-sell, retention, advocacy*) jest typowym przykładem modelu sprawdzającego się w instytucjach, w których mamy do czynienia z dużą liczbą klientów i pracuje się z procesami i zjawiskami masowymi. Dość powszechnie formułowane przez znawców CRM czy zarządzania marketingiem jest stwierdzenie, że pozyskanie nowego klienta, czyli akwizycja (*acquisition*), jest o wiele droższe od utrzymania obecnego klienta, którego można nadal pozyskiwać przez połączone działania *cross-sell*, *up-sell* i *retention*. Pojęcie retencji w zależności od sektora gospodarczego jest także związane z angielskimi pojęciami *churn* w telekomunikacji i *attri-*

tion w bankowości. Przyjmijmy, że odchodzącego klienta będziemy nazywać churnerem.

Rozważmy przypadek firmy telekomunikacyjnej, która obecnie ma 7 mln klientów (*number of customers in stock portfolio*) – patrz tabela 17. Niestety niemożliwe jest bardzo realne przedstawienie przypadku problemu odchodzenia klientów. Potrzebne są zbyt szczegółowe dane, które są objęte tajemnicą firmy. Na podstawie wzmianek prasowych dostępnych w Internecie oraz cenników reklam telewizyjnych udało się zebrać w miarę logicznie uzasadnione liczby, ale należy pamiętać, że są one przykładowe i służą tylko do przybliżenia tematu. Prawdziwa analiza procesu biznesowego jest możliwa tylko przy posiadaniu wszystkich rzeczywistych parametrów.

Przypuśćmy, że wszelkie obliczenia i prognozy odnoszą się do rocznego horyzontu czasowego. W ciągu roku przez kampanie telewizyjne pozyskujemy 90 tys. nowych klientów (*number of customers attracted by TV regular capmaign*) – patrz arkusze o nazwach *churn_simulation.xlsx* i *churn_simulation_gini.xlsx*. Koszt takich kampanii (ATL) wynosi 200 mln PLN – *TV regular campaign cost (ATL)*. Każdy klient przynosi średnio 800 PLN przychodu (*one year income generated by one customer*). Załóżmy, że udało nam się wyznaczyć średni poziom odejść klientów i jest on równy 130 tys. klientów (*number of churners*). Jeśli pozwolimy tym klientom odchodzić, to z roku na rok będziemy tracić około 40 tys. osób. Pierwszym pomysłem może być pogodzenie się z pewną liczbą odejść i poniesienie kosztu kampanii ATL, aby zrównać poziom odchodzących i nowych klientów. Wtedy dodatkowo wydamy jeszcze około 88 mln PLN – *extra cost of TV to keep churners to have constant stock (ATL)*. Obliczając średni przychód z pozyskanych w ten sposób klientów, dochodzimy do ujemnego wyniku na poziomie $-16{,}8$ mln PLN. Model polegający na nieustającym pozyskiwaniu klientów poprzez dość proste metody ATL niestety jest bardzo kosztowny. Okazuje się jednak, że można ponieść pewien dodatkowy koszt utrzymania klientów. Przypuśćmy, że tworzymy usługę (atrakcyjną opcję w abonamencie), za którą dodatkowo zapłacimy naszym klientom. Koszt takiej opcji dla jednego klienta to 90 PLN, można by go rozłożyć na faktyczną kwotę dotyczącą opcji i koszt dotarcia do klienta, ale pierwsza z nich jest najczęściej największa. Niestety nie możemy

Tabela 17. Fragment arkusza kalkulacyjnego. Model odchodzenia klientów

One year time horizon		Percent
Number of customers in stock portfolio	7 000 000	100,00%
Number of customers attracted by TV regular capmaign	90 000	1,29%
Number of churners	130 000	1,86%

TV regular campaign cost (ATL)	200 000 000
TV cost per customers	2 222
One year income generated by one customer	800
Churn campaign cost per customer (non-tv BTL)	90

Extra cost of TV to keep churners to have constant stock (ATL)	88 888 889
Profit of extra TV anti-churn program	-16 888 889

Loss income on churners	104 000 000
Total extra TV cost of anti-churn program (ATL)	288 888 889
Total cost of anti-churn campaigns (non-tv BTL) only for churners	11 700 000

Theoretical profit of anti-churn campaigns when churners are known	92 300 000
Total profit of extra TV anti-churn program	-184 888 889

When churners are not known, and we need to predict and indicate them		
Cost of first 5% of all customers most likely to be churners	31 500 000	Gini
Percent of captured of all churners by the predictive model	33,69%	80,18%
Non-captured	66,31%	
Number of theoretical saved churners after BTL anti-churn campaigns	43 797	
Number of theoretical churners after BTL anti-churn campaigns	86 203	
Profit of BTL anti-churn campaigns	3 537 600	

Źródło: opracowanie własne.

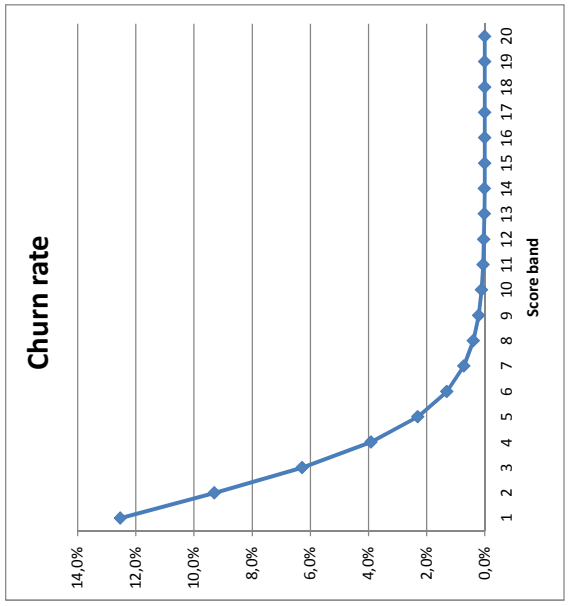
tej opcji sprzedać każdemu z 7 mln klientów, byłoby to zbyt kosztowne. Musimy jakoś zidentyfikować potencjalnych odchodzących klientów. Przypuśćmy, że ich znamy, jesteśmy w stanie wprowadzić taki proces, w którym każdego klienta wyrażającego wolę odejścia dzięki dodatkowej opcji motywujemy do pozostania na kolejny rok u naszego operatora. W tym wypadku zysk ze stosowania kampanii bezpośredniej przeciwko odchodzeniu (ang. *anti-churn* lub *churn detection*) wynosiłby 92 mln PLN i jednocześnie utrzymalibyśmy wszystkich klientów. Ich liczba w naszym portfelu rosłaby zatem z roku na rok.

Niestety nie zawsze istnieje możliwość łatwej identyfikacji klientów odchodzących. Kontakt z klientem przez Call Center, gdy oznajmia, że chce rozwiązać umowę, może być już zbyt późnym zdarzeniem, by klienta odwieźć od jego zamiarów. Trzeba zatem reagować szybciej, zanim klient pomyśli o odejściu. Istnieje taka możliwość przez budowę właściwego modelu predykcyjnego. Jego budowa pozwala zarówno lepiej zarządzać procesem finansowym odejść, jak i często zrozumieć ich przyczyny, co w dalszej perspektywie może pomagać w ulepszaniu procesów, powodując zmniejszenie się liczby zdarzeń niekorzystnych dla klientów, osłabiających wizerunek operatora telekomunikacyjnego. Trzeba także w tym przypadku tworzyć i analizować grupy kontrolne, w których przypadku pozwolimy odchodzić klientom, nie wszczynając żadnych akcji utrzymaniowych. Pozornie będziemy tracić przychody i samych klientów, ale za to będziemy gromadzić cenne dane, które pozwolą nam budować modele i identyfikować przyczyny odejść. Najważniejsze są właściwe proporcje. Pozwolimy, dla przykładu, odejść tysiącom klientów, ale zatrzymamy setki tysięcy.

Przypuśćmy, że potrafimy zbudować bardzo dobry model z mocą predykcyjną 80%. Tak duża moc rozróżniania klientów pozwala zgromadzić w pierwszym 5-procentowym percentylu (5%) prawie aż 34% wszystkich churnerów (*percent of captured of all churners by the predictive model*). Innymi słowy, decydujemy się na wysłanie ofert dodatkowej opcji tylko do 5% z 7 mln klientów, czyli do 350 tys. (patrz rysunek 21).

Możemy przypuścić, że klient odchodzący po otrzymaniu propozycji dodatkowej opcji pozostanie z nami, czyli stanie się nieod-

Rysunek 21. Fragment arkusza kalkulacyjnego. Kalkulacja modelu odchodzenia klientów



Number of cases	7 000 000
Average income on non-chuner	800
Average cost of contact, offer, campaign	90

Gini global	80,18%

Global churn rate	1,86%
Accepted churn rate	12,53%
Acceptance rate	5,00%
Captured percent (Gains)	33,69%
Global cost	630 000 000
Global income	104 160 000
Global profit	-525 840 000

Accepted cost	31 500 000
Accepted income	35 090 833
Accepted profit	3 590 833
Number of offers	350 000
Number of captured churners	43 864
Number of non-captured churners	86 336

Źródło: opracowanie własne.

chodzący. Oznacza to, że kierując kampanię do wybranego precentylu, utrzymamy aż 44 tys. klientów z dodatnim wynikiem w wysokości $3{,}5$ mln PLN. Wykorzystując model predykcyjny, jesteśmy zatem w stanie utrzymać na tyle dużą liczbę klientów, by uzyskać bilans odejść i nowych klientów z tendencją wzrostu liczby wszystkich klientów oraz jednocześnie uzyskiwać rocznie kilka milionów złotych dodatkowo. W praktyce pojawia się jeszcze jeden parametr. Otóż nie każdy klient po otrzymaniu oferty opcji decyduje się na pozostanie w relacji z naszą firmą. Innymi słowy, ponosimy wtedy większe koszty i musimy zwiększyć liczbę wysyłanych ofert, by utrzymywać oczekiwany bilans odejść.

4.3. Pozostałe przykłady zastosowań bez szczegółowych analiz finansowych

Zastosowanie modeli predykcyjnych jest coraz to szersze. Jedynymi barierami ich zastosowań może być ludzka wyobraźnia i brak posiadania odpowiednich danych. Innym problemem jest także zdobycie danych do wykonania testów procesu biznesowego.

Nawet jeśli jeszcze nie gromadzimy właściwych danych, to i tak można już przeprowadzać testy, chociaż w niektórych przypadkach to także jest niemożliwe. W medycynie czy biostatystyce nie zawsze można (lub nawet nie powinno się) przeprowadzać kalkulacje finansowe. Przypuśćmy np., że rozważamy przypadek zachorowalności na raka. Okazuje się, że można z dość dużą mocą predykcyjną (prawie 80% Giniego) przybliżać prawdopodobieństwo śmierci lub zgonu pacjenta ze względu na pierwsze objawy lub pierwszą diagnozę zachorowania na raka (Delen et al., 2005). Modele te są oczywiście oparte na głębokiej diagnozie medycznej. Czy powinno się zatem wyliczać zysk z ocalenia lub śmierci pacjenta? Nawet nie powinno się takiego pytania stawiać. Można natomiast wykorzystywać modele predykcyjne do profilaktyki, gdyż nie tylko prognozują one przyszłą śmierć pacjenta, ale także wyjaśniają tego przyczyny. Można też stosować modele do wyboru terapii – skoro potrafimy prognozować zgon lub przeżycie, to można to uszczegóławiać i pytać o przeżycie pod warunkiem przejścia danej terapii. Można też w ostateczności

przy ograniczonym budżecie służby zdrowia, selekcjonować pacjentów do poddania ich właściwej terapii, która jest bardzo kosztowna.

Inne zastosowania, szczególnie wykorzystywane w telekomunikacji, ubezpieczeniach i bankowości, to wykrywanie nadużyć (ang. *fraud detection*), gdy główną ideą jest identyfikacja klientów pod kątem jakiegoś rodzaju oszustwa, wyłudzenia lub kradzieży.

Szczególny rodzaj działalności przestępczej jest znany pod specjalną nazwą „prania brudnych pieniędzy" (ang. *anti money laundering* – AML). Do przeciwdziałania tej działalności są zobowiązane w związku z tym i monitorowane wszystkie banki pod karą dużej grzywny, wyznaczanej przez Główny Inspektorat Finansowy (GIF).

Nie wolno zapomnieć o metodach skoringowych stosowanych także w bankowości wobec portfeli korporacyjnych i w szczególności małych i średnich przedsiębiorstw (ang. *small medium enterprises* – SME), w których przede wszystkim są znane zastosowania modelu Z-score (Altman, 1968). Wszystkie metody, analizy i strategie procesu akceptacji przedstawione w rozdziale 3 można by powtórzyć, zastępując jedynie słowo „klient" słowem „przedsiębiorstwo". Oczywiście modele są budowane na podstawie zupełnie innych danych i wskaźników, ale problemy takie jak opłacalność procesu akceptacji, obciążenie próby i problem wniosków odrzuconych oraz inna wartość mocy predykcyjnej po zastosowaniu modelu w systemie decyzyjnym są także tu obecne i stanowią podstawowe zagadnienie optymalizacji zarządzania portfelem SME.

Podsumowanie

Nigdy w historii ludzkości nie było czasów lepszych od obecnych, w których analizy danych tak bardzo są potrzebne i oczekiwane. Jest to złota era *Big Data*, gdy konferencji na ten temat pojawia się tak dużo, że prelegenci mają często kłopot ze znalezieniem czasu, aby wziąć w nich wszystkich udział. Internet jest przepełniony serwisami, blogami, masowymi emailami z tematyki *Business Intelligence*, *Business Analytics* czy *Big Data*. Prawie codziennie pojawiają się informacje o nowo powstałym ośrodku, centrum szkoleniowym czy edukacyjnym związanym z *Big Data*. Pobudza to wszystko do refleksji: skoro tak dużo się o tym mówi, to dlaczego tak mało w literaturze czy na konferencjach pojawia się konkretnych przykładów udowadniających użyteczność zaawansowanej analizy danych? Najczęściej podczas prezentacji podkreśla się walory danego rozwiązania technologicznego: że jest elastyczne, skuteczne, szybkie w obsłudze, konfigurowane w intuicyjnym interfejsie, metodami przeciągnij i upuść. Podkreśla się np. to, że obecne technologie potrafią liczyć prognozy dla milionów szeregów czasowych w czasie kilku godzin, ale już mało precyzyjnie wykazuje się ich trafność. Prezentuje się bardzo wygodne narzędzia wizualizacji danych, ale zbyt mało uwagi poświęca się interpretacji danych i wykazaniu użyteczności biznesowej. Można odnieść wrażenie, że zachwyt nad technologią uśpił naszą czujność; zamiast wykorzystywać potęgę obliczeniową do lepszego rozumienia biznesu i jego składowych procesów, kolekcjonujemy gadżety.

Temat książki i podstawowy jej cel – wykazanie przydatności zaawansowanej analizy danych na podstawie danych symulacyjnych – wydaje się bardzo ważny w obecnych czasach. Rozumowanie przedstawiane na kolejnych stronach książki uwypukla przesłanie, że użyteczność analizy danych jest problemem mentalnym, a nie technologicznym. Owszem bez obecnej techniki nie potrafilibyśmy wykonać tak dużej liczby obliczeń, ale najważniejsze pytanie brzmi: co liczyć i jak to wykorzystać w biznesie? Na tak postawione pytanie jest znacznie trudniej odpowiedzieć. Łatwiej jest zrobić prezentację, podczas której metodą klikania tworzy się model predykcyjny pro-

gnozujący przypadki odejścia klientów. Trudniej udowodnić, że ów model naprawdę przynosi korzyści w firmie. Jeszcze trudniej wykazać konkretne kwoty zysku, przychodów czy strat. Coraz częściej dochodzi się do wniosku, że zarówno środowisko naukowe, jak i środowisko konsultingowe nie są gotowe do rzetelnego udowadniania korzyści ze stosowania zaawansowanej analizy danych. Owszem tradycyjną metodą firm konsultingowych jest tworzenie projektów PoC (ang. *proof of concept*), gdy w konkretnej firmie, przy konkretnych uwarunkowaniach, udaje się wykazać wspomnianą przydatność. Ale wnioski z realizacji takiego projektu z reguły są chronione, czyli nie można ich prezentować podczas działań przedsprzedażowych (ang. *presale*), a tym bardziej podczas wykładów dla studentów. Mamy zatem poważną lukę w procesie kształcenia. Pojawiają się grupy tzw. ekspertów, którym dopisało szczęście i znaleźli się akurat w takich okolicznościach, że mieli okazję zetknąć się z rzeczywistymi danymi, co mylnie utożsamiają z doświadczeniem z zakresu know-how. Problem więc w tym, że okoliczności te pojawiają się rzadko i z reguły nikt nie jest dobrze przygotowany do tego, by je w pełni wykorzystać. Obowiązkiem nauczyciela akademickiego czy też naukowca jest działanie upowszechniające wiedzę i naukę, a w tej dziedzinie oznacza to tworzenie różnorodnych symulatorów umożliwiających start wszystkim zdolnym studentom.

Treść kolejnych czterech rozdziałów opracowania jednoznacznie wskazuje, że sformułowane cele zostały zrealizowane. Jest możliwe wykorzystanie danych symulacyjnych w badaniach Credit Scoring i zaprezentowanie ich użyteczności w optymalizacji procesów biznesowych.

Modele skoringowe pozwalają w procesie akceptacji kredytowej osiągać miesięcznie milionowe zyski, jeśli tylko mamy do czynienia z masowymi procesami, czyli odpowiednio dużą liczbą klientów wnioskujących miesięcznie o nowy kredyt, i jeśli także poprawnie określi się wszystkie parametry automatycznego procesu.

Wykorzystanie modeli skoringowych jest bardzo szerokie i to nie tylko w bankowości. Można je stosować zarówno w procesach zarządzania kampaniami reklamowymi, jak i w utrzymaniu odchodzących klientów. Można na ich podstawie powiększać zyski o milionowe kwoty.

Wreszcie dzięki danym symulacyjnym możliwe jest budowanie uproszczonych modeli finansowych procesów biznesowych, spełniających istotną rolę edukacyjną i naukową oraz pozwalających uświadomić sobie istotne elementy procesu.

Podstawowy pomysł przedstawiony w książce polega na uproszczeniu postaci modelu i prezentowaniu go jako narzędzia podziału na 20 równolicznych grup klientów ze zróżnicowanymi wartościami wskaźnika biznesowego takiego jak *response rate* lub *bad rate* dla każdej z grup. Innymi słowy, model predykcyjny oznacza tworzenie grup wartości statystyki interpretowanej biznesowo i porządkowanie ich od jej najmniejszej wartości do największej. Tylko taka własność modelu jest potrzebna w analizowaniu jego biznesowej użyteczności. Jednocześnie taki sposób prezentacji modelu jest banalnie prosty i nie wymaga rozumienia zaawansowanych metod budowy modeli statystycznych. Staje się zatem narzędziem komunikacji inżyniera danych z przedstawicielami biznesu.

Tego typu sposoby przedstawiania rzeczy trudnych powinny na stałe zagościć wśród zagadnień poruszanych w trakcie szkoleń, konferencji i studiów podyplomowych dla szeroko rozumianego biznesu, by jak najlepiej i najpełniej przekonać środowisko, że w obecnych czasach przewagę konkurencyjną uzyska się przede wszystkim dzięki analizie już zgromadzonych danych i wykorzystaniu ich w usprawnianiu istniejących procesów, a następnie że właściwe analizy danych stanu aktualnego pomogą także zdefiniować nowe zadania, propozycje zmian procesów i właściwości nowych produktów.

Dodatek – lista arkuszy kalkulacyjnych

Istotną częścią książki są dołączone arkusze kalkulacyjne w programie Microsoft Excel:

- `acceptance_process_simulation.xlsx` – symulacja procesu akceptacji jednego produktu kredytowego (podrozdział 3.2),
- `gini_curves.xlsx` – obliczenia i wykresy najpopularniejszych statystyk i krzywych do mierzenia mocy predykcyjnej (podrozdział 3.3),
- `collection_amicable_simulation.xlsx`,

  `collection_amicable_simulation2.xlsx`,

  `collection_amicable_simulation3.xlsx` – symulacje procesu windykacji polubownej (podrozdział 3.4),
- `mortgage_simulation.xlsx` – symulacja procesu akceptacji kredytu hipotecznego (podrozdział 3.5),
- `campaign_management.xlsx` – symulacja zarządzania kampaniami reklamowymi (podrozdział 4.1),
- `churn_simulation.xlsx`,

  `churn_simulation_gini.xlsx` – symulacja procesu utrzymania odchodzących klientów (podrozdział 4.2).

Można je pobrać ze strony internetowej:
`https://ssl-administracja.sgh.waw.pl/pl/OW/publikacje/Strony/2015.aspx`.
Każdy arkusz kalkulacyjny znajduje się pod adresem:
`https://ssl-administracja.sgh.waw.pl/pl/OW/publikacje/Documents/`

Bibliografia

Altman E.I. (1968), *Financial ratios, discriminant analysis and the prediction of corporate bankruptcy*, Journal of Finance, 23, s. 589–609.

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. 1072–1081.

Banasik J., Crook J. (2007), *Reject inference, augmentation, and sample selection*, European Journal of Operational Research, 183(3), s. 1582–1594.

Bellotti T., Crook J. (2009), *Credit scoring with macroeconomic variables using survival analysis*, Journal of the Operational Research Society, 60, s. 1699–1707.

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

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, dostęp: 2014.02.19.

Blikle A.J. (2014), *Doktryna jakości – rzecz o skutecznym zarządzaniu*, Helion.

Crouhy M., Galai D., Mark R. (2006), *The Essentials of Risk Management*, McGraw-Hill.

Ćwik J., Koronacki J. (2005), *Statystyczne systemy uczące się*, Wydawnictwo Naukowo-Techniczne.

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.

Dobson A.J. (2002), *An Introduction to Generalized Linaer Models*, Chapman & Hall.

Engelmann B., Hayden E., Tasche D. (2003), *Testing rating accuracy*, Risk, 16, s. 82–86, http://www.risk.net/data/basel/pdf/basel_risk_jan03_2.pdf, dostęp: 2015.02.21.

Finlay S. (2010), *Credit Scoring, Response Modelling and Insurance Rating*, Palgrave Macmillan.

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

Hosmer D.W., Lemenshow S. (2000), *Applied Logistic Regression*, Wiley.

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.

Janc A., Kraska M. (2001), *Credit-scoring: nowoczesna metoda oceny zdolności kredytowej*, Biblioteka Menedżera i Bankowca „Zarządzanie i Finanse".

Kamiński B., Zawisza B. (2012), *Receptury w R*, Oficyna Wydawnicza SGH.

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.

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

Lo V.S.Y. (2002), *The true lift model – a novel data mining approach to response modeling in database marketing*, ACM SIGKDD Explorations Newsletter, 4(2), s. 78–86, http://www.sigkdd.org/sites/default/files/issues/4-2-2002-12/lo.pdf, dostęp: 2015.02.21.

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

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, 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.

Ostasiewicz W. (2012), *Myślenie statystyczne*, Wolters Kluwer Polska.

Payne A. (2002), *The value creation process in customer relationship management*, Cranfield University, White Paper.

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.

Provost F., Fawcett T. (2014), *Analiza danych w biznesie. Sztuka podejmowania skutecznych decyzji*, Helion.

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. (2014a), *Credit Scoring w erze Big-Data*, Oficyna Wydawnicza SGH.

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

Ptak-Chmielewska A. (2013), *Uogólnione modele liniowe*, Oficyna Wydawnicza SGH.

Radcliffe N.J., Surry P.D. (1999), *Differential response analysis: Modelling true response by isolating the effect of a single action*, Proceedings of Credit Scoring and Credit Control VI, Credit Research Centre, University of Edinburgh Management School.

Řezáč M., Řezáč F. (2011), *How to measure the quality of credit scoring models*, Czech Journal of Economics and Finance, 61 (5), s. 486–507, http://journal.fsv.cuni.cz/storage/1228_rezac.pdf, dostęp: 2015.02.21.

Sarner A., Hopkins J. (2013), *Magic quadrant for multichannel campaign management*, Raport techniczny, Technical report, Gartner, http://www.timetomarketcompany.com/magic-quadrant-for-multicha.pdf, dostęp: 2015.02.21.

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

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

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.

Spis rysunków

Spis tabel

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". Autor książki „Credit Scoring w erze Big Data" (OW SGH, 2014) i współautor podręcznika „Przetwarzanie danych w SAS" (OW SGH, 2013).

Odpowiedzialny w dużych bankach grup kapitałowych za budowanie, wdrażanie i monitoring modeli predykcyjnych, zarządzanie ryzykiem kredytowym, tworzenie zautomatyzowanych procesów Customer Relationship Management (CRM), zarządzanie kampaniami i ofertami w tym MultiChannel Campaign Management (MCCM), tworzenie automatycznych procesów budżetowania i planowania.

Publikacja stanowi cenne pogłębienie problematyki związanej z credit scoringiem. Autor podejmuje istotny temat, związany z możliwością wykorzystania modeli predykcyjnych poza bankowością, do optymalizowania decyzji finansowych w samych przedsiębiorstwach, czyli o możliwość szerszego zastosowania tych modeli, z uwzględnieniem ograniczeń, jakie jednak również w tym zakresie występują. Zwracając uwagę na stale zwiększającą się rolę baz danych, stanowiących podstawę do podejmowania racjonalnych decyzji finansowych, należy podkreślić (i czyni to autor), w warunkach niedostatku informacji, istotną rolę danych symulacyjnych jako podstawy do podejmowania takich decyzji. Biorąc pod uwagę ważność i atrakcyjność tematyki, należy uznać próbę jej wzbogacenia o dodatkowe wątki badawcze za niezmiernie wskazaną.

fragment recenzji prof. dr. hab. Jerzego Różańskiego, Uniwersytet Łódzki

OFICYNA WYDAWNICZA
SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE
www.wydawnictwo.sgh.waw.pl

ISBN 978-83-8030-051-4