



MODELOWANIE RYZYKA KREDYTOWEGO

Grzegorz Migut

StatSoft Polska Sp. z o.o.

Wprowadzenie

Ryzyko jest terminem bardzo często używanym, zarówno w języku potocznym, jak i w terminologii naukowej. Można je pokrótce scharakteryzować jako prawdopodobieństwo niepowodzenia naszych działań. Pojęcie ryzyka szczególnie mocno zaznacza się we wszelkiego typu działalności gospodarczej, gdzie zwykle podejmuje się decyzje w oparciu o niepełne informacje. Stąd umiejętność prawidłowej oceny ryzyka ma często decydujący wpływ na powodzenie danego przedsięwzięcia. Typowym przykładem instytucji na co dzień stykającej się z ryzykiem jest bank.

Studiując literaturę dotyczącą bankowości, można odnaleźć szereg prób podziału ryzyka bankowego pod względem źródła jego pochodzenia, przyjętego horyzontu czasowego itp. Wśród tych klasyfikacji występuje ryzyko kredytowe, zarządzanie którym uznawane jest zgodnie za kluczowy element działalności bankowej.

Pojęcie ryzyka kredytowego

Ryzyko kredytowe oznacza niebezpieczeństwo, iż kredytobiorca nie wypełni zobowiązań i warunków umowy, narażając kredytodawcę na powstanie straty finansowej [1]. Nawet najlepsze metody analityczne nie są w stanie w pełni wyeliminować strat związanych z ryzykiem, dlatego istotnym zadaniem dla banku jest umiejętne zarządzanie nim. W zależności od polityki banku i jego pozycji na rynku może on przyjąć odpowiednią strategię kredytową, np. udzielać kredytów mniej lub bardziej ostrożnie. Jednak niezależnie od przyjętej polityki, kluczowym zagadnieniem jest prawidłowa ocena poziomu tego ryzyka. Dlatego też banki podejmują szereg działań mających na celu skuteczne zarządzanie ryzykiem kredytowym. Działania te można podzielić na:

- ♦ działania banku poprzedzające decyzje o przyznaniu kredytu, tj. identyfikacja źródeł ryzyka i ich ocena,
- ♦ działania, jakie podejmuje bank w związku z już przyznanym kredytem, są to m.in. wszelkiego rodzaju działania monitorujące oraz tworzenie funduszy celowych na pokrycie ewentualnych strat związanych z niespłaceniem kredytu.



W niniejszej pracy główny nacisk został położony na możliwości zarządzania ryzykiem kredytowym odnośnie działań poprzedzających przyznanie kredytu. Tego typu działania nazywamy **oceną zdolności kredytowej**. Przez termin ten możemy rozumieć zdolność do terminowego i kompletnego wypełniania zobowiązań oraz warunków umowy kredytowej. Ocena zdolności kredytowej klienta ubiegającego się o kredyt jest zawsze badana w dwóch wymiarach:

- ◆ pod względem formalno-prawnym,
- ◆ pod względem merytorycznym.

W pierwszym etapie oceny zdolności kredytowej klienta analizuje się jego zdolność formalno-prawną do podejmowania zobowiązań kredytowych. Celem tej analizy jest ustalenie prawnej zdolności wnioskodawcy do zaciągania zobowiązań, posiadania przez niego prawnych możliwości zabezpieczenia kredytu, posiadania wymaganych zezwoleń itp. Dopiero po stwierdzeniu tej zdolności przystępuje się do oceny zdolności kredytowej podmiotu pod względem merytorycznym. Zdolność ta jest z kolei analizowana w dwóch odrębnych aspektach [1]:

- ◆ personalnym (charakter, stan rodzinny, stan majątkowy, reputacja, kwalifikacje zawodowe),
- ◆ ekonomicznym (ocena finansów oraz zabezpieczeń).

Wymienione powyżej grupy nie są równoważne. W odniesieniu do kredytów udzielanych dla celów konsumpcyjnych dominuje aspekt personalny, natomiast w przypadku kredytów na działalność gospodarczą większy nacisk kładzie się na aspekty ekonomiczne.

W praktyce bankowej występuje zwyczaj dzielenia klientów na względnie jednorodne grupy i stosowanie wewnątrz tych grup podobnej polityki kredytowej. Może być to podział na klientów sprawdzonych i nowych, osoby fizyczne i podmioty gospodarcze, bądź według gałęzi przemysłu, jaką dana organizacja reprezentuje. Stosuje się też podział ze względu na typ kredytu, jego kwotę oraz okres spłaty.

Metody oceny zdolności kredytowej

W trakcie wieloletniej praktyki banki wypracowały szereg metod oceny zdolności kredytowej, które w skrócie można podzielić na metody [6]:

- ◆ Opisowe – (inaczej tradycyjne lub logiczno-dedukcyjne) polegające na ocenie zdolności kredytowej klienta na podstawie informacji o jego sytuacji ekonomiczno-finansowej.
- ◆ Statystyczno-matematyczne (empiryczno-dedukcyjne) – zdolność kredytową klienta określa się na podstawie cech i zachowań wcześniejszych kredytobiorców.

Podejście opisowe

W podejściu opisowym ocenia się szereg wskaźników uzyskanych na podstawie analizy ekonomicznej (bilansu, rachunku przepływów środków pieniężnych (*cash flow*), stanu



zadłużenia, zabezpieczeń itp.) i oceny personalnej. Wskaźniki te mają na celu informowanie o poszczególnych aspektach działalności wnioskodawcy. Często stosuje się porównania ze wskaźnikami uzyskiwanymi w danej branży, by bardziej zobiektywizować otrzymane wyniki. Pożądane jest również uzupełnianie analizy o prognozy dotyczące kształtowania się przyszłej kondycji firmy. Decyzje podejmowane w oparciu o tego typu oceny są w dużej mierze subiektywne, zależą od indywidualnej oceny przyznającego kredyt.

Często opisywanym w literaturze podejściem jest zasada 6C, tj. sześć kryteriów rozważanych podczas badania zdolność kredytowej [2]:

Character (charakter) - kompetentne zarządzanie i chęć spłaty kredytu,

Capacity (zdolność) - zdolność kredytobiorcy do spłaty kredytu,

Capital (kapitał) - relacja kapitału własnego lub majątku do udzielonego kredytu,

Collateral (zabezpieczenie) - jaka jest rynkowa wartość oferowanego zabezpieczenia,

Conditions (warunki) - zdolności wytwórcze, pozycja na rynku, konkurencja,

Confidence (pewność) - zdolność do zachowania działalności.

Pracownik banku wykorzystuje informacje dotyczące wymienionych charakterystyk potencjalnego kredytobiorcy i podejmuje decyzję - w dużej mierze subiektywną, jakkolwiek opartą na zestawie odpowiednich norm dla danej branży. Jak pokazały badania, tego rodzaju podejścia okazują się zbyt pesymistyczne w porównaniu z opisaną poniżej metodą punktową [3].

W oparciu o opisowy sposób oceny banki wprowadziły metodę punktową (*scoring*), polegającą na przypisaniu otrzymanej wartości wskaźnika pewnej oceny liczbowej, przy czym zakres wartości poszczególnych wskaźników może być różny, w zależności od jego wagi i indywidualnych preferencji banku. Następnie wykonuje się działanie polegające na sumowaniu ocen wskaźników. Otrzymana liczba zawiera się w pewnym ustalonym zakresie (np. 0-60) i określa oszacowany poziom ryzyka dla danego kredytobiorcy. W zależności od sumy punktów klient zostaje zaklasyfikowany do odpowiedniej grupy ryzyka ustalonej ogólnie (i po części subiektywnie) przez bank. Przykładowy podział na grupy ryzyka obrazuje poniższa tabela, opracowana na podstawie [6].

Liczba punktów	Zdolność kredytowa
60-40	niebudząca obaw
39-29	budząca obawy
28-17	zagrożona
16-0	utracona

Otrzymana suma jest podstawą do podjęcia decyzji o przyznaniu bądź nie przyznaniu kredytu (zdarza się jednak, że decydujący wpływ na przyznanie kredytu mają względy pozamerytoryczne). Metoda punktowa jest próbą wprowadzenia większej obiektywizacji w procedurę oceny zdolności kredytowej danego podmiotu. Jej przydatność uwidacznia się wszędzie tam, gdzie występuje duża liczba rutynowych decyzji.



Podjęcie statystyczno-matematyczne

Zupełnie innym (dla statystyka najbardziej interesującym) podejściem jest podejście empiryczno-dedukcyjne. Metody działające w oparciu o tę filozofię oceniają zdolność kredytową danego klienta na podstawie zachowań wcześniejszych kredytobiorców. Podejście to zakłada, że dany kredytobiorca będzie zachowywał się podobnie do historycznych kredytobiorców podobnych do niego. Zadaniem analityka stosującego tego typu metody jest odpowiedni dobór zmiennych opisujących zachowanie kredytobiorcy i zbudowanie na ich podstawie modelu, który potrafiłby rozpoznać, czy dany klient jest wiarygodny czy też nie. Model taki powinien mieć zdolność uogólnienia informacji zawartych w danych historycznych i działać z podobną skutecznością również dla nowych, nieznanymi sobie danych.

Należy pamiętać, że modele zbudowane w oparciu o podejście matematyczno-statystyczne najlepiej sprawdzają się w krótkich okresach czasu. Wraz z jego upływem zmieniają się zależności pomiędzy poszczególnymi parametrami, co może spowodować spadek zdolności prognostycznej modelu. Dlatego też należy w zadanych odstępach czasu reestymować go (czyli obliczać na nowo jego parametry) na podstawie nowych obserwacji.

Podobnego rodzaju modele mogą być budowane, by ocenić zachowanie klienta, któremu już przyznano kredyt. Mają one na celu rozpoznanie, czy pewna nieprawidłowość w spłacaniu kredytu przez klienta jest efektem jego problemów finansowych czy też wynika z przyczyn losowych. Tego typu analizy mają ogromne znaczenie ze względu na możliwość szybkiej reakcji banku w przypadku rozpoznania klienta z kłopotami, co zwiększa szansę na odzyskanie kredytu.

Modelowanie wiarygodności kredytowej

Proces konstruowania modelu składa się najczęściej z kilku etapów. Przed przystąpieniem do zasadniczej części analizy modelujący ma przed sobą dwa ważne zadania: wybór metody, przy pomocy której będzie budowany model, oraz przygotowanie danych, by mogły być użyte w analizie (tzw. preprocessing).

Jeśli chodzi o wybór metody analitycznej, to dostępnych jest szereg metod tradycyjnych, takich jak analiza dyskryminacyjna, analiza logitowa czy analiza probitowa. Dostępne są również metody nowsze, z których należy wymienić: sieci neuronowe, drzewa klasyfikacyjne oraz MARSplines. Metody starsze, wykorzystujące tradycyjne obliczenia statystyczne, stawiają przed badającym szereg wymagań odnośnie przygotowania danych, na przykład dane nie mogą być ze sobą nadmiernie skorelowane, wymagane jest spełnienie założeń o charakterze rozkładu. Metody nowsze uzyskują podobne lub lepsze wyniki, nie stawiając przy tym tak dużych wymagań w odniesieniu do danych. Wyboru metody możemy dokonać na samym początku procesu analizy lub też możemy skonstruować kilka różnych modeli i wybrać ten, który najlepiej spełnia nasze oczekiwania. W niniejszej pracy zaprezentowana zostanie przykładowa analiza za pomocą sieci neuronowych oraz drzew klasyfikacyjnych.



Sieć neuronowa jest narzędziem analizy danych, którego budowa i działanie zainspirowane zostało wynikami badań nad ludzkim mózgiem. Sieć składa się z

- ♦ wejść, gdzie wprowadzane zostają dane,
- ♦ warstw połączonych ze sobą neuronów, w których przebiega proces analizy,
- ♦ wyjścia, gdzie pojawia się sygnał będący wynikiem analizy.

Docierające do neuronów sygnały są w nich przekształcane przez odpowiednią funkcję. Ważnym elementem struktury sieci są wagi, osłabiające lub wzmacniające poszczególne sygnały docierające do neuronów. To właśnie od rodzajów funkcji oraz wag zależą wartości, jakie wygeneruje sieć na wyjściu. Na podstawie zbioru danych sieć uczy się rozpoznawać „złe” i „dobre” kredyty. Poprawnie nauczona sieć posiada umiejętność uogólnienia wiedzy zdobytej na podstawie historycznych obserwacji i dokonywania trafnych prognoz dla nowych danych. Dużą zaletą sieci neuronowych jest jej zdolność do radzenia sobie z modelowaniem zależności o charakterze nieliniowym, a taki właśnie charakter mają zależności opisujące zdolność kredytową. Pewną wadą sieci neuronowych jest działanie na zasadzie czarnej skrzynki: nie jesteśmy w stanie podać reguł i zasad, na podstawie których otrzymano dany wynik.

Kolejną metodą, jakiej możemy użyć, są **drzewa klasyfikacyjne**. Proces budowy drzewa opiera się na zasadzie rekurencyjnego podziału. Zasada ta polega na przeszukiwaniu w przestrzeni cech wszystkich możliwych podziałów zbioru danych na dwie części, tak by dwa otrzymane podzbiory maksymalnie się między sobą różniły ze względu na zmienną zależną (w naszym przypadku zmienną tą jest wiarygodność kredytowa). Podział ten jest kontynuowany, aż do całkowitego podziału przypadków na jednorodne grupy lub spełnienia ustalonych warunków zatrzymania. Reguły, względem których dokonano podziału przestrzeni cech, można w łatwy sposób przedstawić w formie drzewa. Tego typu grafy składają się z wierzchołków i krawędzi. Każdy wierzchołek reprezentuje decyzję o podziale zbioru obiektów na dwa podzbiory ze względu na jedną z cech objaśniających. Ważną zaletą drzew jest zrozumiałość dla człowieka sekwencja reguł decyzyjnych, pozwalająca klasyfikować nowe obiekty na podstawie wartości zmiennych. Atrakcyjną jest również możliwość graficznej prezentacji procesu klasyfikacji. Dodatkową zaletą drzew klasyfikacyjnych jest ich odporność na obserwacje odstające.

W obydwu metodach zalecane jest, aby badany zbiór obiektów podzielić na dwie części – zbiór uczący i zbiór testowy. Modele budowane są na podstawie informacji zawartych w zbiorze uczącym, a ich przydatność określana jest na podstawie zbioru testowego.

Przedstawiane w niniejszej pracy modele zbudowano w oparciu o dane dostępne na witrynie internetowej Uniwersytetu w Monachium (<http://www.stat.unimuenchen.de/service/datenarchiv/kredit/kredit.html>), przedstawiające zbiór 1000 historycznych obserwacji kredytobiorców indywidualnych. W obserwacjach tych wyszczególniono przedstawione poniżej zmienne:

- ♦ *decyzja* – określająca, czy dany klient spłacił kredyt czy nie.

Tę zmienną będziemy traktowali jako skategoryzowaną zmienną zależną.



Kolejne 20 zmiennych:

- ◆ *stan konta* - aktualny stan rachunku,
- ◆ *okres_k* - okres kredytu,
- ◆ *historia* - historia kredytowa klienta,
- ◆ *cel* - przeznaczenie kredytu,
- ◆ *kwota_k* - kwota kredytu,
- ◆ *suma_akt* - suma aktywów,
- ◆ *zatrudnienie* - czas pracy u obecnego pracodawcy,
- ◆ *rata* - wysokość raty,
- ◆ *stan* - stan cywilny,
- ◆ *gwaranci* - gwaranci lub inne osoby wspólnie zaciągające zobowiązanie,
- ◆ *zamieszkanie* - czas zamieszkania,
- ◆ *zabezpieczenie* - najbardziej wartościowe zabezpieczenie,
- ◆ *wiek_k* - wiek,
- ◆ *inne kredyty* - inne niespłacone kredyty,
- ◆ *mieszkanie* - mieszkanie wynajmowane, własnościowe lub inne,
- ◆ *ile kredytów* - liczba wcześniejszych kredytów,
- ◆ *stanowisko* - stanowisko pracy,
- ◆ *osoby* - liczba zaangażowanych osób,
- ◆ *telefon* - posiadanie telefonu,
- ◆ *obcokrajowiec* - informacja o pochodzeniu,

pełnić będzie w analizie rolę skategoryzowanych zmiennych niezależnych (predyktorów).

W zbiorze danych wyszczególniono również trzy zmienne ilościowe (liczbowe):

- ◆ *okres* - okres kredytowy w miesiącach,
- ◆ *kwota* - wysokość kredytu,
- ◆ *wiek* - wiek kredytobiorcy.

Należy zauważyć, że zmienne te zostały przedstawione również jako zmienne skategoryzowane (*okres_k*, *kwota_k*, *wiek_k*). Wszystkie analizy przeprowadzono w środowisku *STATISTICA Data Miner*.

Wstępna analiza danych

Należy pamiętać, że niezależnie od przyjętej metody analizy odpowiednia jakość danych jest kluczowym czynnikiem wpływającym na wyniki modelu. Właściwe przeprowadzenie wstępnej analizy danych jest niezbędnym warunkiem uzyskania pożądanego efektu końcowego, którym jest skonstruowanie modelu opisującego w poprawny sposób badany



fragment rzeczywistości. Znane jest powiedzenie: śmieci na wejściu - śmieci na wyjściu (*garbage in - garbage out*), oddające wiernie tę regułę.

Analizę rozpoczynamy od otwarcia w programie *STATISTICA* pliku *kredit.sta*, zawierającego dane kredytobiorców. Na początek warto sprawdzić, jaka jest **podaż i struktura danych**. Zbyt mała ich ilość może spowodować niewielkie zdolności predykcyjne modelu, spowodowane niewystarczającą ilością informacji zawartą w danych. Kolejnym ważnym czynnikiem jest względnie równy dobór obserwacji z poszczególnych grup ryzyka. Model uzyskany w oparciu o obserwacje, z których zdecydowana większość opisywać będzie sytuację prawidłowej spłaty kredytu, będzie miał tendencje do zbyt optymistycznego uznawania klientów za wiarygodnych kredytowo.

W naszym przypadku dysponujemy grupą tysiąca obserwacji, co wydaje się być liczbą wystarczającą do poprawnego przeprowadzenia analizy. By sprawdzić rozkład zmiennej *decyzja*, w opcji *Statystyki opisowe* uruchamiamy dla niej tabelę licznosci lub histogram. Na ich podstawie możemy stwierdzić, że dane zawierają 700 obserwacji, w których *decyzja* ma wartość TAK, a jedynie 300 przyjmuje wartość NIE. Ta dysproporcja może mieć istotny wpływ na jakość zbudowanego modelu, który przypuszczalnie często będzie się mylił, oceniając złe wnioski.

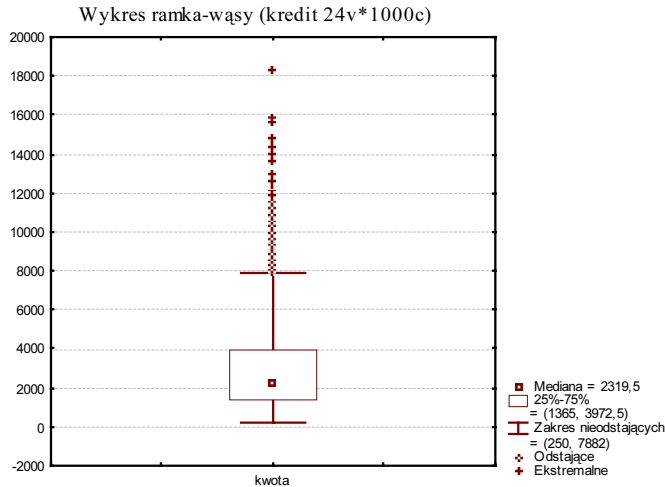
Kolejnym krokiem jest **analiza poprawności i jednorodności danych**. Zgromadzone przez nas dane nie mogą zawierać braków, należy zbadać występowanie obserwacji nietypowych lub błędnych. Podczas doboru danych do modelu należy zwrócić szczególną uwagę, by zawierały one informacje o klientach należących do jednorodnej grupy. Nie ma sensu budowanie łącznego modelu dla klientów indywidualnych i instytucji (choćbyż ze względu na różnice w parametrach oceny), jak również dla klientów o diametralnie różnej wysokości kredytu, ponieważ zachowania poszczególnych grup cechują odmienne zależności.

Jeśli nie mamy pewności, czy zgromadzone przez nas dane są kompletne, możemy uruchomić moduł *Replace missing data*, który zastąpi brakujące dane w wybrany przez nas sposób, np. zastępując je średnią, medianą, wartościami określonymi przez użytkownika lub usuwając przypadki, w których występują brakujące wartości.

Analizę danych odstających¹ spróbujemy prześledzić na przykładzie zmiennej *kwota*. W tym celu uruchamiamy wykres ramka-wąsy i jako zmienną zależną specyfikujemy zmienną liczbową *kwota*. Otrzymany wykres analizujemy pod kontem wartości odstających i ekstremalnych. Zostały one zaznaczone w formie kółek (odstające) i krzyżyków (ekstremalne).

Za odstające zostały uznane wnioski, w których wartość kredytu przekraczała kwotę 7882 DM. Warto rozważyć nieuwzględnianie tych obserwacji w dalszej analizie, ponieważ mogą one mieć negatywny wpływ na jakość modelu.

¹ Za wartości odstające i ekstremalne uznajemy takie wartości, które są oddalone od środka rozkładu w stopniu przekraczającym określoną regułę.



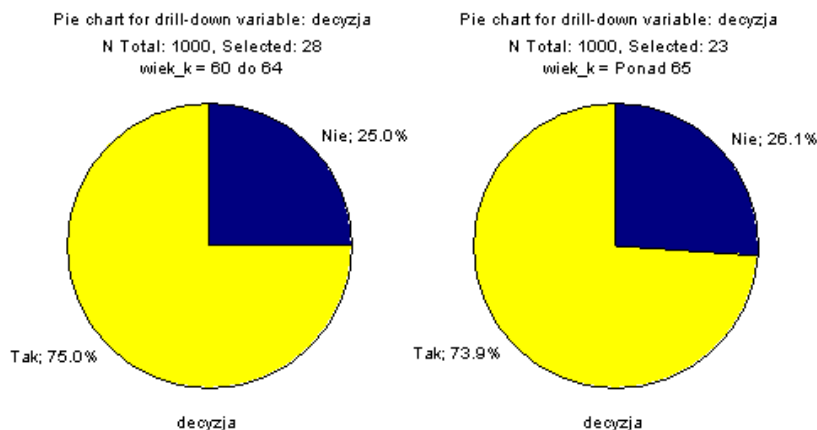
Format zgromadzonych przez nas danych często nie odpowiada wymogom zaplanowanej analizy. Dlatego też we wstępnej analizie danych stosuje się działania polegające na **przekształceniu danych** w zbiór nowych danych, spełniających określone założenia. Tego typu działania mają na celu poprawę jakości modelu oraz skrócenie czasu potrzebnego do analizy.

Jednym ze sposobów przekształcania danych jest normalizacja, polegająca na takim przekształceniu zmiennej, aby była porównywalna z jakimś ustalonym punktem odniesienia, co jest przydatne, gdy niekompatybilność pomiarów pomiędzy zmiennymi może mieć wpływ na wyniki analizy. Wynikiem normalizacji może być na przykład takie przekształcenie zmiennej, by jej wartości zawierały się w przedziale $[0,1]$. Tego typu przekształcenie jest przydatne zwłaszcza przy wykorzystaniu sieci neuronowych, dlatego też moduły sieci neuronowych zawarte w programach *STATISTICA* domyślnie przeprowadzają normalizację zmiennych wejściowych.

Inną metodą jest dyskretyzacja, polegająca na przekształceniu zmiennej liczbowej w zmienną skategoryzowaną. Podczas przeprowadzania tego typu operacji należy zwrócić uwagę, by licznosci w grupach powstałych w wyniku przekształcenia były jak największe. Nie jest prawidłowy podział dwustu historycznych klientów według wieku na 3 grupy, gdy do ostatniej grupy kwalifikuje się tylko jeden klient. Tego typu sytuacja może wpływać niekorzystnie na proces modelowania. Na przykład wszystkie obserwacje należące do danej grupy mogą znaleźć się jedynie w zbiorze testowym, nie uczestnicząc w procesie nauki, co powodować może znaczące błędy predykcji. Innym ważnym aspektem jest, by grupy powstałe w wyniku kategoryzacji były homogeniczne (jednorodne) ze względu na wartość zmiennej objaśnianej.

W naszym zbiorze danych dysponujemy grupą dwudziestu zmiennych skategoryzowanych. W celu przeanalizowania poprawności podziału zmiennych na kategorie dla wybranego arkusza uruchamiamy moduł statystyk opisowych (*Descriptive statistics*) i wybieramy jako

zmiennie grupę zmiennych skategoryzowanych (2-21). Obliczamy dla nich interesujące nas tabele licznosci oraz histogramy. Analizując licznosci w poszczególnych grupach, możemy stwierdzić, że niektóre licznosci są bardzo małe. Przeanalizujemy zmienną *wiek_k*. Dwie jej ostatnie grupy 60-64 lat oraz powyżej 65 lat posiadają niewielkie licznosci, warto zatem porównać procent poszczególnych decyzji w tych grupach i w wypadku podobnych proporcji, rozważyć możliwość połączenia ich w jedną grupę. W tym celu uruchamiamy moduł *Data Miner - Kostki, przekroje i drążenie danych* i wybieramy *Interakcyjne drążenie danych*. Jako zmienne określamy zmienną *decyzja* oraz *wiek_k*, a następnie przeprowadzamy operację *drill-down* dla interesujących nas grup wiekowych. Wyniki można zobrazować za pomocą wykresów kołowych.



Widzimy, że dla obydwu grup rozkład odpowiedzi jest bardzo zbliżony, rozsądne jest połączenie ich w jedną grupę.

Data Miner posiada dodatkowo wbudowane narzędzie umożliwiające automatyczny podział na jednorodne grupy. By je uruchomić, wybieramy moduł *Combining Groups for predictive Data Mining*, którego algorytm oparty jest na drzewach decyzyjnych CHAID. Funkcjonalność tego modułu zostanie zaprezentowana na przykładzie zmiennej *cel*, w której występuje kilka grup o niewielkiej licznosci, np. RTV, wakacje, biznes. Podczas uruchamiania modułu jako zmienną zależną ustalamy zmienną *decyzja*, jako zmienną skategoryzowaną zmienną *cel*, natomiast jako wyjście ustalamy zmienną, w której należy umieścić nowo pogrupowane zmienne. W wyniku przekształcenia zamiast 10 otrzymaliśmy 4 grupy o dostatecznej licznosci. Podobne działania możemy zastosować do pozostałych zmiennych wykorzystywanych w modelu.

Jeden z etapów wstępnej analizy danych ma na celu **określenie charakteru oraz dekompozycję danych**. Podstawowym celem tego etapu jest stwierdzenie, czy pomiędzy zestawem danych wejściowych oraz wartością wyjściową występuje zależność czy też związek pomiędzy tymi wartościami ma charakter przypadkowy [4]. Na samym początku badający powinien ze wszystkich dostępnych mu parametrów opisujących klientów wybrać te, które mogłyby mieć wpływ na wiarygodność kredytową. Przykładowo kolor oczu jest

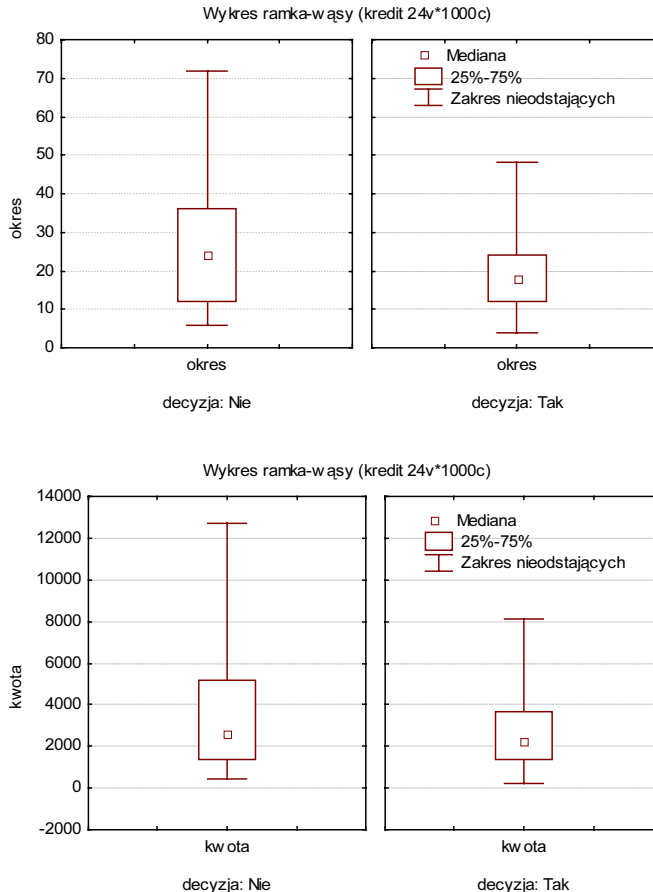


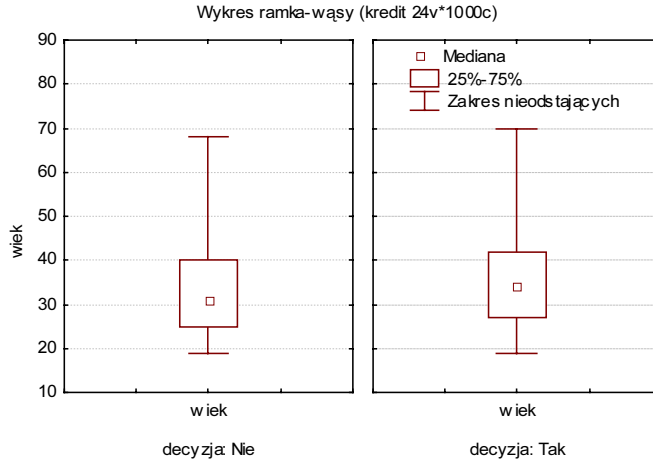
cechą różnicującą grupę kredytobiorców, trudno jednak uznać ją za istotną ze względu na wiarygodność kredytową. Selekcji należy dokonywać bardzo ostrożnie, by przypadkowo nie usunąć parametru istotnie wpływającego na ocenę klienta. Po wybraniu zmiennych mogących potencjalnie mieć wpływ na zmienną objaśnianą sprawdzamy, czy są one z nią w istotny sposób skorelowane. Brak skorelowania jest podstawą do nieuwzględniania zmiennej w modelu.

W celu zbadania związku predyktorów ciągłych i zmiennej zależnej *decyzja* skorzystamy ponownie z wykresu ramka-wąsy (*Categorized Box Plot*) oraz dodatkowo z modułu *Break-down and One-Way ANOVA*. Następnie zmieniamy parametry analiz:

- ♦ dla wykresu ramka-wąsy (*Categorized Box Plots*) ustawiamy jako wartość dla wąsy rozstęp po odrzuceniu punktów odstających ze współczynnikiem 1,
- ♦ dla *Breakdown and One-Way ANOVA* zmieniamy stopień szczegółowości na *Comprehensive*.

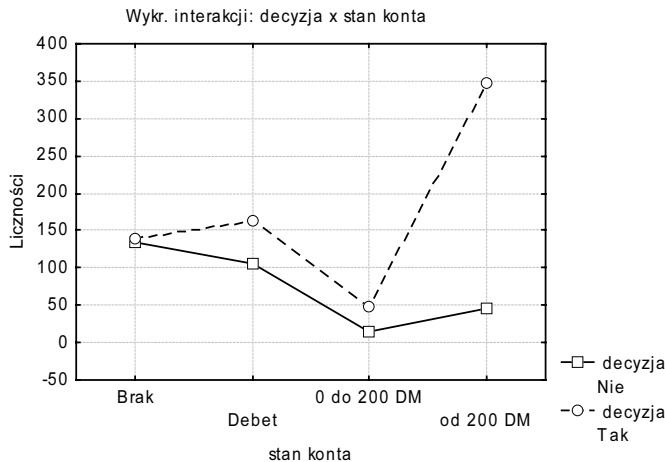
W wyniku otrzymujemy poniższe wykresy:



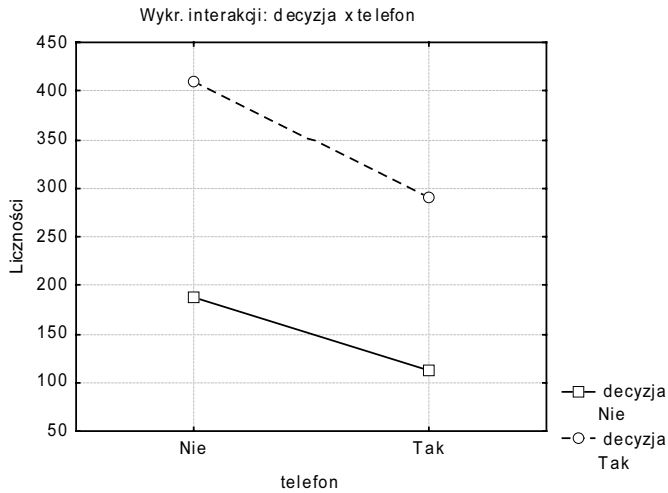


Analizując je, możemy stwierdzić, że wszystkie zmienne mają różne rozkłady w grupach z decyzją pozytywną i negatywną, chociaż mniej wyraźnie zarysowuje się ta zależność w przypadku zmiennej wiek. Widać także, że wraz ze wzrostem kwoty kredytu oraz okresu spłaty rośnie liczba decyzji negatywnych. Podczas analizy testu ANOVA możemy zauważyć, że wszystkie zmienne są w sposób istotny skorelowane z decyzją (we wszystkich przypadkach współczynnik p jest mniejszy od 0,05), dodatkowo istnieje wyraźna dodatnia korelacja pomiędzy zmienną okres a zmienną kwota.

By zbadać współzależności występujące wśród zmiennych skategoryzowanych, wykorzystamy tabele krzyżowe. Jako skategoryzowaną zmienną zależną wskazujemy zmienną decyzja, jako zmienne niezależne - pozostałe zmienne skategoryzowane. Dla wszystkich zmiennych włączamy rysowanie wykresu interakcji. Na wykresie tym przedstawiane są licznosci kategorii zmiennej zależnej w grupach zmiennej niezależnej. Jeśli linie łączące punkty są równoległe, można mówić o niewielkim związku pomiędzy zmiennymi.



W przypadku zmiennej *stan konta* można zaobserwować gwałtowny skok decyzji pozytywnych dla stanu konta powyżej 200 DM. Można więc w tym wypadku mówić o silnej zależności pomiędzy zmienną *decyzja* a stanem konta.



Dla zmiennej *telefon* linie te są praktycznie równoległe, co informuje nas o znikomym wpływie tej zmiennej na decyzję. Podczas budowania modelu możemy zrezygnować z tej zmiennej.

Tego typu analizy są konieczne w przypadkach, gdy stosujemy metody czule na występowanie w modelu zmiennych o znikomym stopniu korelacji ze zmienną zależną (w naszym przykładzie jest nią zmienna *decyzja*) lub występowanie dużej liczby kategorii w przypadku zmiennych skategoryzowanych. Sieci neuronowe oraz drzewa klasyfikacyjne należą do metod odpornych na te mankamenty. Dopuszczają również użycie zmiennych nadmiernie skorelowanych. Należy natomiast zadbać, by dla zmiennych skategoryzowanych, licznosci w poszczególnych grupach były odpowiednio duże.

Budowa modeli

Tę część analizy najwygodniej jest przeprowadzić w przestrzeni roboczej *Data Miner*. Umieszczamy w nim arkusz wejściowy i specyfikujemy zmienne zgodnie z zamieszczonym powyżej opisem danych. Kolejnym krokiem, jaki musimy wykonać, jest podzielenie zbioru danych na zbiory uczący i testowy. Na podstawie zbioru uczącego zostaną ustalone parametry modeli, ich weryfikacja przeprowadzona będzie na podstawie zbioru testowego. By podzielić zbiór danych, uruchamiamy węzeł *Split Input Data into Training and Testing Samples (Classification)*. Domyślnie moduł ten dzieli dane na dwie równoliczne grupy, dla naszych potrzeb zmienimy proporcje tego podziału tak, by zbiór testowy zawierał 20% wszystkich obserwacji. Ponieważ podział na dane uczące i testowe jest losowy, nie zawsze musi on w sposób najlepszy spełniać wymagania analizy. Proces ustalania parametrów może dać różne wyniki, w zależności od tego, jakie dane znajdują się w poszczególnych zbiorach. Dlatego naukę należy powtórzyć kilkakrotnie dla różnych zbiorów uczących



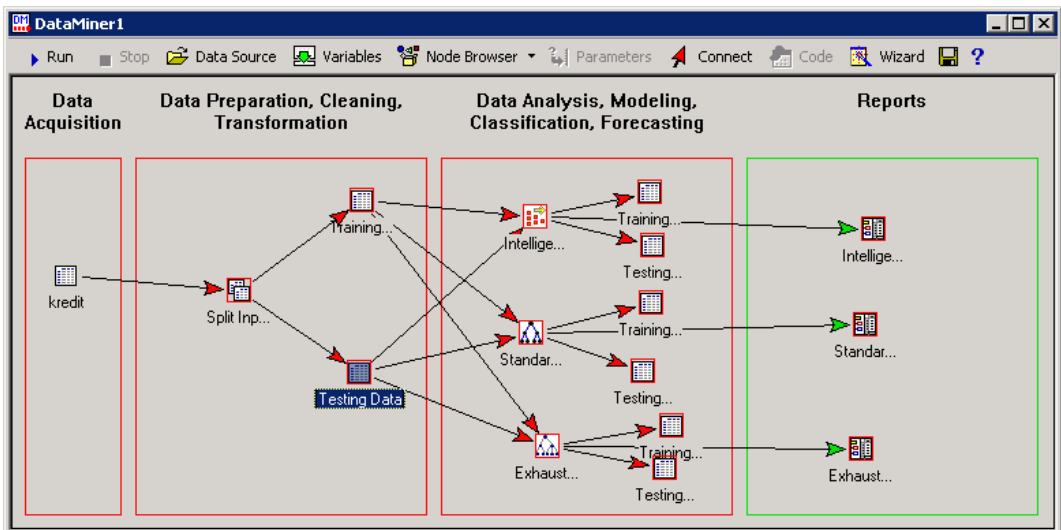
i testowych (ewentualnie zmieniając także proporcje między tymi zbiorami), wybierając ten podział, dla którego modele dają najmniejsze błędy.

Jako narzędzia analizy zastosujemy sieci neuronowe, drzewa klasyfikacyjne C&RT (*Standard Classification Trees with Deployment (C And RT)*) oraz drzewa CHAID (*Exhaustive Classification CHAID with Deployment*). W celu określenia najlepszej architektury sieci wybieramy moduł *Intelligent Problem Solver*. Jest to bardzo użyteczne narzędzie służące do konstruowania modeli sieci neuronowych. Moduł ten automatycznie sprawdza różne typy architektur oraz różne wielkości sieci, wybierając do analizy najlepszą z nich. Zauważamy poszukiwania tego modułu jedynie do sieci o radialnych funkcjach bazowych, zwiększając maksymalną liczbę neuronów bazowych do 50.

W modułach drzew decyzyjnych zmieniamy minimalną liczbę obiektów w węźle, jaka nie może podlegać kolejnym podziałom, na 80. Węzły zawierające mniejszą liczbę obserwacji traktowane będą jako jednorodne i nie będą dalej dzielone.

Warto zauważyć, że wszystkie te metody (podobnie jak inne zawierające w swojej nazwie frazę *with deployment*) umożliwiają zapisanie zbudowanego modelu w postaci kodu C i PMML oraz *STATISTICA Visual Basic*.

Podłączamy wybrane moduły odpowiednio do danych uczących i testowych, a następnie uruchamiamy proces analizy, wybierając polecenie *Run Dirty Nodes* z menu *Run* lub z menu podręcznego przestrzeni roboczej *STATISTICA Data Miner*, co spowoduje uruchomienie analiz tylko dla nowych lub zmienionych węzłów. Wyniki działania modeli pojawiają się w arkuszach wyjściowych osobno dla danych uczących oraz danych testowych.



Ocena modeli

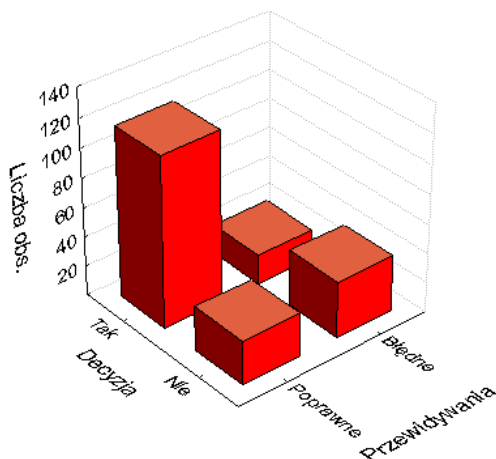
Po wykonaniu analizy należy zweryfikować poprawność zbudowanych modeli. Można zrobić to przy pomocy tradycyjnych narzędzi pakietu *STATISTICA*. Sprawdzamy w ten



sposób skuteczność analizy dla drzew klasyfikacyjnych C&RT. W tym celu wybieramy arkusz powstały w wyniku tej analizy dla wartości testowych. Klikamy prawym przyciskiem myszy na arkuszu i wybieramy opcję *View document*. Otrzymamy arkusz, dla którego zastosować możemy różnego typu analizy. Przykładowo: z grupy statystyk opisowych możemy wybrać opcję *Tabele wielozdzielcze*. Po kliknięciu przycisku *Określ tabele (wybierz zmienne)* w pierwszej liście wybieramy zmienną numer 3 (*decyzja*), natomiast w drugiej - zmienną numer 2, opisującą, czy dana odpowiedź była poprawna czy błędna. By zobaczyć, jak dla różnych decyzji będzie wyglądać procentowy udział trafnych przewidywań, musimy na karcie *Opcje* zaznaczyć pola *Procenty w całości* oraz *Procenty w wierszach*. Uzyskane wyniki zamieszczono w poniższej tabeli.

Tabela licznosci Standard Classification Trees (C And RT)				
Licznosc oznacz. komorek > 10				
(Nie oznaczono sum brzegowych)				
	decyzja	Odpowiedzi Poprawne	Odpowiedzi Bledne	Wiersz Razem
Liczba	Nie	30	38	68
% z wiersza		44,12%	55,88%	
% z calosci		14,71%	18,63%	33,33%
Liczba	Tak	116	20	136
% z wiersza		85,29%	14,71%	
% z calosci		56,86%	9,80%	66,67%
Liczba	Ogól grp	146	58	204
% z calosci		71,57%	28,43%	

Analizując otrzymane wyniki, możemy stwierdzić, że dla 71,57% przypadków model przewidział prawidłowe odpowiedzi. Można również zaobserwować, że model częściej myli się podczas rozpoznawania przypadków, dla których rzeczywista decyzja była negatywna (wśród tej grupy jedynie 44,12% prawidłowych klasyfikacji). Uzyskane wyniki możemy także zaprezentować w postaci histogramu 3D dwóch zmiennych:



Do łącznej weryfikacji wszystkich zbudowanych modeli można użyć specjalnego węzła analitycznego *Compute Best Predicted Classification from all Models*. Wybieramy ten



węzeł i łączymy go z węzłem danych testowych (*Testing Data*). Efektem działania tej analizy jest arkusz *Final Prediction for decyzja* opisujący wyniki działania poszczególnych modeli. Dodatkowo zaprezentowane zostaną wyniki uzyskane w wyniku działania modelu opartego na głosowaniu modeli.

By dokonać oceny tych modeli, musimy w otrzymanym arkuszu wybrać zmienne potrzebne nam do tego celu. Pierwotnie zmienne pokrywają się z początkową specyfikacją, ponieważ domyślnie arkusz ten stosowany jest do dalszego uczenia. Jako skategoryzowaną zmienną zależną wybieramy zmienną decyzja, natomiast jako zmienne niezależne ustalamy zmienne powstałe w wyniku przewidywania poszczególnych modeli. Po określeniu zmiennych zaznaczamy opcję *Always use this selection*, by przy odświeżaniu modelu nie zostały one zamienione domyślnym wyborem. Do tak przygotowanego arkusza podłączamy węzeł *Goodness of Fit*. Przed jego uruchomieniem należy zmienić jego ustawienia na ocenę modeli dla zmiennych skategoryzowanych. Wyniki działania tego modułu umożliwiają nam ocenę poszczególnych modeli oraz modelu zagregowanego, co jest podstawą do wyboru ostatecznego modelu i ustalenia, czy uzyskiwane przez niego wyniki są zadowalające (przykładowo: czy spełniają założenia odnośnie poziomu dopuszczalnego przez nas błędu).

Tablica błędnych odpowiedzi dla zmiennej decyzja				
decyzja	Drzewa klasyfikacyjne C&RT % Błędnych	Drzewa klasyfikacyjne CHAID % Błędnych	Intelligent Problem Solver % Błędnych	Głosowanie modeli % Błędnych
	Tak	14,81481	17,03704	28,88889
Nie	55,22388	59,70149	41,79104	53,73134

Możemy zauważyć, że najlepsze odpowiedzi dla decyzji Tak są generowane przez model powstały w oparciu o głosowanie modeli, w wypadku decyzji Nie najlepiej sprawdzają się sieci neuronowe. Ponieważ model często się myli, sugerując przyznanie kredytów osobom niewiarygodnym, natomiast z większą trafnością wyznacza osoby, którym nie powinno się przyznawać kredytu, jego użyteczność uwidacznia się szczególnie we wstępnej fazie analizy do oddzielenia kredytobiorców, którym na pewno nie należy przyznać kredytu. Osoby uznane przez model za wiarygodne wymagają dodatkowych analiz mogących w sposób bardziej wiarygodny przydzielić kredytobiorców do odpowiedniej grupy.

Warta uwagi jest możliwość zapisania całego projektu data mining w pliku i udostępniania go innym użytkownikom *STATISTICA Data Miner*. Jeśli projekt ten umieścimy w repozytorium *Web STATISTICA*, to będzie można z niego korzystać i udoskonalać go w środowisku internetowym.

Literatura

1. *Bankowość Podręcznik akademicki*, red. Jaworski W., Zawadzka Z., Poltext, Warszawa 2001.
2. Borys G., *Zarządzanie ryzykiem kredytowym w banku*, Wydawnictwo naukowe PWN, Warszawa-Wrocław 1996.



3. Gruszczyński M., Modele i prognozy zmiennych jakościowych w finansach i bankowości, Szkoła Główna Handlowa, Warszawa 2002.
4. Lula P., *Jednokierunkowe sieci neuronowe w modelowaniu zjawisk ekonomicznych*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków 1999.
5. Sokołowski A., Demski T., Analizy statystyczne i data mining z zastosowaniem oprogramowania StatSoft, StatSoft Polska, Kraków 2003.
6. Zawadzka Z., Zarządzanie ryzykiem w banku komercyjnym, Poltext, Warszawa 2000.