

Scoring kredytowy a modele data mining

Grzegorz Migut, Janusz Wątroba

RYZIKO
W INSTYTUCJI
FINANSOWEJ



Wprowadzenie

Jedną z najbardziej charakterystycznych cech otaczającej nas rzeczywistości jest niepewność. Dotyczy to zarówno zjawisk i procesów przyrodniczych, jak i tych, które są związane z szeroko pojętym otoczeniem gospodarczym i społecznym. Nie będzie chyba zbyt-niej przesady w stwierdzeniu, że cała ludzka aktywność jest bardzo mocno powiązana z przewidywaniem. Przyczyny występowania niepewności mogą być bardzo różne. Sama natura zjawisk i procesów implikuje bardzo często występowanie niepewności. Niekiedy jej przyczyną jest niepełna lub nieścisła informacja albo wręcz całkowity brak informacji.

Mimo, iż potrzeba poskromienia niepewności pojawiła się w działalności człowieka bardzo dawno temu, to jednak skuteczne ilościowe sposoby jej oceny pojawiły się stosunkowo niedawno (Rao 1994). Zagadnienie to ma szczególne znaczenie w sytuacjach decyzyjnych. Podejmowanie decyzji w warunkach niepewności zawsze jest obciążone ryzykiem. Stąd bierze się powszechne zapotrzebowanie na narzędzia wspierające procesy decyzyjne. Wymagania stawiane tym narzędziom to przede wszystkim minimalizacja ryzyka błędnej decyzji (koszty błędnych decyzji są coraz wyższe), szybkość działania (warunki konkurencji) oraz możliwość uwzględniania różnych informacji jakościowych i ilościowych.

Wiele spośród współczesnych dziedzin działalności gospodarczej jest powszechnie kojarzonych z występowaniem niepewności. Wystarczy wspomnieć o finansach, bankowości i ubezpieczeniach.

Ryzyko kredytowe

Działalność banku jest nierozzerwalnie związana z występowaniem ryzyka. Można byłoby chyba zaryzykować stwierdzenie, że istotą działalności banku jest identyfikacja różnych rodzajów ryzyka, ocena jego wielkości oraz stosowanie odpowiednich procedur zarządzania ryzykiem. Biorąc pod uwagę specyfikę funkcjonowania banku można tutaj wspomnieć o ryzyku związanym z płynnością finansową, poziomem stóp procentowych, działalnością kredytową czy też ryzykiem występującym na rynku kapitałowym (np. inwestycje). Konieczność odpowiedniego podejścia do zagadnienia ryzyka nie jest wyłącznie sprawą samego banku. Za zapewnienie bezpiecznego operowania powierzony-

Zbudowanie dobrego modelu scoringowego nie jest zadaniem łatwym i wymaga zwykle bardzo przemyślanego zaprojektowania całego przedsięwzięcia. Proces budowy analitycznych modeli jest tylko jednym z etapów tego procesu, w dużym stopniu uzależnionym od jakości i rzetelności zebranych danych. Trafność oceny zdolności kredytowej zależy zatem od tego, co zostało w taki model wbudowane. W niniejszej pracy zaprezentowana zostanie przykładowa analiza za pomocą sieci neuronowych oraz drzew klasyfikacyjnych



mi środkami finansowymi oraz utrzymywanie odpowiednich standardów w zakresie ryzyka jest także odpowiedzialny bank centralny. Ważną rolę pełni też tutaj odpowiednie prawodawstwo (Janc i Kraska 2001).

Spśród różnych form działalności banku szczególną wagę przywiązuje się do bezpieczeństwa działalności kredytowej. Efektem tego są formalne wymogi stawiane bankom przez ustawę Prawo bankowe. W jej świetle bank ma obowiązek zbadania zdolności kredytowej osoby lub jednostki gospodarczej starającej się o kredyt. Bankom pozostawiono natomiast dużą swobodę w zakresie wyboru odpowiednich metod oceny zdolności kredytowej. Z punktu widzenia potrzeb banku najważniejsze wymagania stawiane metodom wspomagającym ocenę zdolności kredytowej to:

- łatwa dostępność informacji stanowiących podstawę oceny ryzyka,
- sprawne przetwarzanie tych informacji, z uwzględnieniem czasu i kosztów,
- możliwość łatwej interpretacji proponowanych reguł oraz ustalenia jednoznacznej decyzji dotyczącej przyznania kredytu.

W literaturze są omawiane różne podejścia stosowane przy ocenie zdolności kredytowej (Janc i Kraska 2001, Gruszczyński 2002). Mimo różnic, na jakie zwracają uwagę niektórzy autorzy, podejścia te są w zasadzie bardzo podobne. Chodzi generalnie o to, żeby ocenić, czy dana osoba starająca się o przyznanie kredytu daje gwarancję jego spłaty. W tym celu pracownik podejmujący decyzję o przyznaniu kredytu gromadzi pewne informacje charakteryzujące kredytobiorcę, a następnie w oparciu o określoną metodykę ocenia dany wniosek kredytowy.

Za najpopularniejszą metodę oceny ryzyka kredytowego uznać można scoring kredytowy. Metoda ta polega na wyznaczeniu na podstawie różnych charakterystyk kredytobiorcy pewnej punktowej oceny, która następnie wykorzystywana jest do klasyfikowania kredytobiorcy do grupy o określonym poziomie ryzyka.

Scoring kredytowy

Z technicznego punktu widzenia ilościowa ocena ryzyka kredytowego polega na obliczeniu prognozy dla jakościowej zmiennej o rozkładzie dwumianowym (jest to zmienna objaśniana). Wartości zmiennej oznaczają przynależność osoby starającej się o kredyt do jednej z dwóch kategorii:

- osoby, którym nie należy przyznawać kredytu (ze względu na duże ryzyko dla banku)
- osoby, którym można kredyt przyznać (niewysokie ryzyko).

Lista potencjalnych zmiennych objaśniających (predyktorów) obejmuje różnego rodzaju informacje jakościowe i ilościowe zarówno na temat samego klienta, jak i je-

go otoczenia. W praktyce punktowa ocena ryzyka oznacza wartość liczbowa, z pewnego przedziału (wygodnie jest przyjąć zakres od 0 do 100 punktów). Wartość prognozy obliczona dla konkretnego klienta oznacza szacunkowy poziom ryzyka danego wniosku kredytowego.

Ocenę punktową uzyskuje się na podstawie opracowanego wcześniej modelu scoringowego. Kształt takiego modelu ustala się na podstawie doświadczeń banku z osobami, które poprzednio starały się o przyznanie kredytu. 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.

Metody analizy danych stosowane w scoringu kredytowym

Skonstruowanie dobrego modelu, tzn. takiego, który pozwala trafnie zakwalifikować dany wniosek kredytowy do jednej z dwóch klas (zły i dobry) nie jest zadaniem łatwym. Warunkiem koniecznym jest tutaj posiadanie takich informacji o osobie czy firmie starającej się o kredyt, które rzeczywiście pozwalają ocenić zdolność kredytową. Nawet najbardziej wyszukane metody modelowania nie zapewnią zbudowania dobrego modelu na podstawie nieadekwatnych bądź nierzetelnych danych. Warto o tym szczególnie pamiętać przy tworzeniu koncepcji zbierania odpowiednich informacji o kredytobiorcach.

W tym miejscu wypada wreszcie wyjaśnić, dlaczego w tytule artykułu znalazło się określenie „modele *data mining*”. Najważniejszy powód polega na tym, że przy konstruowaniu modeli scoringowych dobiera się odpowiednie informacje o kredytobiorcach w taki sposób, aby na ich podstawie móc trafnie oceniać, czy dany kredyt dobrze czy źle rokuje (tak więc bierze się pod uwagę własności predykcyjne tych informacji). Nie opieramy się zatem na jakiejś istniejącej teorii (z zakresu bankowości lub psychologii konsumenta) czy też postulowanych mechanizmach kształtujących ryzyko kredytowe lecz poszukujemy pewnych wzorców lub prawidłowości w zebranych wcześniej danych. Właśnie dane stanowią punkt wyjścia przy budowie modelu, a nie *a priori* sformułowane hipotezy. Taki sposób prowadzenia analizy danych jest charakterystyczny dla dziedziny nazywanej *data mining*. Podejście to znajduje zastosowanie przede wszystkim w tych obszarach badań empirycznych, gdzie z różnych względów nie jest możliwe przeprowadzanie ściśle zaplanowanych i dobrze kontrolowanych eksperymentów (np. ze względu na koszty lub powody etyczne) lub występuje brak wystarczająco uzasadnionych teorii lub też złożoność zjawisk jest zbyt

duża (np. w medycynie, naukach społecznych, ekonomii, finansach czy ubezpieczeniach).

Stosowanie tych metod nie wymaga co prawda spełnienia dość krępujących założeń, jakie są stawiane przy klasycznym (statystycznym) podejściu do analizy danych, ale niejako w zamian za to z reguły wymagana jest bardzo duża liczba obserwacji, a ponadto wyniki niełatwo dają się uogólniać na populację. Mimo tego można zaobserwować zarówno stały rozwój tego typu metod, jak i niemalejący zakres praktycznych zastosowań.

Przy budowie modeli wykorzystywanych w scoringu kredytowym stosuje się metody uwzględniające fakt, że zmienna zależna (objaśniana) jest zmienną jakościową (najczęściej jest to zmienna dychotomiczna). Wśród powszechnie stosowanych technik analitycznych można zauważyć zarówno bardziej klasyczne metody, np. regresję logistyczną czy analizę dyskryminacyjną, jak i techniki zaliczane często do grupy metod określanych terminem *machine learning* (wśród których wymienia się np. uogólnione modele addytywne, techniki drzew decyzyjnych, metodę *Support Vector Machines*, metody wykorzystujące sztuczne sieci neuronowe). Obecnie istnieje już dość bogata literatura na ten temat, zwłaszcza w języku angielskim (m. in. Hastie i wsp. [2001], Krawiec i Stefanowski [2003], Giudici [2003], Lasek [2002], Gruszczynski [2002]).

Praktyczne wykorzystanie tych metod jest coraz bardziej dostępne ze względu na szybki rozwój odpowiedniego oprogramowania – bardzo dobrym przykładem mogą być systemy i programy analizy danych z rodziny STATISTICA firmy StatSoft.

Zalety i wady punktowej oceny ryzyka kredytowego

Omawiana metoda oceny zdolności kredytowej posiada zarówno swoje dobre strony jak i pewne mankamenty (Janc i Kraska 2001). Podstawowa jej zaleta polega na tym, że jej przeprowadzenie jest bardzo łatwe oraz pozwala zaoszczędzić czas niezbędny do przeprowadzenia analizy wniosku kredytowego. Jednocześnie jest to metoda umożliwiająca obiektywną ocenę zdolności kredytowej. Dobrze przygotowany model scoringowy pozwala zmniejszyć liczbę złych kredytów. Dla banku jej stosowanie umożliwia na ogół zwiększenie wydajności pracy urzędników kredytowych oraz obniżenie kosztów obsługi. Z kolei dla osoby starającej się o kredyt stosowanie przez bank oceny scoringowej oznacza zmniejszenie liczby wymaganych dokumentów niezbędnych do przeprowadzenia oceny zdolności kredytowej.

Chociaż generalna ocena scoringu kredytowego jest na ogół bardzo wysoka, to jednak nie sposób pominąć również pewne zarzuty, które są jej stawiane. Czasami stosowane karty scoringowe dyskryminują pewne grupy społeczne. Podobnym zarzutem jest to, że przy budowie karty scoringowej wykorzystuje się jedynie informacje o kredytobiorcach, którym kredyt został przydzielony, a pomija się grupę klientów, których wnioski zostały odrzucone. Przynajmniej z częścią wad przytoczanych

przez różnych autorów można chyba polemizować. Przykładowo można zapobiec możliwości szybkiej dezaktualizacji stosowanego przez dany bank systemu scoringowego poprzez bieżące wprowadzanie do systemu informacji o nowych kredytobiorcach i uaktualnianie wykorzystywanej karty scoringowej. Kolejny zarzut to klasyfikowanie kredytobiorców tylko do dwóch grup ryzyka (kredyt dobry i zły). W tym przypadku istnieje możliwość wprowadzenia trzeciej kategorii (np. nieokreślony). Wymaga to tylko zastosowania odpowiednich metod modelowania w przypadku zmiennych jakościowych (w takim przypadku zmienna objaśniana przyjmuje postać zmiennej jakościowej wielomianowej).

Warto również pamiętać, że scoringu nie powinno się stosować dla wszystkich podmiotów starających się o kredyt. Wykorzystuje się go najczęściej w odniesieniu do klientów indywidualnych oraz małych firm.

Modelowanie wiarygodności kredytowej

W modelach scoringowych wykorzystywanych może być szereg metod *data mining*. Sam proces analizy z wykorzystaniem jednej lub wielu metod tego typu 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 one być użyte w analizie (tzw. *preprocessing*).

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 radzenia sobie z modelowaniem zależności o charakterze nieliniowym, a taki właśnie charakter mają zależności opisujące zdolność kredytową. Pewną wadą sieci neuronowych jest natomiast 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 zrozumiała dla człowieka sekwencja reguł decyzyjnych pozwalająca klasyfikować nowe obiekty na podstawie wartości zmiennych. Atrakcyjna jest również możliwość graficznej prezentacji procesu klasyfikacji. Dodatkową zaletą drzew klasyfikacyjnych jest ich odporność na obserwacje odstające.

W obydwu metodach zalecane jest, aby zbadany 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.uni-muenchen.de/service/datenarchiv/kredit/kredit.html>) przedstawiające zbiór 1000 historycznych obserwacji kredytobiorców indywidualnych (w praktyce taka liczba obserwacji może okazać się niewystarczająca do zbudowania poprawnie działającego modelu). W obserwacjach tych wyszczególniono zmienną decyzja – określającą czy dany klient spłacił kredyt czy nie.

Zmienną tą będziemy traktowali jako skategoryzowaną zmienną zależną. Kolejne 20 zmiennych:

- stan konta – aktualny stan rachunku,
- okres kredytu,
- historia kredytowa klienta,
- przeznaczenie kredytu,
- kwota kredytu,
- 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 kredytu,
- inne niespłacone kredyty,

- mieszkanie – mieszkanie wynajmowane, własnościowe lub inne,
 - liczba wcześniejszych kredytów,
 - stanowisko pracy,
 - liczba zaangażowanych osób,
 - posiadanie telefonu,
 - informacja o pochodzeniu klienta,
- pełnić będzie w analizie rolę skategoryzowanych zmiennych niezależnych (predyktorów). W zbiorze danych wyszczególniono również trzy zmienne ilościowe (liczbowe):
- okres kredytowy w miesiącach,
 - wysokość kredytu,
 - wiek kredytobiorcy.

Należy zauważyć, że zmienne te zostały przedstawione również jako zmienne skategoryzowane (okres kredytu, kwota kredytu, wiek kredytu). 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 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 liczości 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 (choćby ze względu na różnice w parametrach oceny), jak również dla klientów o diametralnie różnej wysokości kredytu, ponieważ zachowania poszczególnych grup cechują odmiennie zależności.

Jeśli zgromadzone przez nas dane nie są kompletne, możemy zastąpić brakujące pozycje w wybrany przez nas sposób, np. zastępując je średnią, medianą, wartościami określonymi przez użytkownika lub wręcz usuwając przypadki, w których występują brakujące wartości.

Analizę danych odstających spróbujemy prześledzić na przykładzie zmiennej wysokość kredytu. W tym celu sporządzimy wykres ramka-wąsy. Otrzymany wykres widoczny na rysunku 1. analizujemy pod kontem wartości odstających i ekstremalnych. Zostały one zaznaczone w formie kółek (odstające) i krzyżyków (ekstremalne).

Za odstające zostały uznane te wnioski, w których wartość kredytu przekraczała kwotę 7882 marek. Warto rozważyć nieuwzględnianie tych obserwacji

w dalszej analizie, ponieważ mogą one mieć negatywny wpływ na jakość modelu.

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 przekształcenie zmiennej, by jej wartości zawierały się w przedziale [0,1]. Tego typu przekształcenie jest zwłaszcza przydatne przy wykorzystaniu sieci neuronowych.

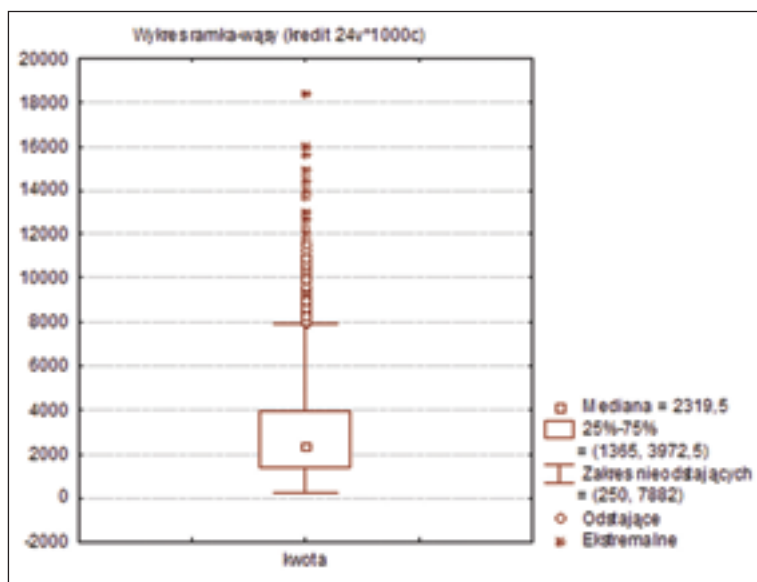
Inną metodą jest dyskretyzacja, polegająca na przekształceniu zmiennej liczbowej w zmienną skategoryzowaną. Podczas przeprowadzania tego typu operacji

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.

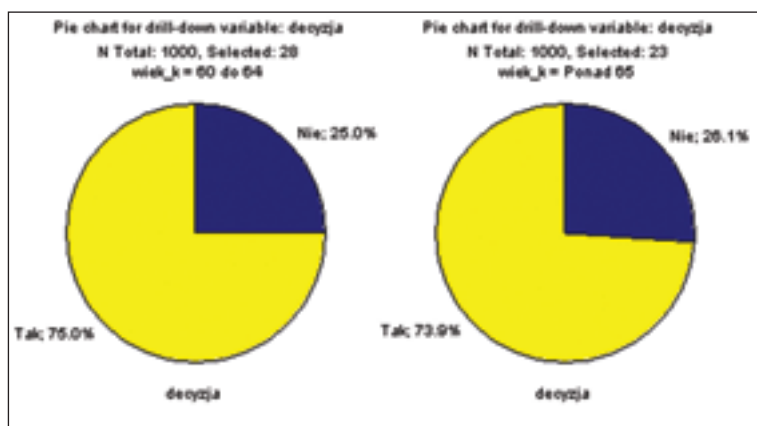
W naszym zbiorze danych dysponujemy grupą dwudziestu zmiennych skategoryzowanych. Analizując licznosci w poszczególnych grupach możemy stwierdzić, że niektóre licznosci są bardzo małe. Na przykład badając zmienną wiek zauważamy, że dwie ostatnie grupy 60-64 lat oraz powyżej 65 lat posiadają niewielkie licznosci. Warto zatem porównać procent poszczególnych decyzji w tych grupach (rysunek 2.) i w wypadku podobnych proporcji, rozważyć możliwość połączenia ich w jedną grupę.

Podział na jednorodne grupy można również wykonać za pomocą specjalnie do tego przygotowanego modułu *Combining Groups for predictive Data Mining*, opartego o drzewa decyzyjne CHAID. Np. wykorzystując ten algorytm do zmiennej cel, w której występuje kilka grup o niewielkiej liczebności np. RTV, wakacje, biznes, otrzymujemy nie dziesięć jak pierwotnie, lecz cztery grupy o dostatecznej li-

Rysunek 1. Analiza danych odstających na przykładzie zmiennej kwota



Rysunek 2. Wyniki interakcyjnego drążenia danych





czebności. Grupy o podobnych charakterystykach dla pozostałych zmiennych zostały ze sobą połączone.

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 (Luła [1999]). 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. Tego typu analizy są konieczne w przypadkach, gdy stosujemy metody czułe 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 liczności w poszczególnych grupach były odpowiednio duże.

Budowa modeli

Tę część analizy najwygodniej jest przeprowadzić w przestrzeni roboczej Data Minera. 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. W STATISTICA Data Miner takiego podziału można dokonać za pomocą węzła *Split Input Data into Training and Testing Samples (Classification)*. Domyślnie moduł ten dzieli dane

na dwie równoliczne grupy, dla naszych potrzeb zmienimy jednak proporcje tego podziału, 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 znajdą 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*) oraz drzewa CHAID (*Exhaustive Classification CHAID with Deployment*). W celu określenia najlepszej architektury sieci można wybierać 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.

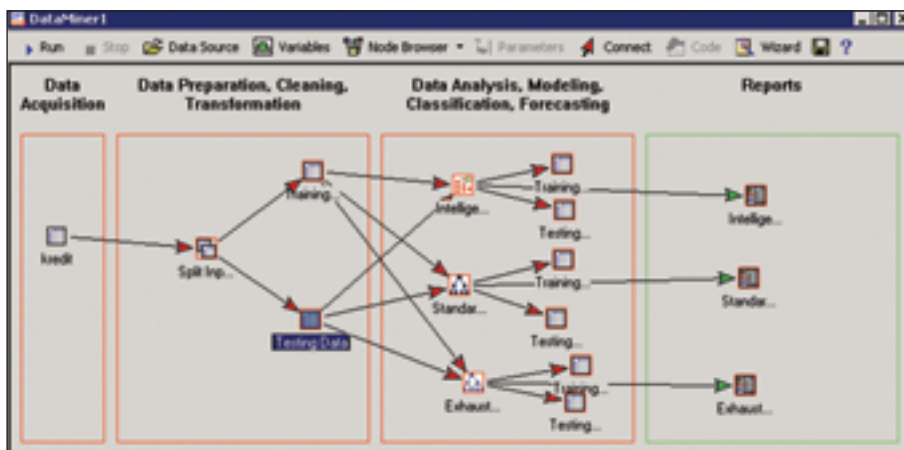
Jeśli w modułach drzew decyzyjnych zmienimy minimalną liczbę obiektów w węzle, jaka nie może podlegać kolejnym podziałom, to węzły te traktowane będą jako jednorodne i nie będą dalej dzielone.

Tak zaprojektowany model używamy do danych uczących się i danych testowych.

Ocena Modeli

Po wykonaniu analizy należy zweryfikować poprawność zbudowanych modeli. Można zrobić to przy pomocy tradycyjnych narzędzi statystycznych. Tabela 1. przedstawia, jak dla różnych decyzji kredytowych będzie wyglądać procentowy udział trafnych przewidywań. Analizując te 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 wyni-

Rysunek 3. Model



ki możemy także zaprezentować w postaci histogramu 3D dwóch zmiennych (rysunek 4.).

Można również przeprowadzić weryfikację wszystkich zbudowanych modeli (sieci neuronowe, drzewa klasyfikacyjne C&RT oraz drzewa CHAID) i modelu opartego na głosowaniu tych modeli (gdy dwa z trzech modeli dają ten sam wynik, to ta decyzja jest podejmowana) Wyniki te zaprezentowane są w tabeli 2.

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 natomiast 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.

Podsumowanie

Trzeba sobie zdawać sprawę, że zbudowanie dobrego modelu scoringowego nie jest zadaniem łatwym i wymaga zwykle bardzo przemyślanego zaprojektowania całego przedsięwzięcia. Proces budowy analitycznych modeli jest tylko jednym z etapów tego procesu, w dużym stopniu uzależnionym od jakości i rzetelności zebranych danych. Trafność oceny zdolności kredytowej zależy zatem od tego, co zostało w taki model wbudowane. Model scoringowy jest budowany w oparciu o dane dotyczące ubiegłych okresów, a zatem musi on być umiejętnie „konserwowany” (tzn. co pewien okres czasu należy sprawdzać, czy jego własności prognostyczne są zachowane).

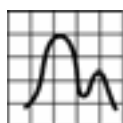
Na zakończenie wypadu przypomnieć, że nawet najlepszy system scoringowy powinien być traktowany jako jeden z elementów wspierających podejmowanie rzeczywistych decyzji kredytowych, a nie jako narzędzie zastępujące decyzje urzędnika kredytowego odpowiedniego szczebla. ■

Janusz Wątroba

*Doktor,
konsultant Data Mining w firmie StatSoft Polska
j.watroba@statsoft.pl*

Grzegorz Migut

*Konsultant Data Mining w firmie StatSoft Polska
g.migut@statsoft.pl*



StatSoft®

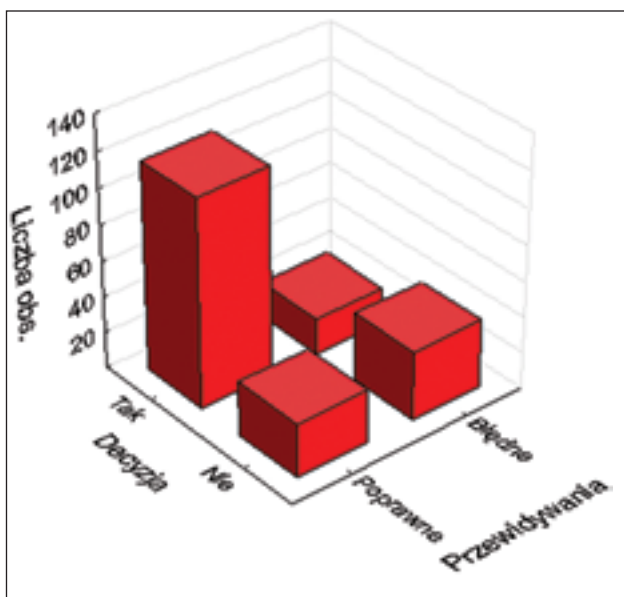
Tabela 1. Trafność decyzji kredytowych

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%	

Tabela 2. Ocena modeli

Tablica błędnych odpowiedzi dla zmiennej decyzja				
	Drzewa klasyfikacyjne C&RT % Blednych	Drzewa klasyfikacyjne CHAID % Blednych	Intelligent Problem Solver % Blednych	Odosowanie modeli % Blednych
decyzja				
Tak	14,91481	17,83704	29,88696	14,07401
Nie	55,22286	59,79146	41,79104	53,73134

Rysunek 4. Histogram dla trafności decyzji kredytowych



Literatura

- [1] Giudici P., 2003, *Applied Data Mining. Statistical Methods for Business and Industry*, Wiley.
- [2] Gruszczyński M., 2002, *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa.
- [3] Hastie T., Tibshirani R., Friedman J., 2001, *The Elements of Statistical Learning*, Springer.
- [4] Janc A., Kraska M., 2001, *Credit-Scoring. Nowoczesna metoda oceny zdolności kredytowej*, Biblioteka Menedżera i Bankowca, Zarządzanie i Finanse, Warszawa.
- [5] Krawiec K., Stefanowski J., 2003, *Uczenie maszynowe i sieci neuronowe*, Wydawnictwo Politechniki Poznańskiej.
- [6] Lasek M., 2002, *Data Mining. Zastosowania w analizach i ocenach klientów bankowych*, Biblioteka Menedżera i Bankowca, Zarządzanie i Finanse, Warszawa.
- [7] Lula P., *Jednokierunkowe sieci neuronowe w modelowaniu zjawisk ekonomicznych*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków 1999.
- [8] Rao C. R., 1994, *Statystyka i prawda*, PWN
- [9] Sokółowski A., Demski T., *Analizy statystyczne i data mining z zastosowaniem oprogramowania StatSoft*, StatSoft Polska, Kraków 2003.