PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 — 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Tomasz Ząbkowski, Wiesław Szczesny
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

KLASYFIKACJA RYZYKA NIEWYPŁACALNOŚCI KLIENTÓW W TELEKOMUNIKACJI

1. Cel badań

Celem prezentowanych badań było zastosowanie drzew klasyfikacyjnych, pojedynczych i zagregowanych przez technikę boosting, do klasyfikacji ryzyka niewypłacalności klienta firmy telekomunikacyjnej. Niniejsze opracowanie ma za zadanie także stwierdzenie zasadności budowy modeli klasyfikacyjnych w zdefiniowanych podzbiorach celem poprawy trafności. W niniejszych analizach do tego celu została wykorzystana informacja o statusie zawodowym klientów, gdzie na podstawie wydzielonych grup została przeprowadzona klasyfikacja mająca na celu stwierdzenie przynależności do danej klasy, związanej z wystąpieniem lub niewystąpieniem ryzyka niewypłacalności. Decyzja o podziale klientów według wcześniej zdefiniowanych grup zawodowych była następstwem przeprowadzenia wstępnej analizy, której wyniki wykazały istotną różnicę w poziomie ryzyka niewypłacalności wśród analizowanych grup zawodowych.

2. Ryzyko niewypłacalności

Terminy „ryzyko niewypłacalności" oraz „ryzyko kredytowe" pojawiają się zamiennie, zwykle w kontekście związanym z działalnością bankową, i wskazują przede wszystkim na niebezpieczeństwo niewypłacalności kredytobiorcy bądź ryzyko bankructwa [Altman 1994; Janc, Kraska 2001; Saunders 1999; Witkowska 2006]. Jednakże określenie to nie jest zarezerwowane wyłącznie do opisu charakterystyki sektora bankowego, gdyż ryzyko kredytowe istnieje wszędzie tam, gdzie odbiorcy dóbr bądź usług otrzymują towary, a moment płatności za nie zostaje odroczony w czasie. Z tego typu sytuacją spotykają się m.in. firmy telekomunikacyjne – użytkownicy systemu abonamentowego korzystają z usług, po czym dopiero po zamknięciu okresu rozliczeniowego otrzymują fakturę. W prezentowanych

rozważaniach ryzyko niewypłacalności zostało zdefiniowane w ten sposób, że zobowiązania niespłacone przez klienta w ciągu 90 dni po wystawieniu faktury były klasyfikowane jako zagrożone. Tym samym na koncie klienta pojawiał się znacznik sugerujący wszczęcie procesu windykacyjno-komorniczego.

Rynek telekomunikacyjny boryka się z problemami ryzyka kredytowego oraz niewypłacalności dopiero od kilkudziesięciu lat, przy czym operator telekomunikacyjny ma znacznie mniej informacji o swoich klientach niż np. banki, co powoduje, że szacowanie ryzyka kredytowego staje się problemem niełatwym, obarczonym dodatkowo wyjątkową dynamiką badanego rynku. W literaturze pewne prace ukazują zastosowania metod klasyfikacyjnych w wypadku problemów rynku telekomunikacyjnego, takich jak przeciwdziałanie nadużyciom [Estevez i in. 2006] oraz szacowanie ryzyka niewypłacalności [Canalli 2001; Daskalaki i in. 2003; Ezawa i in. 1996]. Prace te akcentują specyfikę rynku telekomunikacyjnego oraz dynamikę, która często jest problemem we wdrażaniu rozwiązań z obszaru ryzyka kredytowego.

W obliczu skali zjawiska związanego z dużą liczbą niesolidnych klientów oraz wobec dużej liczby klientów obsługiwanych przez operatorów podstawą do oceny ryzyka nie mogą być stosowane metody traktujące każdy przypadek indywidualnie. Obecnie do modelowania tego typu problemów wykorzystuje się metody pozwalające dokonać odwzorowania bardzo złożonych funkcji, przy czym istotnym elementem jest także ich łatwość zastosowania i prostota interpretacji wyników. Z punktu widzenia potrzeb firmy telekomunikacyjnej wymagania stawiane metodom wspomagającym ocenę ryzyka niewypłacalności to głównie: (a) sprawne przetwarzanie informacji wejściowych będących podstawą oceny oraz (b) możliwość łatwej interpretacji wyniku oraz ustalenia jednoznacznej decyzji dotyczącej przynależności do danej klasy. Do grona takich technik można zaliczyć m.in. drzewa klasyfikacyjne, będące techniką polegającą na stopniowym podziale wielowymiarowej przestrzeni cech na rozłączne obszary aż do uzyskania maksymalnej ich homogeniczności ze względu na wartość zmiennej objaśnianej. Decyzja o zastosowaniu drzew klasyfikacyjnych do rozwiązania problemu niewypłacalności klientów była podyktowana przede wszystkim potrzebą uzyskania wyników (reguł) łatwo interpretowalnych, które można łatwo zastosować w praktyce gospodarczej. Ponadto praca ta ma za zadanie stwierdzić zasadność stosowania modeli klasyfikacyjnych nie dla danych całościowych, lecz w wyodrębnionych podgrupach. Tego typu podejście do segmentacji badanej populacji przed przystąpieniem do budowy modelu zostało m.in. ukazane w pracach [Jiang, Tuzhilin 2006], w zagadnieniu klasyfikacji zachowań konsumentów [Roar i in. 2003], w problemach klasyfikacji i optymalizacji procesu należności w firmie telekomunikacyjnej [Moore 1989] do modelowania decyzji zakupowych.

3. Zastosowane techniki klasyfikacji

Przedstawiony problem oceny ryzyka niewypłacalności użytkownika telekomunikacyjnego sprowadzał się do rozwiązania problemu klasyfikacyjnego, związanego z przypisaniem obiektom przynależności do danej klasy (grupy). W związku z tym, że znane są wartości klas, z których pochodzą obiekty, rozważany był przypadek tzw. uczenia z nauczycielem, zwany inaczej klasyfikacją wzorcową.

Do tworzenia modeli klasyfikacyjnych wykorzystano drzewa klasyfikacyjne (decyzyjne), będące techniką polegającą na stopniowym podziale wielowymiarowej przestrzeni cech na rozłączne obszary aż do uzyskania maksymalnej ich homogeniczności ze względu na wartość zmiennej objaśnianej. Drzewa klasyfikacyjne (decyzyjne) pojawiły się w literaturze w kontekście badań socjologicznych w związku z publikacją w 1963 r. opracowania J.N. Morgana i J.A. Sonquista w czasopiśmie „Journal of the American Statistical Association" [Koronacki, Ćwik 2005]. Oparte na nich algorytmy są na obecny moment najlepiej znane oraz najczęściej wykorzystywane. Struktura drzew decyzyjnych pozwala na konstrukcje najogólniejszych reguł, umożliwiając przy tym niezwykle efektywną ich implementację w wielu zastosowaniach. Według definicji drzewa klasyfikacyjnego z pracy [Koronacki, Ćwik 2005] jest to dowolny spójny graf acykliczny, w którym (1) krawędzie takiego grafu nazywane są gałęziami, (2) wierzchołki, z których wychodzi co najmniej jedna krawędź, nazywamy węzłami (przy czym węzeł drzewa, który nie ma żadnych węzłów macierzystych, nazywamy korzeniem), (3) wierzchołki niebędące węzłami nazywane są liśćmi. Drzewo takie ma dodatkową interpretację dla węzłów, gałęzi i liści, gdzie węzły odpowiadają testom dokonanym na wartościach atrybutów przykładów, gałęzie – możliwym wynikom tych testów, natomiast liście odpowiadają etykietom klas danego problemu decyzyjnego.

Budowa drzewa jest procesem wieloetapowym, podczas którego analizuje się wszystkie zmienne niezależne i wybiera jedną, która daje najlepszy podział węzła skutkujący wydzieleniem najbardziej homogenicznych grup. Głównym kryterium podziału przestrzeni cech jest funkcja (która jest maksymalizowana), dzięki której możliwa jest ocena jakości podziału. Zwykle różnorodność klas w węźle mierzy się za pomocą miar, takich jak:

(1) proporcja błędnych klasyfikacji $P_m(T) = 1 - \hat{p}_i$, (1)

(2) indeks Giniego $G_m(T) = 1 - \sum_{i=1}^{n} \hat{p}_i^2$, (2)

(3) entropia $E_m(T) = -\sum_{i=1}^{n} \hat{p}_i \log \hat{p}_i$, (3)

gdzie m jest węzłem drzewa T, zawierającym obserwacje z n klas, a $\hat{p}_i$ oznacza proporcję obserwacji z klasy i w węźle m. Stosowane podczas budowy drzew algorytmy podziału dotyczą przede wszystkim sposobu wyboru podziału dla węzła, zasad podejmowania decyzji o utworzeniu liścia lub węzła oraz technik uwzględ-

niania zaburzeń w przypadkach uczących. W niniejszym opracowaniu wykorzystywany był algorytm CART (*Classification and Regression Tree*), będący metodą wyczerpującego poszukiwania podziałów jednowymiarowych.

Drugą techniką wykorzystaną w niniejszym opracowaniu jest tzw. wzmacnianie (boosting). Uznaje się, że boosting jest ogólną metodą, u podstaw której leży idea poprawy dokładności działania dowolnego algorytmu uczącego. Polega na wykorzystywaniu sekwencji prostych modeli, przy czym każdy kolejny model przykłada większą wagę do obserwacji źle zaklasyfikowanych przez poprzednie modele. W pracy tej został wykorzystany podstawowy algorytm boostingu noszący nazwę AdaBoost. Przy założeniu, że:

(I) dysponujemy zbiorem uczącym U o m elementach pochodzących z K klas; zbiór $U = \{(x_i, y_i), i = 1, 2, \ldots, m\}$ i $y_i \in Y = \{1, 2, \ldots, K\}$, gdzie x_i – i-ta obserwacja, a y_i – klasa i-tej obserwacji;

(II) dany algorytm uczący jest funkcją tworzącą klasyfikator na podstawie podzbiorów zbioru U, z uwzględnieniem wcześniej nadanych wag;

(III) stała L oznacza maksymalną liczbę iteracji,

algorytm ten ma następujące etapy: (1) inicjalizacja wag dla każdej obserwacji o wartości $1/m$, a następnie iteracyjnie (2) normalizacja wag, (3) wyznaczenie klasyfikatora, (4) wyznaczenie błędu klasyfikatora, (5) wyznaczenie nowych wag z uwzględnieniem błędu.

Proces iteracji kończy się, gdy błąd klasyfikatora przekroczy wartość 0,5 lub gdy numer iteracji równy jest stałej L. Ostateczny klasyfikator wyznaczany jest na podstawie głosowania większościowego. Opisana powyżej procedura wzmacniania została oparta w niniejszym opracowaniu na drzewach decyzyjnych, z uwzględnieniem założenia, że liczba drzew przy douczaniu jest adekwatna do liczby danych i złożoności problemu.

Efektywność badanego problemu klasyfikacji została oceniona za pomocą dwóch miar. Pierwsza miara opisuje procent poprawnych klasyfikacji (PPK) na podstawie zbudowanego modelu i tym samych decyduje o jakości klasyfikacji. Miara ta ma następującą postać:

$$PPK = \frac{n_{00} + n_{11}}{n} \times 100, \tag{4}$$

gdzie n – liczba obserwacji, n_{00} – liczba obserwacji, dla których $\hat{y}_i = y_i = 0$, n_{11} – liczba obserwacji, dla których $\hat{y}_i = y_i = 1$ oraz $\hat{y}_i$, y_i to odpowiednio wartości uzyskane oraz wartości obserwowane zmiennej zależnej, dla $i = (1, 2)$.

Zwykle oprócz miary (4) do stwierdzenia skuteczności modelu wykorzystuje się miarę (5) opisującą udział poprawnie rozpoznanych przypadków (UPK) w poszczególnych percentylach posortowanych względem malejącego prawdopodobieństwa zajścia zdarzenia:

$$UPK_{pct} = \frac{n_{11}}{n} \times 100, \tag{5}$$

gdzie n – liczba obserwacji, n_{11} – liczba obserwacji, dla których $\hat{y}_i = y_i = 1$ oraz $\hat{y}_i$, y_i to odpowiednio wartości uzyskane oraz wartości obserwowane zmiennej zależnej, dla $i = (1, 2)$.

Popularność tej miary jest uzasadniona względami biznesowymi. W przypadku, kiedy model ma być stosowany na części populacji, istotną miarą mówiącą o jakości modelu jest trafność w zależności od wielkości bazy, do której możemy dotrzeć. Często nie ma możliwości, np. z powodu kosztów, dotarcia do całej populacji w celu ograniczenia ryzyka związanego z niesolidnymi użytkownikami, którzy nie uiszczają należności za dostarczone usługi. W związku z tym, że modele klasyfikacyjne grupują przypadki w zależności od prawdopodobieństwa przynależności do danej klasy, informacja ta może posłużyć do uporządkowania tychże przypadków według malejącego ryzyka zajścia przewidywanego zdarzenia i tym samym wyliczenia miary (5).

4. Dane empiryczne

Dane wykorzystane do budowy modeli klasyfikacyjnych stanowiły losową próbkę 20 tys. klientów operatora telekomunikacyjnego, przy czym w ramach tej zbiorowości zostało wydzielonych pięć równolicznych grup według zmiennej określającej status zawodowy. Grupy obejmowały następujący podział: osoby fizyczne prowadzące działalność gospodarczą, osoby fizyczne pracujące, uczący się (uczeń, student), emeryci/renciści, osoby prowadzące działalność rolniczą. Definiując analizowaną zbiorowość jako zbiorowość C, zawierającą N obiektów (klientów), każdy obiekt C_i możemy opisać przez zbiór atrybutów $X = \{x_1, x_2, ..., x_n\}$ obejmujących m.in. cechy demograficzne, zmienne transakcyjne (płatności, charakterystyki wykonywanych połączeń) oraz różnego typu agregaty (sumy, średnie) będące pochodną zmiennych transakcyjnych. Dodatkowo każdy obiekt podlegający klasyfikacji reprezentowany był przez zmienną klasyfikującą y, określającą przynależność obiektu do jednej z dwóch rozważanych grup. Zmienna ta jest postaci zero-jedynkowej i przyjmuje wartość 1, jeśli nastąpił brak płatności za usługi telekomunikacyjne dostarczone przez operatora, oraz 0 w przypadku przeciwnym. Tabela 1 prezentuje wyniki przeprowadzonej wstępnej analizy, wskazującej na istnienie istotnych różnic w poziomie ryzyka niewypłacalności wśród analizowanych grup zawodowych.

Tabela 1. Charakterystyka danych według wydzielonych grup zawodowych

Status zawodowy	Klasyfikacja płatności		Udział procentowy należności zagrożonych
	zagrożona	normalna	
Os. fiz. prowadząca działalność	661	3 339	16,5
Os. fiz. pracująca	393	3 607	9,8
Emeryt/rencista	323	3 677	8,1
Uczący się	618	3 382	15,5
Działalność rolnicza	299	3 701	7,5
Ogółem	2294	20 000	11,5

Źródło: obliczenia własne.

Łączna liczba dostępnych zmiennych niezależnych wynosiła 43. W celu uproszczenia eksperymentu praktycznego zdecydowano się na zastosowanie procedury eliminacji cech na podstawie analizy macierzy współczynników korelacji pomiędzy zmiennymi diagnostycznymi a zmienną grupującą. W związku z tym w puli dostępnych zmiennych pozostały te, które wykazały statystycznie istotną korelację na poziomie istotności 0,05. Ponadto jeśli współczynnik korelacji między dwiema cechami diagnostycznymi był większy niż 0,7, w dalszych badaniach uwzględniano tę, dla której współczynnik korelacji ze zmienną decyzyjną był większy. Tym sposobem potencjalny zbiór zmiennych został ograniczony do 9 cech, które były wykorzystane do badań. Zmienne te to: y – zmienna grupująca (1 – płatność zagrożona, 0 – płatność normalna), oraz zmienne $x1$-$x9$, które opisywały użytkowników na podstawie dotychczas zaobserwowanych charakterystyk, dotyczących m.in. płatności, regulowania należności w terminie, liczby usług telekomunikacyjnych i częstości korzystania z nich, długości trwania relacji oraz demografii (wiek, miejsce zamieszkania według GUS).

5. Eksperyment

Podczas eksperymentu dokonano budowy modeli klasyfikacyjnych na podstawie całego zbioru dostępnych danych (20 tys.) oraz w wyodrębnionych pięciu podgrupach (każda o liczności 4 tys.) względem kryterium zawodowego, opisanego we wcześniejszych punktach pracy oraz w tab. 1. We wszystkich przypadkach dostępny zbiór danych został w sposób losowy podzielony na dwie części w następujących proporcjach: 60% obserwacji trafiło do zbioru uczącego, a 40% do zbioru testowego – służącego do przeprowadzenia ostatecznej oceny zbudowanych modeli. Prezentowane wyniki dotyczą wskazań osiągniętych na zbiorze testowym. Eksperyment został zapoczątkowany przez zbudowanie modeli drzew klasyfikacyjnych z dostępnymi zmiennymi $\{x_1, \ldots, x_9\}$ dla danych całościowych oraz we wszystkich wyodrębnionych podgrupach. Przyjęto następującą symbolikę dla utworzonych modeli: T – model drzewa klasyfikacyjnego opracowanego na danych całościowych, $\{T_1, ..., T_5\}$ – modele drzew klasyfikacyjnych opracowane na danych według wydzielonych grup zawodowych, $T_{1\text{-}5}$ – suma modeli $\{T_1, ..., T_5\}$.

Zestawienie wyników klasyfikacji modeli zawierają tab. 2 oraz 3.

Tabela 2. Macierz klasyfikacji dla modelu T dla danych całościowych

	Klasyfikacja uzyskana				
Klasyfikacja obserwowana		1	0	PPK	PPK razem
	1	359	586	38,0%	91,2%
	0	116	6939	98,3%	

Źródło: obliczenia własne.

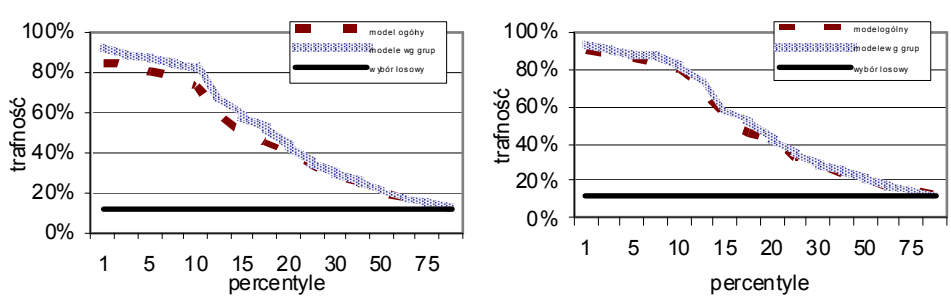
Tabela 3. Macierz klasyfikacji dla modeli dla danych pogrupowanych

			Klasyfikacja uzyskana			
	Modele		1	0	PPK	PPK razem
Klasyfikacja obserwowana	Os. fiz. prowadząca działalność (model T_1)	1	137	137	50,0%	89,3%
		0	34	1292	97,4%	
	Os. fiz. pracująca (model T_2)	1	75	91	45,2%	93,4%
		0	14	1420	99,0%	
	Emeryt/rencista (model T_3)	1	58	70	45,3%	94,3%
		0	21	1451	98,6%	
	Uczący się (model T_4)	1	166	103	61,7%	90,4%
		0	51	1280	96,2%	
	Działalność rolnicza (model T_5)	1	53	55	49,1%	95,9%
		0	10	1482	99,3%	
	Razem ($T_{1\text{-}5}=T_1+T_2+T_3+T_4+T_5$)	**1**	**489**	**456**	**51,7%**	**92,7%**
		0	**130**	**6925**	**98,2%**	

Źródło: obliczenia własne.

Zwraca uwagę fakt, że przy porównaniu uzyskanych wyników dla modeli opartych na danych całościowych T oraz pogrupowanych $T_{1\text{-}5}$ w przypadku miary PPK modele te wykazują podobną skuteczność klasyfikacji, odpowiednio 91,2% oraz 92,7%. Jednakże modele oparte na danych pogrupowanych wykazują znacznie lepszą trafność przypadków ze statusem 1, czyli tych zagrożonych niewypłacalnością, 38% wobec 51,7%.

Biorąc pod uwagę miarę UPK, czyli badanie trafności opracowanych modeli w poszczególnych percentylach, przekonujemy się, że modele na danych pogrupowanych znacznie trafniej wskazują przypadki, z którymi jest związane ryzyko niewypłacalności. Interpretując wyniki porównania modeli przedstawione na rys. 1, można zauważyć, że dysponując nakładami pozwalającymi na monitoring 10% populacji użytkowników i posługując się modelami uzyskanymi w podgrupach, osiągniemy trafność na poziomie 70%. W przypadku modelu na podstawie danych ogólnych osiągnięto poziom klasyfikacji 60%.



Rys. 1. Trafność modeli a) (T wobec $T_{1\text{-}5}$) oraz b) po boostingu według UPK w percentylach
Źródło: obliczenia własne.

Kolejny etap eksperymentu dotyczył opracowania modeli drzew klasyfikacyjnych z uwzględnieniem techniki boostingu, czyli wzmacniania. Biorąc pod uwagę liczność zbioru oraz złożoność problemu, ustalono liczbę 150 pojedynczych klasyfikatorów przy douczaniu. Otrzymane wyniki (tab. 4) wykazały, że zarówno dla modelu opartego na danych całościowych T, jak i dla modeli pogrupowanych $T_{1\text{-}5}$ udało się poprawić trafność klasyfikacji.

Tabela 4. Porównanie klasyfikacji przed zastosowaniem i po zastosowaniu boostingu

Model na podstawie danych		PPK	PPK po boostingu
całościowych T		91,2%	92,8%
Pogrupowane	Os. fiz. prowadząca działalność T_1	89,3%	91,2%
	Os. fiz. pracująca T_2	93,4%	94,0%
	Emeryt/ rencista T_3	94,3%	94,3%
	Uczący się T_4	90,4%	91,6%
	Działalność rolnicza T_5	95,9%	96,8%
	Razem	92,7%	92,9%

Źródło: obliczenia własne.

Jednocześnie, śledząc trafność modeli po zastosowaniu boostingu, tym razem na podstawie miary (5), można zauważyć, że trafność modelu opracowanego na podstawie danych ogólnych zbliżyła się do tych osiągniętych w grupach. Mimo wszystko model będący zestawieniem modeli opracowanych w podzbiorach wykazuje większą trafność.

6. Podsumowanie

Otrzymane w wyniku eksperymentu rezultaty pozwalają na sformułowanie następujących wniosków. Drzewa decyzyjne oraz drzewa ze wzmacnianiem w zastosowaniu do klasyfikacji ryzyka niewypłacalności wydają się skutecznym narzędziem, które w istotny sposób pozwala na skuteczniejsze zabezpieczenie przychodów.

Zaproponowane podejście potwierdza zasadność podziału zbiorowości według określonych kryteriów w celu budowy modeli klasyfikacyjnych w podzbiorach. Analiza wykazała, że budując modele dla określonych grup, w przeciwieństwie do jednego modelu ogólnego, możemy trafniej klasyfikować zjawisko powstawania należności zagrożonych.

Porównując modele drzew decyzyjnych z modelami otrzymanymi poprzez boosting, można uzyskać bardziej skuteczne i jednocześnie stabilne modele. Przy czym w przypadku modelu ogólnego, boosting znacznie poprawia trafność poprawnych wskazań modelu. Wobec tego celem poprawy klasyfikacji należy rozważyć wzmacnianie modeli lub budować modele na podstawie danych w podzbiorach.

Literatura

Altman E., Giancarlo M., Varetto F. (1994), *Corporate distress diagnostic: comparison using linear analysis and neural networks*, „Journal of Banking and Finance" no 18.

Canalli E. (2001), *Experimenting neural networks to forecasts business insolvency*, „Neural Network World" no 11 (4), s. 349-361.
Daskalaki S., Kopanas I., Goudara M., Avouris N. (2003), *Data mining for decision support on customer insolvency in telecommunications business*, „European Journal of Operational Research" no 145, s. 239-255.
Estevez P., Held C., Perez C. (2006), *Subscription fraud prevention in telecommunications using fuzzy rules and neural networks*, „Expert Systems with Applications" no 31, s. 337-344.
Ezawa K.J., Norton S.W. (1996), *Constructing Bayesian networks to predict uncollectable telecommunication accounts*, „IEEE Expert Systems with Applications" no 11, s. 45-51.
Janc A., Kraska M. (2001), *Credit scoring. Nowoczesna metoda oceny zdolności kredytowej*, Biblioteka Menedżera i Bankowca, Warszawa.
Jiang T., Tuzhilin A. (2006), *Segmenting customers from population to individuals: does 1-to-1 keep your customers forever*, „IEEE Transactions on Knowledge and Data Engineering" no 18(10), s. 1297-1311.
Koronacki J., Ćwik J. (2005), *Statystyczne systemy uczące się*, Wyd. WNT, Warszawa.
Moore L. (1989), *Modelling store choice: a segmented approach using stated preference analysis*, „Transactions of the Institute of British Geographers", vol. 14, no 4, s. 461-77.
Roar G., Eleazer M., Per-Age B. (2003), *One size fits all? Segmenting customer base for maximum returns*, Proceedings of the 2003 Winter Simulation Conference, red. S. Chick, P.J. Sánchez, D. Ferrin, D.J. Morrice, New Orleans, s. 1848-52.
Saunders A. (1999), *Managing credit risk*, John Wiley and Sons, New York.
Witkowska D. (2006), *Discrete choice model application to the credit risk evaluation*, „International Advances in Economic Research" no 12, s. 33-42.

CUSTOMER INSOLVENCY CLASSIFICATION IN THE TELECOMMUNICATION MARKET

Summary

The article presents an application of classification methods (decision trees and decision trees with boosting) for customer insolvency problem on the example of the telecommunication market. Based on characteristics of the customers, models to estimate their credit risk were proposed. Five groups of the customers were proposed based on employment status. The results confirm the usefulness and high performance of the methods proposed. The approach to analyze customers in groups can be considered as a good method for more effective revenue assurance.