

Iwona Staniec
Politechnika Łódzka

DRZEWIA KLASYFIKACYJNE W OCENIE WIARYGODNOŚCI KREDYTOBIORCÓW

1. Wstęp

Ostatnio w Polsce pojawiło się wielu pośredników finansowych, udzielających kredytów celowych bez zbędnych formalności bankowych. Pośrednik, podejmując decyzję o przyznaniu kredytu, podejmuje ryzyko kredytowe zależne od danych uzyskiwanych od osób ubiegających się o kredyt. Badając zdolność kredytową swoich klientów, próbuje ustalić i oszacować specyficzne przy zaciąganiu kredytu „ryzyko jakościowe” (częściowy lub zupełny brak spłaty, zwłokę z zapłatą), by stwierdzić, czy może liczyć na spłatę kredytu zgodną z przepisami.

Ze względu na zdolności kredytowe kredytobiorców dzielimy na „wiarygodnych” – dokonujących całkowitej spłaty kredytu według umowy, i „niewiarygodnych” – niespłacających kredytu lub spłacających go z niedopuszczalnymi opóźnieniami.

2. Charakterystyka danych

W niniejszym artykule dokonano oceny ryzyka kredytowego na podstawie danych uzyskanych od firmy „Gwarant” o kredytach indywidualnych na zakup telewizora. Próba obejmowała 4746 umów kredytowych¹, z których niespłacono 250 (czyli 5,268% wszystkich umów). 4496 umów (czyli 94,732% całości) to kredyty spłacone. Szkodowość w analizowanej próbie wyniosła tylko 5,268%. Celem artykułu jest sprawdzenie efektywności zastosowania drzew klasyfikacyjnych w ocenie uszkodzoności (por. [Staniec 2004, s. 87-95]).

¹ Szczegółowy opis bazy znajduje się w pracy [Szmit, Szmit, Kaniewski 2003, s. 163-172].

Tabela 1. Podział na zbiór uczący i testowy

Zbiór uczący (3000)		Zbiór testowy (1746) ^a	
wiarygodni kredytobiorcy	niewiarygodni kredytobiorcy	wiarygodni kredytobiorcy	niewiarygodni kredytobiorcy
2842 (94,7%)	158 (5,3%)	1654 (94,7%)	92 (5,3%)

Źródło: opracowanie własne.

Każdy obiekt badania był opisany przez następujące atrybuty²:

- pierwsza wpłata – zmienna quasi-ciągła – określała wysokość pierwszej raty,
- płeć – zmienna dyskretna dwuwartościowa,
- liczba rat – zmienna dyskretna wielowartościowa (liczba rat, na które rozłożona była spłata – liczba całkowita z przedziału od 3 do 24),
- wiek kredytobiorcy w chwili zawarcia umowy – zmienna dyskretna wielowartościowa, odpowiadająca wiekowi w latach (kredytobiorcy mieli od 19 do 93 lat),
- kwota kredytu w zł,
- cena telewizora w zł.

Badania przeprowadzono na pełnej bazie, przy czym prawdopodobieństwo przynależności do każdej z grup za pierwszym razem szacowano z próby, a za kolejnym razem przyjmowano identyczne prawdopodobieństwo wynoszące 50%.

3. Drzewa klasyfikacyjne

Rozwiązanie procesu decyzyjnego w niepewnych warunkach, polegające przede wszystkim na podjęciu optymalnej decyzji, może być przedstawione na odpowiednim wykresie, w teorii podejmowania decyzji nazywanym drzewem decyzyjnym. Drzewo decyzyjne jest bardzo wygodnym narzędziem w analizie procesów decyzyjnych, zwłaszcza gdy procesy te są skomplikowane.

Drzewa klasyfikacyjne wykorzystuje się do wyznaczania przynależności przypadków lub obiektów do klas jakościowej zmiennej zależnej na podstawie pomiarów jednego lub więcej predyktorów. Liście drzewa odpowiadają klasom, do których należą obiekty, a gałęzie reprezentują wartości cech, na podstawie których dokonano podziału.

Cechami charakterystycznymi drzew klasyfikacyjnych są [Breinan i in. 1984]:

- hierarchiczna natura, polegająca na tym, że zależność liścia od drzewa, na którym rośnie, można opisać w postaci pytań zadawanych w porządku hierarchicznym przez hierarchię podziałów gałęzi (począwszy od pnia), prowadzących do ostatniej gałęzi, z której wyrasta liść; ostatecznie podjęte decyzje zależą od odpowiedzi na wszystkie poprzednie pytania,
- elastyczność, polegająca na możliwości badania wpływu predyktorów pojedynczo (a nie wszystkich jednocześnie) oraz na wykorzystaniu predyktorów różnych typów i osłabieniu założeń dotyczących ich pomiaru.

² Autorka nie miała wpływu na wybór atrybutów, nieznane są jej również kryteria ich wyboru.

Proces tworzenia drzewa klasyfikacyjnego polega na rekurencyjnym podziale zbioru uczącego na podzbiory aż do uzyskania ich jednorodności ze względu na przynależność obiektów do klas. Dąży się do utworzenia drzewa o jak najmniejszej liczbie węzłów, aby otrzymać jak najprostsze reguły klasyfikacyjne.

Algorytm tworzenia drzewa klasyfikacyjnego polega na [Gatmar 1998, s. 168-169]:

- sprawdzeniu, czy wszystkie elementy zbioru obiektów należą do tej samej klasy; jeżeli tak to, postępowanie należy zakończyć, a jeżeli nie, należy rozważyć wszystkie możliwe podziały danego zbioru na możliwie najbardziej jednorodne podzbiory,
- ocenie jakości każdego z otrzymanych podzbiorów zgodnie z przyjętym kryterium i wybraniu najlepszego z nich,
- podziale zbioru obiektów w wybrany sposób, podział ten jest dokonywany na podstawie wartości cech obiektów,
- wykonaniu wymienionych etapów rekurencyjnie dla każdego z podzbiorów.

W procesie tworzenia drzewa klasyfikacyjnego wyróżnia się cztery etapy [Kim, Loh 2003, s. 512-530]:

- określenie kryteriów trafności przewidywania – za trafne przewidywania uznaje się te, gdzie ponosi się minimalne koszty, zwykle odnoszące się do błędnie zakwalifikowanych obiektów; minimalizacja kosztów nie odpowiada jednak minimalizacji proporcji przypadków zaklasyfikowanych błędnie, jeśli weźmie się prawdopodobieństwa *a priori*³ proporcjonalne do wielkości klas i równe koszty błędnych klasyfikacji dla każdej klasy,
- wybranie podziałów – podział zależy od zmiennych wykorzystywanych do określenia przynależności obiektów do poszczególnych klas; wybierane są one pojedynczo od podziału przy węźle źródłowym poprzez podziały wynikowych węzłów, aż do momentu pojawienia się liści,
- wyznaczenie końca podziałów – zaprzestajemy dokonywania podziałów w momencie, gdy wszystkie węzły końcowe są czyste lub zawierają nie więcej niż określoną minimalną liczbę obiektów.
- wybranie drzewa właściwej wielkości – jest pożądane, aby drzewo wyjaśniało fakty w sposób złożony, ale równocześnie zrozumiały.

Najczęściej programy numeryczne stosują następujące metody podziału [Loh, Shih 1997, s. 815-840]:

- dyskryminacyjne podziały jednowymiarowe,
- dyskryminacyjne podziały z wykorzystaniem kombinacji liniowych,
- metoda CART wyczerpującego poszukiwania podziałów jednowymiarowych.

³ Prawdopodobieństwa *a priori* określają, na ile jest prawdopodobne, że dany przypadek lub obiekt zostanie zakwalifikowany do danej klasy, bez żadnej wstępnej wiedzy na temat wartości zmiennych predykcyjnych w modelu. Prawdopodobieństwa *a priori* stosowane w minimalizacji kosztów mogą poważnie wpływać na klasyfikację przypadków lub obiektów.

Drzewa właściwej wielkości wybiera się najczęściej na podstawie tzw. sprawdzianu krzyżowego lub na podstawie minimalizacji kosztów i złożoności drzewa w sprawdzianie krzyżowym. Procedura wyboru drzewa jest istotna, bo pomaga uniknąć nadmiernego dopasowania lub niedopasowania do danych. Należy pamiętać, że stopa błędnych klasyfikacji w próbie uczącej (koszt resubstytucji) ma tendencję do spadku wraz ze wzrostem wielkości drzewa. Z kolei koszty sprawdzianu krzyżowego szybko osiągają minimum, gdy wielkość drzewa na początku rośnie, ale koszty te zaczynają wzrastać, gdy wielkość drzewa staje się bardzo duża. Wybrane drzewo właściwej wielkości jest zawsze bliskie punktowi przegięcia krzywej, to znaczy bliskie punktowi, w którym zaczyna się wyrównywać początkowy ostry spadek kosztów sprawdzianu krzyżowego przy wzroście wielkości drzewa. Procedura wyboru drzewa sprowadza się do wyboru najprostszego (najmniejszego) drzewa z kosztami sprawdzianu krzyżowego bliskimi minimum.

4. Wykorzystanie drzew klasyfikacyjnych

Eksperymenty przeprowadzono na podstawie pakietu statystycznego STATISTICA™ v. 6.0. W pierwszym przypadku do stworzenia drzewa klasyfikacyjnego wybrano metodę dyskryminacji podziałów jednowymiarowych dla predyktorów nominalnych i porządkowych przy równym prawdopodobieństwie przynależności do obu grup (bo przy szacowanych prawdopodobieństwach na podstawie próby drzewo klasyfikacji wszystkim przyznaje kredyt), błąd klasyfikacji drugiego rodzaju wynosi 100%. Jako regułę stopu wybrano przycinanie przy błędzie złej klasyfikacji, co oznacza, że wybieramy drzewa właściwej wielkości z sekwencji drzew optymalnie przyciętych.

Po uzyskaniu podziałów jednowymiarowych zmienne predykcyjne ułożono w skali od 0 do 100, w zależności od tego, jak ważny jest ich wpływ na zmienną zależną. Wyniki przedstawiono w tab. 2. Jak widać, największy wpływ na zmienną „ryzyko kredytowe” mają cechy: „wiek” (100), „pierwsza wpłata” (93) oraz „cena telewizora” (87).

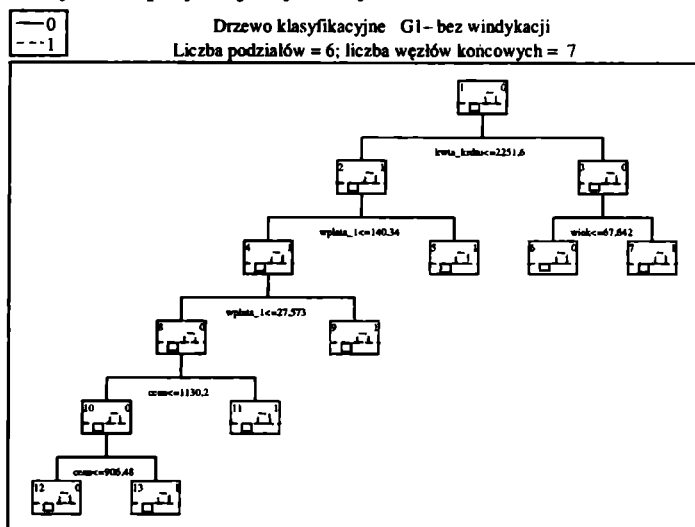
Tabela 2. Ranking ważności predyktorów
(0 = niska ważność; 100 = wysoka ważność)

Cecha	Ranking
Liczba rat	55
Płeć	58
Wiek	100
Pierwsza wpłata	93
Kwota kredytu	75
Cena	87

Źródło: opracowanie własne.

W tym przypadku przebadano 64 różne drzewa. Dla wybranego drzewa koszt sprawdzianu krzyżowego wynosi 0,4 z błędem 0,019, a koszt resubcji – 0,37 z błędem 0,008.

Na początku tworzenia drzewa klasyfikacyjnego wszystkie obiekty przypisano do węzła źródłowego i tymczasowo zakwalifikowano do klasy „niewiarygodnych”. Węzeł źródłowy następnie dzieli się na dwa nowe węzły: obiekty, dla których kwota kredytu jest mniejsza od 2251,62 zł, są tymczasowo zakwalifikowane do przyznania, pozostałym nie przyznajemy kredytu.



Rys. 1. Wybrane drzewo klasyfikacyjne
Źródło: obliczenia własne na podstawie STATISTICA.

Tabela 3. Warunki podziału dla każdego węzła

Lp.	Lewostronna	Prawostronna	n klas 0	n klas 1	Przewidywana klasa	Podział	
	gałąź	gałąź				stała	zmienna
1	2	3	168	2832	0	-251,62	kwota kredytu
2	4	5	105	2217	1	-140,338	pierwsza wpłata
3	6	7	63	615	0	-67,6422	wiek
4	8	9	95	1680	1	-27,5725	pierwsza wpłata
5			10	537	1		
6			53	210	0		
7			10	404	1		
8	10	11	87	1412	0	-1130,22	cena
9			8	268	1		
10	12	13	60	898	0	-906,484	cena
11			7	514	1		
12			60	186	0		
13			20	812	1		

Źródło: obliczenia własne.

Węzeł nr 2 dzieli się według pierwszej wpłaty na dwa podwęzły o numerach 4 i 5. Obiekty o pierwszej wpłacie mniejszej od 140,338 zł zostają przypisane do węzła nr 4, który jest podzielony na dwa kolejne węzły ze względu na wartość pierwszej wpłaty itd. Obiekty niespełniające tego warunku są przypisane do węzła nr 5, który nie podlega dalszemu podziałowi i uznaje obiekty mu przypisane za wiarygodnych kredytobiorców (por. rys. 1 i tab. 3).

Reasumując, wiarygodni kredytobiorcy to klienci, dla których:

- kwota kredytu nie przekracza 2251,62 zł oraz pierwsza wpłata jest większa od 140,338 zł,
- kwota kredytu nie przekracza 2251,62 zł oraz pierwsza wpłata jest mniejsza od 140,338 zł, a większa od 27,5725 zł,
- kwota kredytu nie przekracza 2251,62 zł oraz pierwsza wpłata jest mniejsza od 27,5725 zł, a cena telewizora jest wyższa niż 1130,22 zł,
- kwota kredytu nie przekracza 2251,62 zł oraz pierwsza wpłata jest mniejsza od 27,5725 zł, a cena telewizora jest niższa niż 1130,22 zł, ale wyższa niż 906,48 zł,
- kwota kredytu przekracza 2251,62 zł oraz pierwsza wpłata jest większa od 67,64 zł.

W drugim przypadku do stworzenia drzewa klasyfikacyjnego wybrano metodę CART, wyczerpującego poszukiwania podziałów jednowymiarowych przy równym prawdopodobieństwie przynależności do obu grup. Za regułę stopu przyjęto również przycinanie przy błędzie złej klasyfikacji.

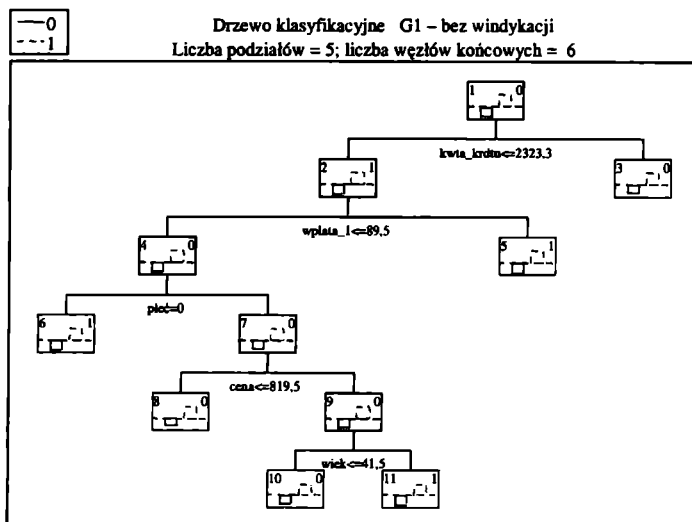
Analiza rankingu predyktorów pokazuje, że największy wpływ na zmienną „ryzyko kredytowe” mają cechy: „wiek” (100), „kwota kredytu” (95) oraz „płeć” (92).

Tabela 4. Przewidywane klasy i warunki podziału dla węzłów

Lp.	Lewostronna	Prawostronna	n klas		Przewidywana	Podział		
	gałąź	gałąź	0	1		stała	zmienna	kateg.
1	2	3	168	2832	0	-2323,31	kwota kredytu	
2	4	5	108	2279	1	-89,5	wpłata_1	
3			60	153	0			
4	6	7	96	2024	0		płeć	0
5			12	755	1			
6			31	1366	1			
7	8	9	65	766	0	-819,5	cena	
8			26	171	0			
9	10	11	39	1195	0	-41,5	wiek	
10			23	131	0			
11			16	1180	1			

Źródło: obliczenia własne na podstawie STATISTICA.

Na początku tworzenia drzewa klasyfikacyjnego wszystkie obiekty przypisano do węzła źródłowego i tymczasowo zakwalifikowano do klasy „niewiarygodnych”. Węzeł źródłowy następnie dzieli się na dwa nowe węzły: obiekty dla których kwota kredytu jest mniejsza od 2323,31zł zostają tymczasowo uznane za wiarygodne, pozostałym nie przyznano kredytu.



Rys. 2. Wybrane drzewo klasyfikacyjne
Źródło: obliczenia własne na podstawie STATISTICA.

Węzeł nr 2 dzieli się według pierwszej wpłaty na dwa podwężły o numerach 4 i 5. Obiekty o pierwszej wpłacie mniejszej od 89,5 zł zostają przypisane do węzła nr 4, który jest podzielony na dwa kolejne węzły ze względu na płeć itd. Obiekty nie spełniające tego warunku są przypisane do węzła nr 5, który nie podlega dalszemu podziałowi i kwalifikuje obiekty mu przypisane jako wiarygodnych kredytobiorców (por. rys. 2 i tab. 4).

Reasumując, wiarygodni kredytobiorcy to klienci, dla których:

- kwota kredytu nie przekracza 2323,31 zł oraz pierwsza wpłata jest większa od 89,5 zł,
- kwota kredytu nie przekracza 2323,31 zł, pierwsza wpłata jest mniejsza od 89,5 zł i klient jest kobietą,
- kwota kredytu nie przekracza 2323,31 zł, pierwsza wpłata jest mniejsza od 89,5 zł, cena telewizora przekracza 819, zł, klient jest mężczyzną i ma więcej niż 41,5 lat.

Po dokonaniu oceny zaprezentowanych drzew z użyciem błędów klasyfikacji otrzymaliśmy wyniki przedstawione w tab. 5.

Tabela 5. Błędy klasyfikacji w % dla próby uczącej i testowej

Prawdopodobieństwo	Próba ucząca			Próba testowa		
	E1	E2	E	E1	E2	E
0,053	0	100	5,3	0	100	5,3
D1 0,5 (Koszt SK = 0,46; odch. std. = 0,028)	23,36	32,74	14,83	28,05	22,10	22,38
D2 0,5 (Koszt SK = 0,39; odch. std. = 0,028)	27,46	35,12	17,13	20,73	21,50	21,47

Źródło: obliczenia własne na podstawie STATISTICA.

W przypadku 50% prawdopodobieństwa przynależności do każdej z grup wiarygodności drzewo klasyfikacyjne dla próby uczącej daje ogólny błąd klasyfikacji dla pierwszego drzewa 14,83%, a dla próby testowej jest on nieznacznie wyższy – wynosi 22,38%; a dla drugiego drzewa odpowiednio – 17,13 oraz 21,47%.

Skonstruowane drzewa dają gorsze wyniki niż strategia przyznania wszystkim kredytu. Nie powinny być one zatem wykorzystywane do oceny wiarygodności, ale mogą być przydatne w charakterystyce klientów zagrożonych ryzykiem niewypłacalności.

5. Podsumowanie

Zaprezentowana metoda drzew klasyfikacji (dla prawdopodobieństwa określonego na podstawie próby) wszystkie wnioski kredytowe oceniła jako „wiarygodne”, błąd drugiego rodzaju był równy 100%. W przypadku równego prawdopodobieństwa przynależności do każdej z grup drzewo klasyfikacji dało wysokie błędy ogólnej klasyfikacji. Najlepsze drzewo klasyfikacyjne dla próby testowej daje błąd klasyfikacji równy 21,47%. Dla próby uczącej błąd ten wynosi 17,13%.

Mimo wysokich błędów klasyfikacji, drzewa mogą być wykorzystywane w ocenie ryzyka kredytowego, gdyż pozwalają na identyfikację kredytobiorców. Za stosowaniem drzew decyzyjnych przemawia również to, iż wyraźnie identyfikują one cechy niewiarygodnego kredytobiorcy. Do tych cech należy zaliczyć:

- młody wiek kredytobiorcy przy wysokich kwotach kredytu,
- wiek mężczyzny poniżej 40 roku życia przy niskich kwotach kredytu.

Wydaje się, że podstawowym problemem w ocenie wiarygodności na podstawie tej próby jest brak danych opisujących dany obiekt i z logicznego punktu widzenia istotnie wpływających na ocenę wiarygodności, np. miesięcznych dochodów, liczby osób w rodzinie itp.

Literatura

- Breiman L., Friedman J.H., Olshen R.A., Stone C.J., *Classification and Regression Trees*, The Wadsworth & Brooks/Cole Statistics/Probability Series 1984.
- Gatnar E., *Symboliczne metody klasyfikacji danych*, PWN, Warszawa 1998.
- Kim H., Loh W.Y., *Classification Trees with Bivariate Linear Discriminant Node Models*, „Journal of Computational and Graphical Statistics” 2003, 12, s. 512-530.
- Loh W. Y., Shih Y. S., *Split Selection Methods for Classification Trees*, „Statistica Sinica” 1997, 7, s. 815-840.
- Mazur A., Staniec I., *Metody klasyfikacji*, materiały konferencyjne, VII Krajowa Konferencja nt. „Komputerowe wspomaganie badań naukowych”, Polanica Zdrój 2000, s. 171-176.

- Misztal M., *Wykorzystanie drzew klasyfikacyjnych do wspomagania procesów podejmowania decyzji*, Materiały na seminaria StatSoft Polska Sp. z o.o., Kraków 2000.
- Pluta W., *Wielowymiarowa analiza porównawcza w badaniach ekonometrycznych*, PWE, Warszawa 1977.
- Staniec I., *Analiza dyskryminacyjna w ocenie wiarygodności kredytobiorcy*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1022, AE, Wrocław 2004, s. 87-95.
- StatisticaTM PL*, t. 3, StatSoft, 1997.
- Szmit A., Szmit M., Kaniewski M., *Analiza ryzyka kredytowego dla potrzeb pośrednika finansowego*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 988, AE, Wrocław 2003, s. 163-172.

CLASSIFYING TREES IN EVALUATION OF THE INBORROWER'S CREDITWORTHINESS

Summary

The evaluation by use classifying tree of the credit risk has been presented in this article. It is based on data gained from „Gwarant” company for individual credits on purchase of television set. The test contained 4746 credit contracts. Among them 250 contracts (5,268 %) were not repaid and 4496 contracts (94,732%) were paid off. Therefore a mischievousness was only 5,268 % in analysing test.