

Mariola Chrzanowska

Wyższa Szkoła Ekonomii i Administracji w Kielcach

ZASTOSOWANIE WYBRANYCH MODELI ZMIENNYCH JAKOŚCIOWYCH DO KLASYFIKACJI KREDYTOBIORCÓW

1. Wstęp

Większość klientów chętnie kupuje według zasady: „korzystam teraz, zapłacę później”. Częstą konsekwencją takiej postawy jest niekontrolowany wzrost zadłużenia klientów. Pojawiają się systematycznie rosnące straty przedsiębiorstw z tytułu nieściągalnych wierzytelności. Priorytetem dla kredytodawców staje się skuteczne zarządzanie ryzykiem i badanie wypłacalności oraz wiarygodności finansowej ich klientów [InFoscore 2003, s. 1]. Polskie doświadczenia ze złymi kredytami w pełni potwierdzają tę zasadę [Kurkliński, 1994, s. 1]. Dlatego ważnym czynnikiem w pracy z klientem – potencjalnym kredytobiorcą – jest badanie jego zdolności kredytowej oraz po udzieleniu kredytu monitorowanie spłaty zadłużenia [Matuszyk 2004, s. 30, 36]. Do analizy zdolności kredytowej osoby fizycznej bank musi dysponować odpowiednimi danymi statystycznymi i wskaźnikami pozwalającymi na ocenę kredytobiorcy. Najczęściej stosuje się metodę określaną jako *credit-scoring* [Jaworski, Krzyżkiewicz, Koniński 1999, s. 237]. Ocena punktowa to wiarygodne i skuteczne narzędzie, wykorzystywane m.in. do oceny ryzyka związanego z nowymi i obecnymi klientami oraz zwiększenia możliwości identyfikacji potencjalnych klientów z akceptowalnym poziomem ryzyka [www.bik.pl...]. Kredytowa ocena punktowa jest statystyczną metodą służącą do określenia prawdopodobieństwa terminowej spłaty kredytu przez klienta banku. Oceny takiej dokonuje się z użyciem kart punktowych, dzięki którym otrzymujemy wynik dla kredytobiorcy poddawanego ocenie. Najwyższy możliwy wynik liczbowy przypisywany jest klientom, w przypadku których zachodzi najwyższe prawdopodobieństwo niezakłóconej spłaty zadłużenia, natomiast niskie wyniki przypisywane są klientom, w przypadku, których istnieje duże prawdopodobieństwo wystąpienia trudności w obsłudze zadłużenia. Ma to na celu maksymalne roz-

różnienie pomiędzy kredytobiorcami obsługującymi zadłużenie w sposób prawidłowy i kredytobiorcami obsługującymi zadłużenie nieprawidłowo [www.bik.pl...].

2. Opis danych

W pracy przedstawiono badanie spłaty kredytu przez osoby, które wzięły kredyty okolicznościowe w banku X w marcu 2004 r. Są to osoby, które przeszły wstępną kwalifikację. Analizie poddano próbę złożoną ze 162 osób. Według stanu w końcu lipca 2004 r. 82 osoby regularnie spłacają kredyt, a 80 kredytobiorców zalega ze spłatami. Kredytobiorcy zostali scharakteryzowani za pomocą następujących cech: płeć; kwota kredytu (w zł); zawód (rolnik, pracownik, biznesmen); wynagrodzenie (netto); okres spłaty (w latach); liczba osób na utrzymaniu; stan cywilny (żonaty/zamężna, wolny); wiek; terminowa spłata (tak, nie)

W pierwszym etapie badania istotność zależności pomiędzy zmienną objaśnianą i potencjalnymi zmiennymi objaśniającymi. Efektem było przyjęcie do dalszej analizy tylko tych potencjalnych zmiennych objaśniających, które okazały się istotnie współzależne ze zmienną egzogeniczną. Następnie dla wybranych zmiennych zbudowano kilka modeli logitowych i probitowych, spośród których wybrano te, które dokonały najlepszej klasyfikacji.

3. Model logitowy

Modele logitowe stosuje się wówczas, gdy nie są spełnione założenie niezbędne dla regresji liniowej [Janc, Kraska 2001, s. 74]. Są to modele regresyjne, opisujące relację między zmienną wynikową-dychotomiczną a zmiennymi objaśniającymi [Dąbrowski 2003, s. 9]. Model logitowy zapisać można w postaci [Gruszczyński 2001, s. 17, 58]:

$$\ln \frac{P_i}{1 - P_i} = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (1)$$

gdzie: $\boldsymbol{\beta}$ – wektor parametrów modelu,

$\mathbf{x}_i$ – wektor wartości zmiennych egzogenicznych,

$\mathbf{x}_i^T \boldsymbol{\beta}$ – wskaźnik opisujący i -tego kredytobiorcę,

P_i – prawdopodobieństwo¹ przyjmowania przez binarną zmienną y wartości 1.

Wzór (1) można zapisać jako [Borkowski, Dudek, Szczesny 2003, s. 212]:

$$y = \ln \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ki} + \varepsilon_i \dots \quad (2)$$

¹ W tym przypadku P_i oznacza prawdopodobieństwo spłaty kredytu.

Zmienna zależna y jest zmienną binarną (zero-jedynkową), a prawdopodobieństwo, że $y = 1$ określa się za pomocą rozkładu logistycznego. Dla i -tej obserwacji $y_i = 1$ przyjęto, że kredyt jest spłacany w terminie, a $y_i = 0$ oznacza, że klient zalega ze spłatą kredytu.

Logitem nazywamy logarytm ilorazu szans przyjęcia oraz nieprzyjęcia wartości 1 przez zmienną y . Jeśli szanse obu zdarzeń są jednakowe ($P(y_i = 1) = P(y_i = 0) = 0,5$) to logit równa się zero. Logit może przyjmować dowolną wartość ze zbioru liczb rzeczywistych, przy czym dla $P_i > 0,5$ przyjmuje wartość ujemną, a dla $P_i < 0,5$ dodatnią [Gruszczyński 2001, s. 19].

4. Model probitowy

W modelu probitowym nieobserwowalną wielkość prawdopodobieństwa P_i można zapisać jako wartości dystrybuanty rozkładu normalnego $N(0,1)$ w punktach $\mathbf{x}_i^T \boldsymbol{\beta}$.

Uzyskany w ten sposób model probitowy ma postać [Cramer 2003, s. 23; Chow 1995, s. 309]:

$$P_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mathbf{x}_i^T \boldsymbol{\beta}} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{t^2}{2}\right) dt = F(\mathbf{x}_i^T \boldsymbol{\beta}) \quad (3)$$

Wartości kombinacji liniowej $\mathbf{x}_i^T \boldsymbol{\beta}$ dla określonego poziomu prawdopodobieństwa (przy założeniu, że $\mathbf{x}_i^T \boldsymbol{\beta} \sim N(0; 1)$) nazywa się profitami lub normitami. Są to wartości funkcji odwrotnej do F [Gruszczyński 2001, s. 18].

Historycznie model probitowy poprzedza model logitowy, jednak ze względu na większą użyteczność częściej wykorzystuje się model logitowy. Do oceny modelu probitowego stosuje te same miary dopasowania jak w modelu logitowym.

5. Mierniki dopasowania modelu logitowego i probitowego

Jednym z kryteriów oceny modelu logitowego jest ocena wiarygodności uzyskanych wyników. Miarą wiarygodności jest ujemny, dwukrotny logarytm wiarygodności wyników $-2\ln L$. Dla modelu doskonałego (pełnego), którego wiarygodność jest równa 1, wartość wyrażenia $-2\ln L = 0$. Poprawność rozwiązania oceniana jest za pomocą wzoru:

$$\chi^2 = -2\ln L_{UR} - [-2\ln L_R] \quad (4)$$

gdzie: L_{UR} oznacza wartość funkcji wiarygodności dla pełnego modelu,

L_R oznacza wartość funkcji wiarygodności dla modelu tylko z wyrazem wolnym.

Dla dużych prób różnica ta ma rozkład zbliżony do rozkładu chi-kwadrat o k stopni swobody. Duża wartość statystyki chi-kwadrat, tj. powyżej wartości odczy-

tanej z tablic, wskazuje na istotność podanego rozwiązania (por. [Rószkiewicz 2002, s. 177; Cramer 2003, s. 38]).

Inną miarą oceny modelu logitowego jest tzw. statystyka Walda. Bada ona istotność poszczególnych współczynników modelu. Określona jest za pomocą wzoru (por. [Rószkiewicz 2002, s. 177; Cramer 2003, s. 38; Dobosz 2001, s. 257]):

$$W = \left(\frac{b_i}{s(b_i)} \right)^2, \quad (5)$$

gdzie: b_i – oszacowana wartość parametru stojącego przy i -tej zmiennej,
 $s(b_i)$ – błąd standardowy.

Jak wynika ze wzoru (5), im wyższa wartość statystyki Walda, tym mocniejsze są podstawy do uznania oszacowanego parametru za istotny. Wartość statystyki teoretycznej odczytuje się z rozkładu chi-kwadrat dla 1 stopnia swobody.

Inną miarą jakości (modelu logitowego) jest tzw. iloraz szans [Gruszczyński 2001, s. 83], który wyznacza się jako:

$$\text{iloraz_szans} = \frac{\text{iloczyn_poprawnie_zakwalifikowanych_przypadków}}{\text{iloczyn_niepoprawnie_zakwalifikowanych_przypadków}}. \quad (6)$$

Kolejną miarą oceny jakości modelu jest zliczeniowy R^2 , będący miarą analogiczną do współczynnika determinacji stosowanego w MNK:

$$R^2 = \frac{\text{liczba_przypadków_zaklasyfikowanych_poprawnie}}{\text{łączna_liczba_przypadków}}. \quad (7)$$

6. Dobór zmiennych do modelu

W modelu posłużono się kryteriami stosowanymi w literaturze. Przyjęto, że do budowy modelu będą użyte zmienne silnie skorelowane między sobą i jednocześnie słabo skojarzone z innymi zmiennymi egzogenicznymi [Borkowski, Dudek, Szczesny 2003, s. 63]. Dobór odbywa się na podstawie analizy macierzy, obrazującej poziom i kierunek współzależności między wszystkimi możliwymi parami zmiennych X i Y . Jeśli co najmniej jedna ze zmiennych w parze jest zmienną zero-jedynkową, wówczas współczynnik korelacji należy zastąpić odpowiednią miarą asocjacji [Gruszczyński 1999]. Najczęściej stosuje się:

- współczynnik skojarzenia Yule'a, jeśli obie zmienne są binarne²,
- test t dla różnicy średnich, w przypadku gdy jedna ze zmiennych jest zmienną zero-jedynkową, a drugą jest zmienną mierzalną na skali przedziałowej,
- współczynnik korelacji Persona, jeśli obie zmienne są ilościowe.

² Szczegółowy opis procedury znajduje się w [Gruszczyński 2001].

Wykonany test t różnicy średnich, wykazał istotne skorelowanie zmiennej zależnej ze zmienną *kwota*. Istotne skorelowanie ze zmienną *terminowa spłata* (obliczone za pomocą współczynnika skojarzenia Yule'a) wykazały zmienne *pleć* oraz *stan cywilny*. Zależność między zmienną egzogeniczną a zmiennymi opartymi na skali nominalnej badana była za pomocą testu niezależności chi-kwadrat. W tej procedurze istotna zależność ze zmienną grupującą wykazała zmienna *zawód*. Do budowy modelu użyto następujących kombinacji zmiennych:

- *pleć* + *kwota_kredytu* + *stan cywilny*,
- *pleć* + *zawód* + *stan cywilny*,
- *pleć* + *kwota_kredytu*,
- *pleć* + *stan cywilny*,
- *kwota* + *stan cywilny*,
- *zawód* + *stan cywilny*.

7. Wyniki badań – modele logitowe

Spośród wszystkich estymowanych modeli najlepszą klasyfikacją obu grup kredytobiorców wykazały modele oparte na drugiej i szóstej kombinacji zmiennych.

Model logitowy I

$$\hat{y} = -346,91 + 2,21 \cdot \text{plec} + 0,95 \cdot \text{zawod} + 0,25 \cdot \text{stan_cywilny}, \quad (8)$$

Wartość funkcji wiarygodności = 88,304307600,

$\chi^2(3) = 47,946$ $p = 0,00000$,

Iloraz szans = 5,4174.

Tabela 1. Wyniki klasyfikacji kredytobiorców – model I

Terminowa spłata	Przewidywane		Procent poprawnie sklasyfikowanych klientów
	nie	tak	
Nie	69	11	86,25
Tak	44	38	46,34

Źródło: opracowanie własne.

Wyniki estymacji modelu I (tab. 1) potwierdzają jego przydatność do klasyfikacji. Wysoka wartość chi-kwadrat (dla której prawdopodobieństwo błędnego odrzucenia hipotezy zerowej o nieistotności wszystkich parametrów modelu jest równe 0) świadczy o lepszej jakości tego modelu niż modelu, który zawiera tylko wyraz wolny. Innymi słowy – klasyfikacja uzyskana za pomocą tego modelu jest lepsza od klasyfikacji przypadkowej (co potwierdza wysoka wartość ilorazu szans).

Model logitowy II

$$\hat{y} = -115 + 1 \cdot \text{zawod} + 0,55 \cdot \text{stan_cywilny}, \quad (9)$$

Wartość funkcji wiarygodności = 101,54387853,

$$\text{Chi}^2(2) = 21,467 \quad p = 0,00002.$$

Iloraz szans = 4,1892.

Wartość statystyki chi-kwadrat w tym modelu wskazuje (tak jak w poprzednich prezentowanych modelach) na istotność oszacowanych parametrów. Wartość ilorazu szans większa od jedynki potwierdza, że klasyfikacja uzyskana za pomocą modelu III jest lepsza od przypadkowej (tab. 2).

Tabela 2. Wyniki klasyfikacji kredytobiorców – model II

Terminowa spłata	Przewidywane		Procent poprawnie sklasyfikowanych klientów
	nie	tak	
Nie	62	18	77,50
Tak	37	45	54,88

Źródło: opracowanie własne.

Tabela 3. Wyniki klasyfikacji kredytobiorców – model III

Terminowa spłata	Przewidywane		Procent poprawnie sklasyfikowanych klientów
	nie	tak	
Nie	46	34	57,50
Tak	38	44	53,66

Źródło: opracowanie własne na podstawie obliczeń w programie Statistica PL.

Model I poprawnie przypisał 86% przypadków z grupy niewywiązującej się terminowo z zobowiązań. W modelu II uzyskano 77,5% prawidłowo zakwalifikowanych osób z grupy niespłacającej terminowo oraz 54,88 % osób spośród wszystkich terminowo spłacających (tab. 2). Łączna trafność prognozy (*ex post*) dla omawianych modeli wynosi 66% ($R^2 = 0,66$), co dla danych jakościowych uznaje się za poprawne.

8. Wyniki badań – modele probitowe

Klasyfikacja danych za pomocą modeli probitowych okazała się znacznie trudniejsza niż klasyfikacja modeli logitowych. Zastosowanie kombinacji zmiennych wykorzystanych przy budowie modeli logitowych okazało się niemożliwe. Parametry modelu albo nie były możliwe do oszacowania za pomocą MNW albo wartość statystyki chi-kwadrat oraz statystyki Walda wskazywała na nieistotność oszacowanych parametrów modelu. W tej sytuacji zdecydowano się zbudować modele probitowe złożone tylko z jednej zmiennej, która jest istotnie powiązana ze zmienną egzogeniczną. Niestety wyniki klasyfikacji za pomocą modeli probitowych są gorsze niż modeli logitowych. W związku z tym w opracowaniu przedstawiono najlepszy model spośród uzyskanych modeli probitowych.

Tabela 4. Wyniki estymacji modelu probitowego

	Ocena parametru	Błąd standardowy	Statystyka Walda	Prawdopodobieństwo
Wyraz wolny	-0,469740	0,20366	5,3201	0,02236
Wiek	0,011825	0,004088	8,366331	0,00435
Wartość funkcji wiarygodności = 111,22839930 $\chi^2(1) = 3,0982$ $p = 0,04$ Iloraz szans = 1,56				

Źródło: opracowanie własne na podstawie obliczeń w programie Statistica PL.

Uzyskany za pomocą metody największej wiarygodności model probitowy jest poprawny. Wartość statystyki chi-kwadrat świadczy o stosunkowo wysokim prawdopodobieństwie odrzucenia hipotezy zerowej o nieistotności oszacowanych parametrów. Potwierdzają to wartości statystyki Walda dla zmiennej grupującej i wyrazu wolnego.

Model III poprawnie przypisał 46 osób (57,5%) spośród osób niespłacających oraz 44 osoby (53,66%) spośród osób, które spłacają kredyt regularnie. Trafność prognozy dla modelu probitowego wynosi 55%.

9. Podsumowanie

W tabeli 5 przedstawiono zestawienie wszystkich modeli ze względu na przyjęte kryteria:

Tabela 2. Ocena prezentowanych modeli

	Chi-kwadrat	Iloraz szans	R^2
Model I	47,946	5,4174	66%
Model II	21,467	4,1892	66%
Model III	3,098	1,56	55%

Źródło: opracowanie własne.

Spośród wszystkich prezentowanych modeli najlepszy okazał się model I. Istotność oszacowanych parametrów dla tego modelu potwierdza wysoka wartość statystyki chi-kwadrat. Wysoka trafność łącznej prognozy *ex post* potwierdza tę tezę. Najlepszy spośród prezentowanych modeli jest model probitowy. Potwierdza to niska wartość chi-kwadrat oraz wielkość współczynnika R^2 zbliżona do przypadkowej.

Powstaje pytanie: jakie wnioski płyną z badania dla banku X? Wydaje się, że inspektorzy kredytowi powinni zwracać większą uwagę na cechy jakościowe przyszłych kredytobiorców. Są one bardziej stabilne niż cechy o charakterze mierzalnym.

Prezentowane w pracy modele logitowe i model probitowy są jednymi z bardziej znanych modeli jakościowych. W pracy omówiono jedynie wybrane aspekty analizy tych modeli. Choć wyniki badania można uznać jedynie za zadowalające, należy pamiętać, że cechy uwzględnione w badaniu miały przede wszystkim charakter jakościowy. W modelach zbudowanych głównie za pomocą cech jakości-

wych trudno uzyskać wyniki porównywalne z klasyfikacją wyłącznie za pomocą ciągłych zmiennych mierzalnych.

Literatura

- Borkowski B., Dudek H., Szczęśny W., *Ekonometria. Wybrane zagadnienia*, PWN, Warszawa 2003.
- Chow G.C., *Ekonometria*, PWN, Warszawa 1995.
- Cramer J.S., *Logit Models from Economics and Other Fields*, Cambridge University Press, Cambridge 2003.
- Cramer J.S., *An Introduction to the Logic Model for Economists*, Timberlake Consultants Ltd, London 2001.
- Dobosz M., *Statystyczna analiza wyników badań*, EXIT, Warszawa 2001.
- Dąbrowski A., *Analiza danych jakościowych*, www.math.uni.wroc.pl/~dabr/dydaktyka.html (30 czerwca 2003 r.).
- Gruszczyński M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Monografie i Opracowania nr 490, SGH, Warszawa 2001.
- Gruszczyński M., *Scoring logitowy w praktyce bankowej a zagadnienie koincydencji*, „Bank i Kredyt” 1999, nr 5.
- InFoscore, listopad 2003.
- Janc A., Kraska M., *Credit-scoring. Nowoczesna metoda oceny zdolności kredytowej*, Biblioteka Menadżera i Bankowca, Warszawa 2001.
- Jaworski W.L., Krzyżkiewicz Z., Koniński B., *Banki – rynek, operacje, polityka*, Warszawa 1999.
- Kurkliński L., *Polityka kredytowa banków wobec podmiotów gospodarczych*, *Bankowe ABC* nr 4, „Bank i Kredyt”, sierpień 1994.
- Matuszyk A., *Credit Scoring*, CeDeWu, Warszawa 2004.
- www.bik.pl/biknewdom/ocena_punktowa.htm (2 sierpnia 2004 r.).
- Rószkiewicz M., *Metody ilościowe w badaniach marketingowych*, PWN, Warszawa 2002.

APPLICATION OF SELECTED MODELS FOR THE BORROWERS' CLASSIFICATION

Summary

The work presents research on paying off the occasional credits received in March 2004. The sample consists of 162 borrowers. According to the data from the end of July 2004, 82 borrowers pay off the credit regularly and 80 fail to pay off the credit. The aim of the report is to prepare a dichotomic classification of borrowers with the use of the logit and probit models.