Jerzy Marzec[1]

ZDOLNOŚCI DYSKRYMINACYJNE MODELU DWUMIANOWEGO ZE SKOŃCZONĄ MIESZANKĄ ROZKŁADÓW NORMALNYCH W OCENIE NIESPŁACALNOŚCI KREDYTÓW

1. Wprowadzenie

Jednym z podstawowych obszarów działalności tradycyjnego banku jest transfer środków pieniężnych pozyskanych od deponentów, a skierowanych do kredytobiorców. W ostatnich latach obserwuje się w Polsce gwałtowany wzrost wartości udzielanych kredytów detalicznych. Od maja 2005 r. zadłużenie gospodarstw domowych w bankach i monetarnych instytucjach finansowych działających w Polsce jest wyższe od zadłużenia przedsiębiorstw[2]. Towarzyszy temu relatywnie mniejszy wzrost wielkości depozytów gospodarstw domowych, a w efekcie stosunek wartości zaciągniętych kredytów do wielkości posiadanych przez nich środków (w formie lokat bankowych, na rachunku bieżącym) systematycznie rośnie. W marcu 2004 roku ten iloraz wynosił prawie 0,5, zaś we wrześniu 2007 roku kształtował się na poziomie 0,98, więc można się spodziewać, że już wkrótce przekroczy wartość jeden.[3] Obserwując gwałtowny rozwój tego zjawiska warto postawić pytanie, czy tej szerokiej dostępności kredytów zgłaszanej przez banki towarzyszy rzetelna kontrola bieżącej i przyszłej zdolności kredytobiorców do spłaty zaciągniętych zobowiązań? Umowa kredytowa nakłada na kredytobiorcę obowiązek spłaty w wyznaczonych momentach czasu rat kapitałowo-odsetkowych. Niedotrzymanie tego warunku ma skutki dla obydwu stron. Wówczas kredytobiorca ma do zapłacenia karne odsetki według wyższej stopy niż oprocentowanie kredytu, zaś na banku spoczywa obowiązek tworzenia rezerw celowych w związku z zaistniałym ryzykiem utraty środków pieniężnych, które powierzył mu deponent. Wzrost realnego kosztu kredytu pogarsza sytuację finansową kredytobiorcy, więc będzie mu trudno uzyskać kolejną pożyczkę albo otrzyma ją na bardzo niekorzystnych warunkach, np. po dużo wyższej cenie. Natomiast dla banku niespłacone kredyty oznaczają obniżenie wskaźnika wypłacalności, czyli ograniczenie jego dalszej działalności kredytowej. W skrajnym przypadku może nastąpić niewypłacalność gospodarstwa domowego, co dla banku oznacza przynajmniej częściową utratę zainwestowanego kapitału. Proces odzyskiwania pełnej kwoty swoich wierzytelności może być długi i kosztowny. Zatem w interesie

[1] Dr Jerzy Marzec – Katedra Ekonometrii, Uniwersytet Ekonomiczny w Krakowie, e-mail: marzecj@uek.krakow.pl.
Artykuł powstał w ramach badań statutowych (nr 2/KE/2/08/S/419) finansowanych przez Uniwersytet Ekonomiczny w Krakowie.
[2] Źródło: dane z NBP o należnościach i zobowiązaniach monetarnych instytucji finansowych.

zarówno banku, jak i gospodarstw domowych jest, aby ryzyko związane ze spłatą rat kapitałowo-odsetkowych było jak najmniejsze. Istnieją narzędzia, które umożliwiają pomiar ryzyka kredytowego związanego z pojedynczą umową kredytową, a następie jego minimalizację. W tym celu banki stosują scoring kredytowy [A. Janc, M. Kraska 2001]. Podstawowymi narzędziami stosowanymi przy ocenie wniosków kredytowych są modele statystyczne (logitowy, probitowy, analiza dyskryminacyjna, drzewa klasyfikacyjne), sieci neuronowe i programowanie matematyczne; [L. Thomas 2000, L. Thomas, R. Oliver, D. Hand 2005, E. Rosenberg, A. Gleit 1994]. Ideą tych modeli jest wykorzystanie historycznych informacji o klientach banku (o charakterze społeczno-demograficznym, o sposobie dotychczasowej współpracy, przebiegu spłaty kredytu itp.) w celu prognozowania poziomu ryzyka w przypadku bieżących wniosków kredytowych dotychczasowych lub nowych klientów.

W Polsce są prowadzone badania empiryczne dotyczące oceny niespłacalności kredytów. Wyniki tych badań przedstawiają m.in.: Chrzanowska, Kompa i Witkowska [2005], Chrzanowska i Witkowska [2005], Misztal [2006], Staniec [2000, 2005], Staniec i Szmit [2004a,b, 2005], Szmit, Szmit i Kaniewski [2003], Staniec i Witkowska [2002], Witkowska i Chrzanowska [2004, 2006, 2007]. W zdecydowanej większości stosuje się w tych pracach sieci neuronowe, drzewa klasyfikacyjne bądź analizę dyskryminacyjną. Z modelu logitowego korzysta się rzadko, wyłącznie w celach porównawczych, zaś model probitowy jest całkowicie pomijany. Wydaje się, że w literaturze polskiej zaawansowane modele danych jakościowych nie znajdują zrozumienia, skoro zdarza się, jak w artykule Szmit, Szmit i Kaniewski [2003], że analizę spłacalności kredytów przeprowadza się na podstawie regresji liniowej. Powyższe prace mają charakter czysto empiryczny, główny nacisk jest kładziony na porównywanie wyników klasyfikacji otrzymanych za pomocą wspomnianych metod. Przykładem innych badań, poświęconych dostępności kredytów dla polskich przedsiębiorstw, jest praca I. Tymoczko i M. Pawłowska [2007]. Innym obszarem zastosowań w ekonomii modeli dwumianowych jest prognozowanie upadłości przedsiębiorstw [E. Mączyńska M. Zawadzki 2006, D. Wędzki 2005].

Celem niniejszego artykułu jest prezentacja modelu dwumianowego (dla zerojedynkowej zmiennej endogenicznej) opartego na skończonej mieszance rozkładów normalnych. Stanowi on uogólnienie modelu probitowego, który obok modelu logitowego jest najbardziej znaną konstrukcją dla jakościowej zmiennej endogenicznej[4]. W warstwie empirycznej przedstawimy wykorzystanie wspomnianego modelu do oceny niespłacalności kredytów detalicznych, czyli jego zastosowanie w bankowym scoringu kredytowym. Dokonamy porównania wyników uzyskanych na podstawie

[3] Jednym ze źródeł niższego tempa wzrostu depozytów gospodarstw domowych jest alokacja ich oszczędności w inne formy inwestycji finansowych (np. w bony skarbowe, obligacje, jednostki uczestnictwa w funduszach inwestycyjnych, indywidualne konto emerytalne).

próby podstawowej z rezultatami dla próby odłożonej. Intencją autora jest pokazanie, iż model dwumianowy oparty na mieszance jest lepszy wobec tradycyjnych narzędzi statystycznych wykorzystywanych w scoringu kredytowym: modelu probitowego i logitowego.

2. Model dwumianowy z mieszanką rozkładów normalnych

Jednym z możliwych kierunków uogólnienia modelu probitowego – podstawowej konstrukcji w przypadku modeli danych jakościowych – jest zastosowanie skończonej mieszanki rozkładów normalnych. Jej wykorzystanie w celu aproksymacji nieznanej postaci funkcji gęstości składnika losowego pozwala uwzględnić takie własności jak: wielomodalność, asymetrię oraz zachowanie się tego rozkładu w otoczeniu wartości modalnej w odniesieniu do rozkładu normalnego - wyostrzenie albo spłaszczenie. Mieszanki dają szeroki wachlarz możliwości, gdy klasyczne założenia nie są spełnione, np. w sytuacji niejednorodnej próby, występowania heteroscedastyczności. Ich użycie jest bardzo interesującym podejściem, zarówno z punktu widzenia metodyki, jak i badań empirycznych. Literatura poświęcona modelom statystycznym opartym na mieszankach jest bogata, [D. Titterington, A. Smith, U. Makov 1985], ale mimo to rzadko spotyka się ich zastosowanie w modelach zmiennych jakościowych. Zatem warto wspomnieć o takich przykładach jak P. Austin M. Escobar [2002], A. Erkanli D. Stang P. Müller [1993], S. Frühwirth-Schnatter R. Frühwirth [2006], J. Geweke M. Keane [1999], P. Qu Y. Qu [2000], Marzec [2008].[5] Wykorzystanie skończonej mieszanki rozkładów normalnych stanowi inny kierunek uogólnienia modelu probitowego, w odniesieniu do propozycji wykorzystania rozkładu *t* Studenta [Marzec 2006]. Oba te uogólnienia stanowią ciekawą propozycję metodologiczną także w odniesieniu do modeli służących do prognozowania upadłości przedsiębiorstw.

Przedmiotem rozważań jest model dwumianowy

$$y_t = \begin{cases} 1 & \text{dla} \quad z_t \geq 0 \\ 0 & \text{dla} \quad z_t < 0 \end{cases} \quad \text{gdzie} \quad z_t = \boldsymbol{x}_t \cdot \boldsymbol{\beta} + \varepsilon_t, \qquad (1)$$

zaś x_t jest wektorem k zmiennych egzogenicznych (lub ich znanych funkcji) charakteryzujących obserwację o numerze t, $\boldsymbol{\beta}$ jest wektorem k nieznanych parametrów. Kluczowe założenie mówi, że składniki losowe ε_t posiadają identyczne i niezależne rozkłady określone przez skończoną mieszankę rozkładów normalnych.

Funkcja gęstości i dystrybuanta jednowymiarowej zmiennej losowej ε_t jest określona przez wypukłą kombinację funkcji gęstości i dystrybuant składowych mieszanki. Jeżeli przez $p_j\left(\cdot \middle| \mu_j, \tau_j\right)$

[4] Przypomnijmy, iż rozkłady normalny i logistyczny są jednomodalne i symetryczne.
[5] We wspomnianych pracach, podobnie jak w niniejszym artykule, estymację parametrów mieszaki przeprowadzono na gruncie wnioskowania bayesowskiego.

oznaczymy funkcję gęstości zmiennej losowej o rozkładzie normalnym, o wartości oczekiwanej μ_j i precyzji τ_j (odwrotności wariancji), zaś π_j są wagami, tzn. $\pi_j \in (0;1)$ i $\sum_{j=1}^{J} \pi_j = 1$, to rozkład zmiennej losowej ε_t jest określony przez gęstość

$$p(\varepsilon_t|\boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j \cdot p_j(\varepsilon_t|\mu_j, \tau_j), \tag{2}$$

gdzie $J>1$ jest liczbą składników mieszanki, $\boldsymbol{\theta}$ to wektor kolumna zawierający wszystkie parametry μ_j, τ_j i π_j dla $j=1,\ldots,J$. Zaprezentowana struktura jest mieszanką zarówno względem średniej, jak i parametru skali; zob. J. Geweke M. Keane 1999.

W niniejszym artykule rozważamy model dychotomiczny, więc funkcja wiarygodności ma postać

$$p(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{t=1}^{T} \left(1 - F(-\boldsymbol{x}_t\boldsymbol{\beta}|\boldsymbol{\theta})\right)^{y_t} F(-\boldsymbol{x}_t\boldsymbol{\beta}|\boldsymbol{\theta})^{1-y_t}, \tag{3}$$

gdzie dystrybuanta zmiennej losowej ε_t w punkcie $-\boldsymbol{x}_t\boldsymbol{\beta}$, $F(-\boldsymbol{x}_t\boldsymbol{\beta}|\boldsymbol{\theta})$, jest określona formułą

$$F(a|\boldsymbol{\theta}) = \Pr(\varepsilon_t < a) = \sum_{j=1}^{J} \pi_j \cdot F_j(a|\mu_j, \tau_j), \tag{4}$$

przy czym $F_j(a|\mu_j, \tau_j)$ oznacza wartość w punkcie a dystrybuanty zmiennej o rozkładzie normalnym, określonym przez parametry μ_j i τ_j. Zatem prawdopodobieństwo zdarzenia $y_t=1$, p_t, jest równe $1 - F(-\boldsymbol{x}_t\boldsymbol{\beta}|\boldsymbol{\theta})$.

W przypadku modeli opartych na mieszankach fundamentalnym zagadnieniem jest identyfikowalność parametrów. W celu jej zapewnienia wymaga się, aby zostały uwzględnione restrykcje $\mu_{j-1}<\mu_j$, albo $\tau_{j-1}<\tau_j$, albo $\pi_{j-1}<\pi_j$ dla $j=2,\ldots,J$; zob. [J. Diebolt J. C.P. Robert 1994, J. Geweke 2007, J. Geweke M. Keane 1999, G. Koop 2003 i P. Qu Y. Qu 2000]. Restrykcje te są konsekwencją własności wielomodalności funkcji wiarygodności, a w literaturze zagadnienie to nosi nazwę *label switching problem*. Z uwagi na własności funkcji wiarygodności stosowanie metody największej wiarygodności nie jest wskazane.

Rozważając model dwumianowy musimy uwzględnić dodatkowe warunki identyfikowalności. Analogicznie jak w modelu probitowym, restrykcja dotyczy wyrazu wolnego w równaniu regresji dla zmiennej ukrytej z_t albo wartości oczekiwanej zmiennej ε_t. Najprościej przyjąć, iż brak jest sztucznej zmiennej „1”, gdyż w przeciwnym przypadku należałoby ustalić $\mu_{j^*}=0$ dla jakiegokolwiek j^*. Wymaga się także, aby dla pewnego (dowolnego) j^* spełniony był warunek: $\tau_{j^*}=1$ [J. Geweke M. Keane 1999]. Oba wymienione warunki są niezbędne dla identyfikowalności, gdyż w przeciwnym przypadku istniałoby nieskończenie wiele kombinacji parametrów, dla których funkcja

wiarygodności osiągałaby identyczne maksima. W tym kontekście warto przypomnieć, iż dla J=1 uzyskujemy model probitowy, w którym zakłada się, że τ_j =1 i μ_j =0, gdy występuje wyraz wolny.

W niniejszym artykule rozważyliśmy mieszankę dwuelementową (J=2) i przyjęliśmy, że nie występuje wyraz wolny, zaś restrykcje parametryczne mają postać $\tau_2>\tau_1$ i τ_1=1. Druga równoważna parametryzacja to $\tau_1<\tau_2$=1. W celu estymacji modelu dwumianowego z mieszanką wykorzystano podejście bayesowskie, stosując metody Monte Carlo Markov Chain. Wnioskowanie bayesowskie jest uzasadnionym narzędziem zwłaszcza w sytuacji, gdy badacz posiada silną wiedzę a priori o parametrach. W tym przypadku wynika ona z restrykcji koniecznych dla zapewnienia identyfikowalności parametrów rozważanego modelu. Te zaś uwzględnia się w rozkładzie a priori. Konstrukcja rozkładów a priori, numeryczne metody uzyskiwania rozkładów a posteriori i testowanie modeli są prezentowane m.in. w artykułach Diebolt J. C.P. Robert [1994], J. Geweke [2007], J. Geweke M. Keane [1999], G. Koop [2003], J. Marzec [2008] i K. Roeder L. Wassermann [1997].

3. Konstrukcja danych

Badania empiryczne przeprowadzono o dane obejmujące 39034 rachunków kredytowych osób fizycznych. Zbiór danych zawierał informacje o udzielonych kredytach hipotecznych i konsumpcyjnych, udzielonych w okresie 01.01.2000-30.09.2001 r. Zmienna endogeniczna y_t została zdefiniowana na podstawie kategorii należności, tj. y_t=1 w przypadku należności zagrożonych (poniżej standardu, wątpliwych i straconych), zaś y_t=0 oznacza rachunek zakwalifikowany do kategorii należności normalnych[6]. Ustalenie kategorii ryzyka dla kredytów udzielonych w 2000 r. nastąpiło na podstawie informacji na 30.09.2001 r., zaś udzielonych w 2001 r. - wg stanu na 30.09.2002 r.

Przyjęto, iż okres między momentem udzielenia kredytu a pobraniem informacji o kategorii ryzyka wynosi przynajmniej 9 miesięcy, aby kredytobiorca miał czas na podjęcie decyzji, czy spłacać raty kapitałowo-odsetkowe zgodnie z umową czy nie. Z drugiej strony okres ten był nie dłuższy niż 21 miesięcy, co wynikało z dostępności danych oraz faktu, iż średni okres spłaty kredytu wg umowy kredytowej wynosi 30 miesięcy. Odstąpienie od umowy przez kredytobiorcę oznacza, że raty są spłacane nieregularnie (z pewnym opóźnieniem) albo wcale. Należy podkreślić, iż prezentowana analiza ma charakter statyczny, zaś sam proces spłaty rat kapitałowo-odsetkowych ma charakter dynamiczny. Analiza dynamiczna nie jest w pełni możliwa, gdyż większość zmiennych egzogenicznych, które charakteryzują kredytobiorcę lub udzielony kredyt bankowy nie

[6] Zob. Uchwała nr 8/1999 Komisji Nadzoru Bankowego z 22 grudnia 1999 r.

zmienia się w czasie, z wyłączeniem jego dochodów, które gwałtownie zmieniając się mogą powodować zakłócenia ze spłatą rat kapitałowo-odsetkowych.

Estymację parametrów modelu przeprowadzono na podstawie zbioru liczącego 35611 kredytów detalicznych, które udzielono w okresie 01.01.2000-30.06.2001 r. Natomiast informacje o 3423 rachunkach udzielonych między 01.07.2001 a 30.09.2001 r. potraktowano jako próbę odłożoną. Obie próby wykorzystano do określenia dopasowania rozważanych modeli i ich własności dyskryminacyjnych.

Jako potencjalne zmienne egzogeniczne wyjaśniające ryzyko kredytowe przyjęto, jak we wcześniejszych pracach autora, płeć (x_{t1}), wiek kredytobiorcy (x_{t2}), wpływy na rachunki typu ROR (x_{t3}), posiadanie rachunku ROR (x_{t4}), informację o tym, czy kredytobiorca posiada karty płatnicze lub kredytowe wydane przez badany bank (x_{t5}), sposób udzielenia kredytu (przez pośrednika kredytowego albo bezpośrednio w oddziałach banku, x_{t6}), typ kredytu (konsumpcyjny albo hipoteczny, x_{t7}), okres trwania umowy kredytowej (x_{t8}), kwota przyznanego kredytu (x_{t9}), waluta kredytu (x_{t10}) i podstawowe źródło dochodu uzyskiwanego przez kredytobiorcę. Ostatnia zmienna ma charakter kategorii mierzonej na skali nominalnej, więc rozważając cztery przypadki przyjęto, iż źródłem dochodu jest umowa o pracę, gdy $x_{t11}=x_{t12}=x_{t13}=0$, renta lub emerytura, gdy $x_{t12}=1$ i $x_{t11}=x_{t13}=0$, własna działalność, umowa o dzieło lub umowa zlecenie - $x_{t12}=0$ i $x_{t11}=x_{t13}=0$ albo inne źródło (np. stypendium), gdy $x_{t11}=x_{t12}=0$ i $x_{t13}=1$.

W tabeli 1 przedstawione są podstawowe informacje o rachunkach kredytowych i ich właścicielach w przypadku próby podstawowej oraz odłożonej. Udział złych kredytów ($y_t=1$) w próbie podstawowej wynosi 23%, zaś w drugiej tylko 6,4%. Średnie wartości takich zmiennych jak: płeć, wiek kredytobiorcy, okres i typ kredytu, są na zbliżonym poziomie. W przypadku pozostałych zmiennych objaśniających widoczne są różnice między wartościami średnich dla każdej z prób.

Tabela 1. Podstawowe informacje o danych

Charakterystyka \ Próba	Podstawowa	Odłożona
Liczba rachunków	35611	3423
Udział rachunków o kategorii normalnej ($y_t=0$)	77%	95,6%
Struktura wg płci (udział mężczyzn)	53%	50%
Wiek kredytobiorcy (w latach)	40,30	39,53
Kwartalne wpływy na rachunki ROR (w tys. zł)	7,21	12,99
Posiadający ROR	53%	84%
Posiadający karty płatnicze	31%	54%
Klient/kredyt pozyskany przez pośrednika	41%	7%
Typ kredytu (konsumpcyjny)	94%	88%
Okres kredytu (w latach)	2,61	2,53
Kwota przyznanego kredytu (w tys. zł)	10,06	12,94
Waluta kredytu (udział PLN)	97%	93%

Źródło: Opracowanie własne.

4. Ocena zdolności dyskryminacyjnych – wyniki empiryczne

Oceny parametrów mieszanki i probitu

Na podstawie próby podstawowej przeprowadzono estymację dwóch modeli dwumianowych: probitowego i opartego na dwuskładnikowej mieszance rozkładów normalnych. Bayesowskie oceny punktowe dla parametrów poszczególnych specyfikacji przedstawia tabela 2.

Zaprezentowane wyniki wskazują, iż z punktu widzenia statystycznego wprowadzone uogólnienie modelu probitowego było zasadne. Wartości oczekiwane a posteriori, dla kluczowych parametrów definiujących mieszankę, charakteryzują się stosunkowo niewielkimi odchyleniami standardowymi. Wskazuje to na ich statystyczną istotność. Warto zwrócić uwagę, że choć wartość wagi π_1 wynosi zaledwie 0,018, to niepewność związana z jej estymacją punktową jest bardzo mała. Obserwujemy zgodność znaków ocen parametrów mieszanki i modelu probitowego, z wyłączeniem zmiennej *ROR*. Formalne testowanie obu modeli z wykorzystaniem prawdopodobieństwa a posteriori pokazuje, iż dane zdecydowanie świadczą na rzecz modelu z mieszanką, a odrzucają model probitowy, zob. J. Marzec [2008]. Zatem ten pierwszy jest wysoce prawdopodobny a posteriori i charakteryzuje się lepszym dopasowaniem do danych.

Tabela 2. Wartości oczekiwane E(·|y) i odchylenia standardowe D(·|y) a posteriori parametrów poszczególnych modeli

×	Model mieszanki		Model probitowy	
Zmienna lub parametr	E(·|y)	D(·|y)	E(·|y)	D(·|y)
„1”	-	-	-1,261	0,124
Płeć (mężczyzna: x_{t1}=1)	0,006	0,004	0,047	0,018
Wiek (x_{t2} w setkach lat)	-0,170	0,023	-0,853	0,084
Wpływy (x_{t3} w setkach tys. zł)	-4,690	0,338	-1,404	0,139
ROR (posiada: x_{t4}=1)	0,083	0,012	-0,285	0,037
Karty (posiada: x_{t5}=1)	-0,009	0,010	-0,121	0,032
Pośrednik (x_{t6}=1)	0,248	0,020	1,254	0,032
Typ (konsumpcyjny: x_{t7}=1)	0,210	0,123	0,362	0,112
Okres (x_{t8} w dziesiątkach lat)	-0,054	0,014	-0,125	0,056
Kwota (x_9 w setkach tys. zł)	0,016	0,011	0,103	0,029
Waluta (PLN: x_{t10}=1)	0,142	0,124	0,380	0,129
Źródło dochodu (x_{t11})	-0,026	0,006	-0,103	0,029
Źródło dochodu (x_{t12})	0,033	0,012	0,303	0,040
Źródło dochodu (x_{t13})	-0,094	0,018	-0,243	0,075
μ_1	6,070	0,588	-	-
μ_2	-0,381	0,126	-	-
τ_2	26,424	4,000	-	-
π_1	0,018	0,001	-	-

Źródło: obliczenia własne.

Ocena trafności decyzji kredytowych

Z punktu widzenia praktycznego - zastosowania obu modeli w systemach scoringu kredytowego – istotnym zagadnieniem jest określenie ich zdolności dyskryminacyjnych poprzez porównanie trafności uzyskiwanych prognoz niespłacenia kredytów. W przypadku klienta o numerze *f* przedmiotem prognozy może być prawdopodobieństwo niespłacenia kredytu p_f=Pr(y_f=1)

lub zmienna zerojedynkowa y_f (1 – kredyt nie jest spłacany, 0 – w przeciwnym przypadku). Prognozę dla p_f uzyskujemy bezpośrednio z modelu (1). Jednakże z punktu widzenia praktyki bankowej to prognoza zmiennej y_f jest kluczowa, bo decyduje czy w przypadku potencjalnego kredytobiorcy negatywnie ocenić wniosek kredytowy (odmówić przyznania kredytu) czy pozytywnie (udzielić go). Prognozy tej zmiennej uzyskuje się na podstawie reguły: $\hat{y}_f = 1$, gdy $\hat{p}_f > p^*$ oraz $\hat{y}_f = 0$, gdy $\hat{p}_f \leq p^*$, gdzie p^* oznacza maksymalny, akceptowalny poziom prawdopodobieństwa niespłacenia kredytu, gdy bank decyduje się go udzielić. Wartość p^* ustaliliśmy na poziomie niezbilansowania próby [M. Gruszczyński 2001, s. 81], więc p^*=0,23. W konsekwencji przyjęliśmy, iż $\hat{y}_f = 1$, gdy $\hat{p}_f > 0{,}23$ oraz $\hat{y}_f = 0$, gdy $\hat{p}_f \leq 0{,}23$. Nie rozważaliśmy kosztów błędnych decyzji: odmowy kredytu rzetelnemu klientowi oraz przyznania kredytu kredytobiorcy, który go nie spłaci. Koszty tych obu błędnych decyzji są różne, więc przyjęcie progowej wartości dla p_f na poziomie 0,5 uważamy za bezzasadne. Uwzględnienie kosztów błędnych decyzji, w celu określenia takiego (minimalnego) poziomu prawdopodobieństwa stanu natury (wypłacalności klienta), który by uzasadniał przyznanie kredytu, proponuje się w pracy J. Osiewalski [2007]. Przyjęta w niniejszych badaniach graniczna (maksymalna) progowa wartość prawdopodobieństwa niespłacenia kredytu, p^*=0,23, odpowiada marży kredytowej na poziomie 17,5 punktów procentowych według propozycji J. Osiewalski [2007, tabela 2].

Tabele 3 i 4 prezentują oceny zdolności dyskryminacyjnych badanych modeli, w postaci liczebności oraz procentowej struktury przypadków poprawnego i błędnego zakwalifikowania kredytów w zależności od tego, czy klienci je spłacają czy nie. W modelu z mieszanką, spośród 27443 spłacanych kredytów i 8168 kredytów nie spłacanych, zostało poprawnie zakwalifikowanych odpowiednio 19858, czyli 72,4% ogółu dobrych rachunków i 7101, czyli 86,9% złych kredytów. Syntetyczny miernik, *count* R^2, wyrażający procentowy udział poprawnych prognoz do wszystkich prognoz jest wysoki i dla obu modeli wynosi 0,76 =(19858+7101)/35611. Wyniki prognoz, które uzyskano na podstawie specyfikacji probitowej są nieznacznie lepsze, jednakże tylko w odniesieniu do kredytów, które są spłacane nie zaś tych, należących do grupy należności zagrożonych. Mając na uwadze, iż zysk wynikający z udzielenia kredytu, który jest spłacany, jest kilkakrotnie mniejszy od straty ponoszonej przez bank, gdy kredytobiorca go nie spłaca, to prognozy uzyskane na podstawie modelu z mieszanką są lepsze. Ponadto z uwagi na fakt, że poszczególne kredyty posiadają różną wartość, porównano także trafność prognozowania odnosząc się do kwoty udzielonych kredytów. Wówczas w modelu mieszanki, spośród spłacanych w terminie kredytów o łącznej wartości 311,4 mln zł poprawne prognozy uzyskano dla grupy rachunków kredytowych o wartości 274,5 mln zł. Wśród 46,9 mln zł kredytów zagrożonych trafne prognozy dotyczyły kwoty 32,9 mln zł. Był to wynik łącznie o 5,2 mln zł lepszy niż w tradycyjnym modelu probitowym. W modelu mieszanki

udział wartości kredytów poprawnie zidentyfikowanych stanowi 85,79% portfela, zaś w probitowym jest niższy i wynosi 84,34%. Jest to kolejny argument, iż prognozy uzyskiwane na podstawie modelu mieszanki niosą więcej korzyści dla banku niż model probitowy.

Tabela 3. Tabela trafności ocen zmiennej y_t dla próby podstawowej (p^*=0,23)

Model	Z mieszanką		Probitowy	
W ujęciu ilościowym (w nawiasie - wg wartości udzielonych kredytów w mln zł)				
Stan faktyczny / Prognoza	Spłaca kredyt: y_t =0	Nie spłaca: y_t =1	Spłaca kredyt: y_t =0	Nie spłaca: y_t =1
$\hat{y}_t = 0$	**19858 (274,5)**	1067 (14,0)	**19884 (269,6)**	1111 (14,3)
$\hat{y}_t = 1$	7585 (36,9)	**7101 (32,9)**	7559 (41,8)	**7057 (32,6)**
Razem	27443 (311,4)	8168 (46,9)	27443 (311,4)	8168 (46,9)
W ujęciu procentowym (w nawiasie – w odniesieniu do wartości udzielonych kredytów)				
Stan faktyczny / Prognoza	Spłaca kredyt: y_t =0	Nie spłaca: y_t =1	Spłaca kredyt: y_t =0	Nie spłaca: y_t =1
$\hat{y}_t = 0$	**72,4% (88,1%)**	13,1% (29,9%)	**72,5% (86,6%)**	13,6% (30,4%)
$\hat{y}_t = 1$	27,6% (11,9%)	**86,9% (70,1%)**	27,5% (13,4%)	**86,4% (69,6%)**

Źródło: obliczenia własne.

Zbadano także trafności prognoz dla próby odłożonej. W tym przypadku oba modele prawie identycznie prognozują niespłacenie kredytów. Uzyskują one bardzo wysoką trafność prognoz dla kredytów, które w rzeczywistości są spłacane, ale dużo niższą w odniesieniu do kredytów nie spłacanych. W przypadku aż 95,3% (dla mieszanki) i 95,7% (dla probitu) łącznej liczby kredytów, które są spłacane, uzyskujemy poprawne prognozy, iż można było ich udzielić. Dotyczy to około 98,9% wartości portfela dobrych (spłacanych) kredytów. W konsekwencji, z uwagi na wysoki udział dobrych kredytów, miernik *count* R^2 przyjmuje wysoką wartość 0,89 dla obu modeli. Jednakże trudniej prognozowane są decyzje klientów, którym udzielono kredyty, a którzy ich nie spłacają. Błędną prognozę uzyskano dla co drugiego kredytobiorcy z tej grupy, tzn. z 200 kredytów błędnie zostało zakwalifikowanych aż 115. Wartość nie spłacanych kredytów, w przypadku których model rekomendował ich udzielenie, stanowi 84,1% wartości portfela kredytów zagrożonych. Dotyczy to obu modeli.

W innych polskich badaniach otrzymano niższe wartości zliczeniowego R^2: 56%-73% u Misztal [2006] oraz 75%-74% u Witkowskiej i Chrzanowskiej [2006], gdy stosowano drzewa decyzyjne. W przypadku modelu logitowego odsetek poprawnych decyzji wynosił 71%, zaś dla sieci neuronowych - 68%; zob. Witkowska i Chrzanowska [2004]. W przypadku innych badań - Featherstone, Roessler i Barry [2006] - uzyskano niższy wskaźnik poprawnie sklasyfikowanych rachunków, bo wynoszący tylko 65,4%, gdy dla dużo liczniejszej próby zastosowano model logitowy. Natomiast w badaniach Lee i Liu [2002] odsetek wszystkich poprawnie zakwalifikowanych kredytów hipotecznych wynosił 77,25% i 89% w zależności od przeznaczenia pożyczek: na kupno nowych czy używanych nieruchomości. W innych badaniach wielkość tej

miary jakości modelu kształtuje się na poziomie od 76% do 80%; zob. Banasik, Cook i Thomas [2003]. Wobec przedstawionych wyników z literatury przedmiotu można zgodnie stwierdzić, iż trafność dyskryminacji uzyskana na podstawie badanych modeli jest wysoka.

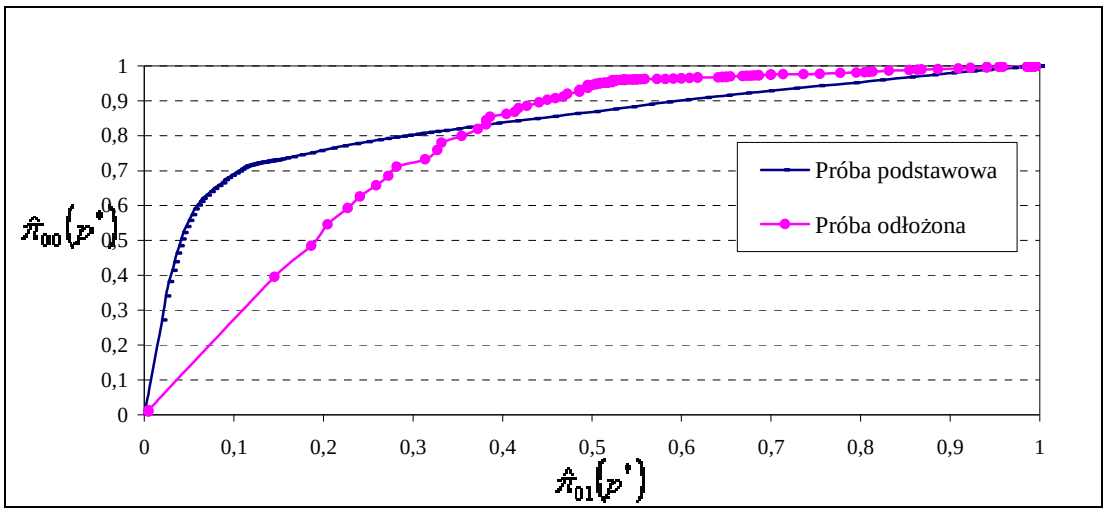
Tabela 4. Tabela trafności prognoz zmiennej y_f dla próby odłożonej (p^*=0,23)

Model	Z mieszanką		Probitowy	
W ujęciu ilościowym (w nawiasie - wg wartości udzielonych kredytów w mln zł)				
Stan faktyczny / Ocena	Spłaca kredyt: y_f=0	Nie spłaca: y_f=1	Spłaca kredyt: y_f=0	Nie spłaca: y_f=1
$\hat{y}_f = 0$	**3053 (42,0)**	115 (1,6)	**3066 (41,9)**	116 (1,6)
$\hat{y}_f = 1$	150 (0,4)	**105 (0,3)**	137 (0,5)	**104 (0,3)**
Razem	3203 (42,4)	220 (1,9)	3203 (42,4)	220 (1,9)
W ujęciu procentowym (w nawiasie – w odniesieniu do wartości udzielonych kredytów)				
Stan faktyczny / Ocena	Spłaca kredyt: y_t=0	Nie spłaca: y_t=1	Spłaca kredyt: y_t=0	Nie spłaca: y_t=1
$\hat{y}_t = 0$	**95,3% (98,9%)**	52,3% (84,1%)	**95,7% (98,8%)**	52,7% (84,2%)
$\hat{y}_t = 1$	4,7% (1,1%)	**47,7% (15,9%)**	4,3% (1,2%)	**47,3% (15,8%)**

Źródło: obliczenia własne.

Krzywa ROC

Innym miernikiem służącym do określenia stopnia dopasowania modelu do danych jest pole pod krzywą ROC (Receiver Operating Characteristic) [D. Hosmer S. Lemeshow 2000, s. 160, Stein 2005]. Krzywa ta prezentuje zależność pomiędzy częstością poprawnych i błędnych prognoz w zależności od wszystkich możliwych wartości p^*, $p^* \in (0;1)$; zob. rysunek 1. Na osi rzędnych znajduje się odsetek ilości kredytów spłacanych (y_t=0), $\hat{\pi}_{00}(p^*)$, w przypadku których dla ustalonego p^* ocena dla y_t jest poprawna, czyli $\hat{y}_t = 0$. Na osi odciętych odkładamy odsetek kredytów nie spłacanych (y_t=1), $\hat{\pi}_{01}(p^*)$, dla których przy ustalonym p^* pojęto błędną decyzję o udzieleniu kredytu $\hat{y}_t = 0$. Pole pod krzywą ROC wynosi 0,84, co świadczy o „doskonałych" zdolnościach dyskryminacyjnych modelu [D. Hossmser S. Lemeshow 2000, s. 162]. W przypadku modelu probitowego miernik dopasowania przyjmuje bardzo zbliżoną wartość 0,83. W przypadku próby odłożonej jego wartość w obu przypadkach jest niższa i wynosi 0,78, lecz można ten wynik uznać za bardzo dobry.



Rysunek 1. Krzywa ROC: zależność między $\hat{\pi}_{00}(p^*)$ a $\hat{\pi}_{01}(p^*)$ dla $p^* \in (0;1)$
Źródło: obliczenia własne.

Inne mierniki dopasowania modelu do danych

W uzupełnieniu prezentujemy, w tabeli 5, informacje o średnich ocenach prawdopodobieństwa niespłacenia kredytu p_t dla dwóch grup kredytów - są spłacane albo nie. W przypadku idealnej klasyfikacji na przekątnej tej tabeli powinny znajdować się jedynki. Dla grupy rzetelnych kredytobiorców średnia prognoza spłacenia kredytu ($1–p_t$) wynosi 0,83 w przypadku próby podstawowej i 0,94 dla próby odłożonej. Są to bardzo dobre wyniki.

Tabela 5. Średnie prawdopodobieństwo zakwalifikowania kredytu do jednej z dwóch kategorii (T_j to liczebność zbioru obserwacji $y_t = j$ dla $j \in (0, 1)$)

Próba	podstawowa		odłożona	
Stan faktyczny / Ocena	Dobry kredyt (y_t =0)	Zły kredyt (y_t =1)	Dobry kredyt (y_t =0)	Zły kredyt (y_t =1)
$(T_j)^{-1} \sum_t \Pr(y_t=0)$ dla t: $y_t=j$	**0,83**	0,56	**0,94**	0,74
$(T_j)^{-1} \sum_t \Pr(y_t=1)$ dla t: $y_t=j$	0,17	**0,44**	0,06	**0,26**
Suma	1,00	1,00	1,00	1,00

Źródło: obliczenia własne.

W grupie kredytobiorców, którzy mają trudności ze spłatą, to prawdopodobieństwo kształtuje się na poziomie 0,44 w próbie podstawowej i tylko 0,26 w drugiej próbie. Te wyniki nie są zadawalające i potwierdzają wcześniej uzyskane rezultaty, iż omawiane modele są bardzo skuteczne, gdy prognoza dotyczy kredytu solidnego kredytobiorcy, natomiast nie są już tak efektywne w przypadku klienta, który odstępuje od podpisanej wcześniej umowy kredytowej.

Obliczyliśmy także klasyczne, syntetyczne mierniki służące do porównań dopasowania modeli dychotomicznych, tj. średnią wartość prawidłowej prognozy prawdopodobieństwa, czyli R^2 Ben-Akivy i Lermana, wyrażoną formułą $R^2_{BL} = T^{-1}\sum_{t=1}^{T} y_t \hat{p}_t + (1-y_t)(1-\hat{p}_t)$ oraz skorygowaną wartość tego miernika (λ Cramera), uwzględniającą niezbilansowanie próby, daną wzorem $\lambda = T_1^{-1}\sum_{t:y_t=1} \hat{p}_t - T_0^{-1}\sum_{t:y_t=0} \hat{p}_t$, [M. Gruszczyński 2001, J. Cramer 2003]. Wartość miernika R^2_{BL}, określonego przez średnią ważoną ocenę prawdopodobieństwa poprawnego zakwalifikowania

kredytu, gdzie wagami są udziały zer i jedynek w próbie, wynosi 0,74 dla próby podstawowej i 0,89 dla próby odłożonej. Mniejsza skuteczność modeli w klasyfikacji złych kredytów powoduje, iż λ jest równe 0,27 dla pierwszej próby, zaś w drugiej próbie jest niższe, bo wynosi zaledwie 0,1. Dla modelu probitowego wartości tych mierników są na takim poziomie jak w mieszance. Rezultaty te świadczą, iż prognozowanie zdarzenia polegającego na niespłaceniu kredytu, gdy relatywnie rzadko ma miejsce w próbie, jest zagadnieniem trudnym. Należy pamiętać, iż do estymacji obu modeli wykorzystano dane o wnioskach kredytowych, które zostały pozytywie zweryfikowane na etapie wstępnej oceny wniosków kredytowych, przeprowadzonej przez pracowników banku. Zatem utrudniło to poprawne prognozowanie prawdopodobieństwa niespłacenia kredytów.

5. Podsumowanie

Zaprezentowane wyniki pokazały, że zastosowanie modelu dwumianowego opartego na dwuskładnikowej mieszance rozkładów normalnych jest uzasadnione w świetle posiadanych danych. Ten model jawi się jako skuteczne narzędzie prognozowania niespłacalności kredytów. Uzyskane na jego podstawie prognozy są bardzo trafne, więc mogłyby przynieść wymierne korzyści finansowe dla banku. Model mieszanki jest rekomendowany zwłaszcza w sytuacji, gdy udział złych kredytów jest znaczący, gdyż wówczas jego użycie do prognozowania zachowania się kredytobiorcy przynosi więcej korzyści niż model probitowy. Dla obu specyfikacji wyniki prognoz ax ante dla kredytów, które nie są spłacane przez kredytobiorców, okazały się mniej zadawalające od rezultatów uzyskanych na podstawie próby podstawowej. Jedną z przyczyn jest niewielki odsetek tych obserwacji w próbie. Przedmiotem kolejnych badań byłoby dokonanie pomiaru korzyści finansowych, jakie można uzyskać, stosując ekonometryczne modele danych jakościowych w scoringu kredytowym. Wymaga to uwzględnienia kosztów błędnych decyzji - przyznania bądź odmowy kredytu - podjętych na podstawie prognozy zmiennej y_t.

6. Literatura

Austin P., M. Escobar, *The Use of Finite Mixture Models to Estimate the Distribution of the Health Utilities Index in the Presence of a Ceiling Effects*, "Journal of Applied Statistics" 2002, vol. 30, nr 8, s. 909-923,

Banasik J., J. Crook, L.C. Thomas, *Sample Selection Bias in Credit Scoring Models*, „Journal of the Operational Research Society" 2003, vol. 54, s. 822-832.

Chrzanowska M., D. Witkowska, *Zastosowanie sztucznych sieci neuronowych do klasyfikacji ryzyka kredytowego podmiotów gospodarczych*, „Metody ilościowe w badaniach ekonomicznych V" 2005, Wydawnictwo SGGW, Warszawa, s. 80-89.

Chrzanowska M., K. Kompa, D. Witkowska, *Analiza spłat pewnego kredytu okolicznościowego. Modele logitowe i sieci neuronowe*, „Metody ilościowe w badaniach ekonomicznych V" 2005, Wydawnictwo SGGW, Warszawa, s.67-79.

Cramer J.S, *Logit Models From Economics and Other Fields*, Cambridge University Press, Cambridge 2003.

Diebolt J., C.P. *Robert, Estimation of Finite Mixture Distributions though Bayesian Sampling*, "Journal of the Royal Statistical Society B" 1994, vol. 56, nr 2, s. 363-375.

Erkanli A., D. Stangl, P. Müller, *A Bayesian Analysis of Ordinal Data Using Mixtures*, "Technical Report 93-01 Institute of Statistics and Decision Sciences", Duke University 1993.

Featherstone A.M., L.M. Roessler, P.J. Barry, *Determining the Probability of Default and Risk Rating Class for Loans in the Seventh Farm Credit District Portfolio*, „Review of Agricultural Economics" 2006, 28, s. 4-23.

Frühwirth-Schnatter S., R. Frühwirth, *Auxiliary Mixture Sampling with Applications to Logistics Models*, "Computational Statistics and Data Analysis" 2006, vol. 51, nr 7.
Geweke J., *Interpretation and Inference in Mixture Models: Simple MCMC Works*, "Computational Statistics and Data Analysis" 2007, vol. 51, nr 7, s. 3529-3550.
Geweke J., M. Keane, *Mixture of Normals Probit Models*, [w:] Analysis of Panel and Limited Dependent Variables: A Volume in Honor of G.S. Maddala, red. Hsiao C., Lahiri K., Lee L.F., Pesaran M.H., Cambridge University Press, Cambridge 1999.
Gruszczyński M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Monografie i Opracowania SGH nr 6, Warszawa 2001.
Hosmer D., S. Lemeshow, *Applied Logistic Regression*, Wiley, New York 2000.
Koop G., *Bayesian Econometrics*, Wiley, Chichester 2003.
Lee S.P., D.Y. Liu, *The Determinations of Defaults in Residential Mortgage Payments: A Statistical Analysis*, „International Journal of Management" 2002, vol. 19, nr 2, s. 377-389.
Mączyńska E., M. Zawadzki, *Dyskryminacyjne modele predykcji upadłości przedsiębiorstw*, „Ekonomista" 2006, nr 2, s. 205-235.
Marzec J., *Bayesowski model dwumianowy z mieszanką rozkładów normalnych*, [w:] Metody ilościowe w naukach ekonomicznych, red. A. Welfe, Wydawnictwo SGH w Warszawie, Warszawa 2008, w druku.
Marzec J., *Bayesowski model wielomianowy z rozkładem t Studenta dla kategorii uporządkowanych*, [w:] Metody ilościowe w naukach ekonomicznych, red. A. Welfe, Wydawnictwo SGH w Warszawie. Warszawa 2006, s. 123-144.
Misztal M., *O zastosowaniu metody rekurencyjnego podziału w analizie ryzyka kredytowego*, „Modelowanie preferencji a ryzyko'05" 2006, Wydawnictwo AE w Katowicach, s. 453-468.
Osiewalski J., *Bayesowska statystyka i teoria decyzji w analizie ryzyka kredytu detalicznego*, [w:] Finansowe warunkowania decyzji ekonomicznych, red. D. Fatuła, Wydawnictwo Krakowskiej Szkoły Wyższej w Krakowie, Kraków 2007.
Qu P., Y. Qu, *A Bayesian Approach to Finite Mixture models in Bioassay via Data Augmentation and Gibbs Sampling and Its Application to Insecticide Resistance*, "Biometrics" 2000, 56, s. 1249-1255.
Roeder K., L. Wasserman, *Practical Bayesian Density Estimation Using Mixtures of Normals*, "Journal of the American Statistical Association" 1997, vol. 92, nr. 439; s. 894-902.
Rosenberg E., A. Gleit, *Quantitative Methods in Credit Management: A Survey*, "Operations Research" 1994, vol. 42, nr 4, s.589-613.
Staniec I., *Badanie zdolności kredytowej przy użyciu sztucznych sieci*, „Zeszyty Naukowe Politechniki Łódzkiej" 2000, Organizacja i Zarządzanie, zeszyt 35, nr 876, Łódź, s. 129-139.
Staniec I., D. Witkowska, *Dychotomiczna klasyfikacja kredytobiorców przy użyciu wielowymiarowej analizy dyskryminacyjnej*, „Acta Universitatis Lodziensis. Folia Oeconomica 156" 2002, s. 221-233.
Staniec I., *Drzewa klasyfikacyjne w ocenie wiarygodności kredytobiorców*, „Taksonomia 12. Klasyfikacja i analiza danych – teoria i zastosowania" 2005, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, s.547-555.
Staniec I., M. Szmit, *Ocena wiarygodności kredytobiorców przy użyciu sieci LTF-C*, „Zeszyty Naukowe Wydziału Mechanicznego nr 35" 2004a, Politechnika Koszalińska, s. 216-223.
Staniec I., M. Szmit, *Ocena wiarygodności kredytobiorców przy użyciu analizy dyskryminacyjnej oraz drzew klasyfikacyjnych*, „Zeszyty Naukowe Politechniki Łódzkiej, Organizacja i Zarządzanie nr 40" 2005, Łódź, s. 79-90.
Staniec I., M. Szmit, *Sztuczne sieci neuronowe w ocenie wiarygodności kredytowej klienta*, „Zeszyty Naukowe Akademii Górniczo-Hutniczej w Krakowie" 2004b, zeszyt 3 tom 8, s. 449-457.
Stein R. M., *The Relationship Between Default Prediction and Lending Profits: Integrating ROC Analysis and Loan Pricing*, "Journal of Banking and Finance" 2005, 29, s. 1213-1236.
Szmit A., M. Szmit, M. Kaniewski, Analiza ryzyka kredytowego na potrzeby pośrednika finansowego, „Taksonomia 10. Klasyfikacja i analiza danych – teoria i zastosowania" 2003, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, s. 163-172.
Thomas L.C., *A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers*, "International Journal of Forecasting" 2000, vol. 16, s. 149-172.
Thomas L.C., R.W. Oliver, D.J. Hand, *A survey of the issues in consumer credit modelling research*, "The Journal of the Operational Research Society" 2005, vol. 56, nr 9, s.1006-1015.
Titterington D.M., A.F.M. Smith, U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*, John Wiley, New York 1985.
Tymoczko I., M. Pawłowska, *Uwarunkowania dostępności kredytu bankowego – analiza polskiego rynku*, „Bank i Kredyt" 2007, nr 6, s. 47-68.
Wędzki D, *Zastosowanie logitowego modelu upadłości przedsiębiorstw*, „Ekonomista" 2005, nr 5, s. 691-705.
Witkowska D., M. Chrzanowska, *Drzewa klasyfikacyjne jako metoda grupowania klientów banku*, „Modelowanie preferencji a ryzyko'05" 2006, Wydawnictwo AE w Katowicach, s. 485-496.
Witkowska D., M. Chrzanowska, *Drzewa klasyfikacyjne w rozpoznawaniu kredytobiorców*, „Metody ilościowe w badaniach ekonomicznych VII" 2007, Wydawnictwo SGGW, Warszawa, s. 291-301.
Witkowska D., M. Chrzanowska, *Wybrane metody klasyfikacji kredytobiorców: modele logitowe i sieci neuronowe*, „Modelowanie preferencji a ryzyko'04" 2004, Wydawnictwo AE w Katowicach, s.531-540.