A C T A U N I V E R S I T A T I S N I C O L A I C O P E R N I C I
DOI: http://dx.doi.org/10.12775/AUNC_ECON.2014.012 EKONOMIA XLV nr 2 (2014) 187–203

Pierwsza wersja złożona 26 listopada 2014
Końcowa wersja zaakceptowana 20 grudnia 2014

ISSN
2080-0339

Dorota Grochowina[*]

WPŁYW METOD IMPUTACJI DANYCH NA SKUTECZNOŚĆ KLASYFIKACYJNĄ MODELU LOGITOWEGO ZASTOSOWANEGO DO PROGNOZOWANIA UPADŁOŚCI PRZEDSIĘBIORSTW

Z a r y s t r e ś c i. Prognozowanie upadłości przedsiębiorstw obnaża problem braku danych, który dotyczy głównie podmiotów z problemami finansowymi, pragnących w ten sposób zataić złą kondycję. Jedną z metod uzupełniania niekompletnych danych jest imputacja. Celem pracy jest przedstawianie różnych odmian imputacji danych oraz porównanie ich wpływu na skuteczność klasyfikacyjną jednej ze statystycznych metod prognozowania upadłości – modelu logitowego. Wyniki analizy wykazały, iż najlepszym podejściem jest zastosowanie mediany wyznaczonej osobno dla grupy zdrowych i upadłych przedsiębiorstw.

S ł o w a k l u c z o w e: bankructwo, prognozowanie upadłości, model logitowy, imputacja, szacowanie brakujących danych, skuteczność klasyfikacyjna modelu.

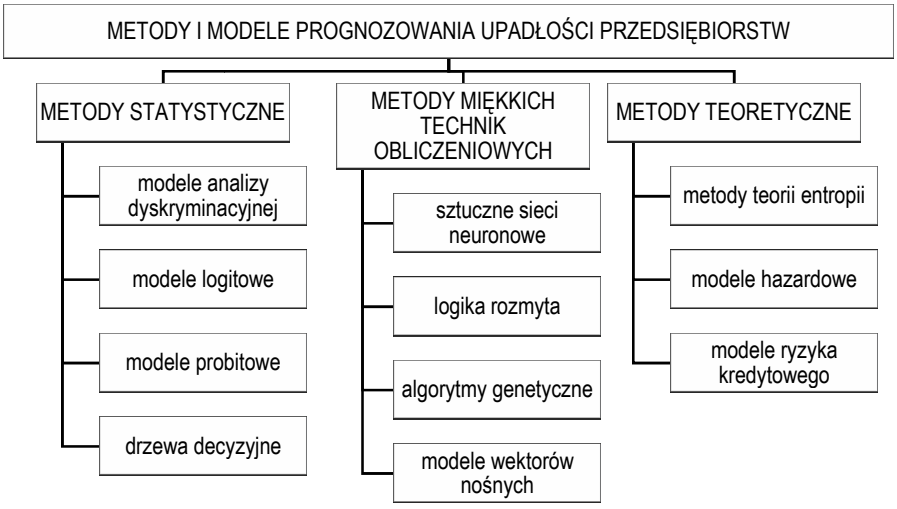
K l a s y f i k a c j a J E L: C53, G17, G33.

WSTĘP

Zmieniająca się sytuacja gospodarcza na świecie, ogólnoświatowe kryzysy ekonomiczne, zmiany preferencji ludności czy złe zarządzanie przedsiębiorstwami – to tylko niektóre z potencjalnych przyczyn pogarszania się sytuacji finansowej czy organizacyjnej firmy, a długotrwałe problemy prowadzą najczęściej do jej bankructwa. Pojęcie upadłości (potocznie nazywane bankructwem) nie ma jednolitej definicji. Lewandowski i Wołowski (2011) definiują upadłość jako prawnie regulowaną procedurę, wszczynaną w razie

[*] Adres do korespondencji: Dorota Grochowina, Uniwersytet Ekonomiczny w Krakowie, ul. Rakowicka 27, 31–510 Kraków, e-mail: dorota.grochowina89@gmail.com.

niewypłacalności dłużnika, w ramach której dochodzone są roszczenia wszystkich jego wierzycieli. Z kolei Prusak (2011) przez bankructwo rozumie stan, kiedy kondycja przedsiębiorstwa nie pozwala mu na kontynuowanie działalności w takim zakresie, żeby było rentowne, płynne i wypłacalne oraz było w stanie nadal konkurować na rynku, bez korzystania z pomocy z zewnątrz. Bankructwo, obok dobrowolnej decyzji właścicieli o zaprzestaniu działalności, stanowi powód likwidacji przedsiębiorstwa. Warto to podkreślić, gdyż likwidacja często utożsamiana jest wyłącznie z działaniem dobrowolnym (Hołda, 2006).



Schemat 1. Podział metod prognozowania upadłości przedsiębiorstw

Źródło: opracowanie własne na podstawie Korol (2010).

Liczba podmiotów zainteresowanych kondycją przedsiębiorstw, jak i skala skutków ich upadłości powodują, że kontrola sytuacji finansowej podmiotów gospodarczych zyskuje na znaczeniu. Skutecznym narzędziem dostarczającym informacji o potencjalnym zagrożeniu bankructwem, ujawniającym pogarszającą się sytuację finansową firmy z odpowiednim wyprzedzeniem czasowym, jest analiza finansowa. Jednakże jej pracochłonność i niejednoznaczność wniosków skłaniała do poszukiwania metod umożliwiających szybką, a zarazem wiarygodną diagnozę kondycji firmy. W konsekwencji powstały metody prognozowania bankructwa przedsiębiorstw, których podział przedstawia Schemat 1. Wśród nich wyodrębnić można grupę metod statystyczno-ekonometrycznych (Korol, 2010), które ułatwiają dokonanie syntetycznej oceny kondycji przedsiębiorstwa. Jednym z przykładów

jest model logitowy, na którym skupiono uwagę w niniejszej pracy. Oparto go na wskaźnikach finansowych, których analiza stanowi część analizy finansowej przedsiębiorstwa.

Zgodnie z ustawą o rachunkowości spółki akcyjne, jako spółki kapitałowe, mają obowiązek składania do właściwego sądu rejonowego corocznych sprawozdań finansowych (DzU 1974, Nr 121). Dodatkowo, obowiązki informacyjne tych spółek akcyjnych, które notowane są na Giełdzie Papierów Wartościowych w Warszawie, regulują:

– Kodeks spółek handlowych (DzU 2000, Nr 94),
– ustawa o ofercie publicznej (DzU 2005, Nr 184),
– ustawa o obrocie instrumentami finansowymi (DzU 2005, Nr 183),
– Rozporządzenie Ministra Finansów w sprawie informacji bieżących i okresowych (DzU 2009, Nr 33).

Obowiązek publikacji sprawozdań ułatwia dostęp do danych finansowych spółek, aczkolwiek nie gwarantuje, że wszystkie z nich z tego obowiązku się wywiążą, ani że sprawozdania będą kompletne. W przypadku niedopełnienia obowiązku notowanie spółki może zostać zawieszone na wniosek Komisji Nadzoru Finansowego (DzU 2005, Nr 183). Z sytuacją brakujących danych można się spotkać głównie w przypadku spółek z problemami finansowymi, które niejednokrotnie próbują w ten sposób ukryć złą kondycję, o czym można się przekonać w trakcie prognozowania upadłości przedsiębiorstw.

W badaniu prognozowania upadłości można zastosować co najmniej kilka metod uzupełniania niepełnych zbiorów danych, których dobór może być determinowany zakresem owych braków. Celem pracy jest przedstawienie kilku możliwych metod uzupełniania brakujących danych oraz porównanie ich wpływu na skuteczność klasyfikacyjną rozważanych modeli logitowych.

1. MODEL LOGITOWY

Regresja logistyczna jest metodą służącą do opisu wpływu zbioru zmiennych objaśniających na zmienną jakościową. Możemy określić prawdopodobieństwo, z jakim w przyszłości wystąpi określony wariant zmiennej objaśnianej, w zależności od działania innych czynników (zmiennych niezależnych) (Zeliaś, Pawełek, Wanat, 2003). Zmienne niezależne mogą mieć charakter zarówno ilościowy, jak i jakościowy, natomiast zmienna zależna musi mieć charakter jakościowy (Hołda, 2006). Liczba wariantów zmiennej zależnej może być skończona lub przeliczalna. Prognozowanie zmiennej jakościowej o wielu wariantach sprowadza się do prognozowania zmiennych

zero-jedynkowych (Zeliaś, Pawełek, Wanat, 2003). Najczęściej stosowany sposób kodowania zmiennych jest następujący:

$$Y = \begin{cases} 1, & \text{jeżeli dany wariant wystąpi,} \\ 0, & \text{jeżeli dany wariant nie wystąpi'} \end{cases} \quad (1)$$

gdzie Y oznacza zmienną objaśnianą (prognozowaną), o rozkładzie prawdopodobieństwa:

$$\begin{aligned} P(Y = 1) &= p, \\ P(Y = 0) &= q, \end{aligned} \quad (2)$$

a p+q=1.

Jeżeli za Z przyjmiemy liniową kombinację zmiennych objaśniających, obrazujących zmienne mające wpływ na zmienną prognozowaną, możemy ją zapisać jako (Rószkiewicz, 2002):

$$Z = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k \,. \quad (3)$$

Wartości oszacowania współczynników α_0, α_1, …, α_k pozwalają na obliczenie wspomnianego wcześniej prawdopodobieństwa wystąpienia danego wariantu zmiennej zależnej Z (Koop, 2011). Funkcja prawdopodobieństwa, w przypadku gdy zmienna zależna jest zmienną dychotomiczną, sprowadza się do funkcji logistycznej (regresji logistycznej, modelu logistycznego) i przyjmuje następującą postać (Rószkiewicz, 2002):

$$p = P(Z) = \frac{e^Z}{1+e^Z} = \frac{1}{1+e^{-Z}}. \quad (4)$$

Wartość minimalna funkcji jest osiągana, gdy zmienna Z dąży do –∞, zaś maksimum – gdy zmienna ta dąży do +∞. Granice te w przypadku tak zdefiniowanej funkcji logistycznej $P(Z)$ przyjmują wartości odpowiednio (Larose, 2008):

$$\lim_{Z \to -\infty} P(Z) = \frac{1}{1+ e^{-Z}} = 0, \quad (5)$$

$$\lim_{Z \to +\infty} P(Z) = \frac{1}{1+ e^{-Z}} = 1. \quad (6)$$

Z powyższego wynika, że wartości funkcji logistycznej mieszczą się w przedziale [0; 1], co umożliwia stosowanie funkcji P(Z) do opisu rozkładu prawdopodobieństwa. Graficznie funkcja przybiera kształt krzywej sigmoidalnej (Larose, 2008).

Do szacowania wartości parametrów w modelu logitowym stosowana jest metoda największej wiarygodności. W tym przypadku bowiem nie możemy skorzystać z metody najmniejszych kwadratów, gdyż ze względu na występowanie w modelu zmiennej dychotomicznej, nie jest spełniony warunek o stałości wariancji.

Przyjmując, że X_1, X_2, ..., X_k stanowią zbiór zmiennych objaśniających, a α_1, α_2, ..., α_k są nieznanymi parametrami modelu, funkcję wiarygodności można zdefiniować jako (Hołda, 2006):

$$L(\alpha) = \prod_{i=1}^{n} p(y_i|\alpha_1, \dots, \alpha_k), \tag{7}$$

gdzie: $p(y_i|\alpha_1, \dots, \alpha_k)$ – prawdopodobieństwo pojawienia się wartości zmiennej objaśnianej y_i przy danym modelu regresji z parametrami $\alpha_1, \dots, \alpha_k$, n – liczba możliwych wartości zmiennej objaśnianej Y.

Metoda ta zakłada, że szukamy takich wartości α, dla których funkcja $L(\alpha)$ przyjmie wartość maksymalną (Koop, 2011).

Na podstawie zależności $p+q=1$ oraz równania (4) prawdziwa jest równość (Gruszczyński, Kuszewski, Podgórska, 2009):

$$q = 1 - P(Z) = \frac{1}{1+ \mathrm{e}^{Z}}. \tag{8}$$

Dokonując przekształceń, korzystając z równości (4) i (8), iloraz $\frac{p}{q}$, zwany również ilorazem szans[1], możemy zapisać jako (Gruszczyński, Kuszewski, Podgórska, 2009):

$$\frac{p}{q} = \frac{P(Y=1)}{P(Y=0)} = \mathrm{e}^{Z}. \tag{9}$$

Logarytmując obie strony równania, otrzymujemy następującą zależność (Gruszczyński, Kuszewski, Podgórska, 2009):

$$\ln\left(\frac{p}{q}\right) = Z = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_k X_k. \tag{10}$$

Tak zapisany model (10) ma charakter liniowy względem parametrów α oraz zmiennych X. Zmienna objaśniana w tym modelu, będąca logarytmem naturalnym ilorazu szans, nazywana jest logitem. Model (10) jest modelem logitowym (funkcją logitową).

2. IMPUTACJA

Imputacja, obok usuwania rekordów oraz ważenia danych (np. kalibracja, postratyfikacja), jest trzecim podejściem możliwym do zastosowania w przypadku niekompletności zbioru danych (Beręsewicz, 2010). Imputacja definiowana jest jako szacowanie brakujących wartości, czyli zastąpienie ich tzw. wartościami imputacyjnymi (Balicki, 2004).

Zgodnie z podziałem dokonanym przez Piaseckiego (2014), można wyróżnić imputację pozycyjną oraz brakujących rekordów. W przypadku bada-

[1] Przez „szansę" rozumiane jest prawdopodobieństwo.

nia upadłości przedsiębiorstw, z imputacją pozycyjną mamy do czynienia w sytuacji brakujących pojedynczych wskaźników finansowych dla danego przedsiębiorstwa, natomiast imputacja brakujących rekordów (nazywana również kompleksową (Młodak, 2010)) wiąże się z brakiem wszystkich wskaźników finansowych dla danego przedsiębiorstwa.

Istnieje wiele odmian imputacji danych oraz ich podziału, których opisu dokonali m.in. Longford (2005) oraz Kalton i Kasprzyk (1982). Metody imputacji można podzielić na dwa podstawowe typy – dedukcyjny oraz statystyczny. Imputacja dedukcyjna, do wyznaczania wartości imputacyjnej, korzysta z zależności między zmiennymi. Ma ona charakter deterministyczny, czyli pozwala na jednoznaczne wyznaczenie wartości imputacyjnej. Z kolei w przypadku imputacji statystycznej wykorzystywana jest pozostała część zbioru danych dotycząca imputowanej zmiennej. Metody imputacji statystycznej podlegają podziałowi na:

- imputację stochastyczną – gdy proces imputacji zawiera element losowości, tym samym dla wejściowego zbioru danych możemy otrzymać różne imputowane wartości,
- imputację deterministyczną – gdy tworzeniu wartości imputowanej nie towarzyszy element losowy, dzięki czemu wartość określona jest jednoznacznie przez metodę oraz wejściowy zbiór danych (tzn. powtarzając imputację dla tego samego zbioru danych, otrzymamy zawsze tą samą wartość).

Ze specyfiki powyższych metod wynika, że metody deterministyczne gwarantują większą precyzję uogólnień w porównaniu do metody stochastycznej, gdyż nie wprowadzają dodatkowo błędu losowego.

W pracy skupiono się m.in. na jednej z odmian imputacji deterministycznej, do której zaliczane są m.in. (Piasecki, 2013; Beręsewicz, 2010):

- imputacja z wykorzystaniem średniej – polega na zastąpieniu brakujących danych wartością średnią wyznaczoną na pozostałej grupie obserwacji (obserwacji kompletnych),
- imputacja z wykorzystaniem mediany – podobnie jak w przypadku imputacji z wykorzystaniem średniej, metoda sprowadza się do zastąpienia brakującej wartości medianą wyznaczoną na kompletnej grupie obserwacji. Możliwe jest zastosowanie mediany wyznaczonej osobno w klasach (grupach) wyłonionych na podstawie przyjętego kryterium. Wówczas mamy do czynienia z imputacją medianą warunkową. W pracy zastosowano oba podejścia – medianę wyznaczoną dla całego zbioru obserwacji oraz medianę wyznaczoną osobno dla zdrowych i upadłych przedsiębiorstw (imputacja medianą warunkową),

– imputacja regresyjna – metoda oparta na równaniu regresji. Wówczas za zmienną objaśnianą w regresji przyjmowana jest zmienna, której dotyczy brak danych, a za zmienne objaśniające przyjmowane są pozostałe zmienne (kompletne). Imputacja sprowadza się do odpowiedniego doboru modelu regresji, oszacowaniu modelu i ostatecznie zastąpieniu brakującej wartości wartością teoretyczną wynikającą z modelu.

3. ZAŁOŻENIA BADANIA

Badanie przeprowadzono na danych dotyczących przedsiębiorstw, których upadłość ogłoszono w okresie od 2009 do 2012 roku. Lata 2007–2008 przyjmuje się za początek ogólnoświatowego kryzysu gospodarczo-finansowego. Z racji tego, by uniknąć zakłóceń wynikających z odmiennej sytuacji na rynku, w analizie pominięto przedsiębiorstwa, których upadłość ogłoszono przed 2009 rokiem. Kierując się kryterium dostępności danych, wzięto pod uwagę tylko spółki notowane na Giełdzie Papierów Wartościowych w Warszawie, co jednak w istotny sposób ograniczyło bazę danych. Informacje na temat ogłoszonych upadłości zaczerpnięto z Ogólnopolskiego Informatora Upadłościowego, dostępnego z poziomu portalu EMIS. Listę upadłych spółek uwzględnionych w badaniu przedstawia tabela 1.

Za źródło informacji o sytuacji finansowej i kondycji przedsiębiorstw przyjęto sprawozdania finansowe odpowiednio z roku oraz 2 lat przed ogłoszeniem upadłości. Do badań wykorzystano wskaźniki finansowe zamieszczone we wspomnianych sprawozdaniach finansowych. Listę wskaźników przedstawiono w tabeli 2. Dokładnego opisu wskaźników finansowych oraz ich interpretacji dokonał m.in. Wędzki (2009), Korol (2013), Podstawka (2010), Bragg (2010) oraz Jerzemowska (2013). W celu zapewnienia porównywalności zmiennych do modeli wprowadzono jedynie wielkości względne. Dlatego też, z przedstawionej listy, w badaniu pominięto wskaźnik kapitału pracującego, który wyrażony jest w jednostce pieniężnej (zł).

Zgodnie z literaturą przedmiotu przyjęło się dzielić próbę badawczą na dwie rozłączne grupy – uczącą i testową. Pierwsza z nich służy do budowy modeli oraz weryfikacji ich skuteczności klasyfikacyjnej. Z kolei na podstawie drugiej grupy dokonywana jest ocena skuteczności klasyfikacyjnej oszacowanych modeli (Hołda, 2006). Jednakże, ze względu na ograniczoną próbę badawczą, nie dokonano podziału i analiza nie obejmowała próby testowej, czego efektem może być zawyżona trafność modeli.

Tabela 1. Wykaz upadłych spółek uwzględnionych w prognozowaniu upadłości

Branża	Spółka	Branża	Spółka
Budownictwo	ABM Solid S.A.	Przemysł drzewny i papierniczy	Drewex S.A.
	Hydrobudowa Polska S.A.		Swarzędz Meble w likwidacji S.A.
	Intakus S.A.	Przemysł farmaceutyczny	Grupa Kolastyna S.A.
	PBG S.A.	Przemysł – inne	Huta Szkła Gospodarczego IRENA S.A.
	Budostal-5 S.A.		Krościeńskie Huty Szkła Krosno S.A.
Handel detaliczny	Bomi S.A.	Przemysł lekki	Zakłady Lniarskie ORZEŁ S.A.
	Monnari Trade S.A.	Przemysł metalowy	Odlwenie Polskie S.A.
Handel hurtowy	Advadis S.A.	Przemysł surowcowy	Dolnośląskie Surowce Skalne
	Firma Handlowa JAGO S.A.	Usługi inne	Zakłady naprawcze Taboru Kolejowego w Łapach S.A. w upadłości likwidacyjnej
Hotele i restauracje	Polrest S.A.		
Informatyka	Pronox Technology S.A.		
	Techmex S.A.		

Źródło: opracowanie własne.

Tabela 2. Wykaz wskaźników finansowych publikowanych w sprawozdaniach finansowych przez spółki notowane na Giełdzie Papierów Wartościowych w Warszawie

Lp.	Wskaźnik	Lp.	Wskaźnik
1	Marża zysku brutto ze sprzedaży	11	Rotacja zapsów
2	Marża zysku operacyjnego	12	Cykl operacyjny
3	Marża zysku brutto	13	Rotacja zobowiązań
4	Marża zysku netto	14	Cykl konwetrsji gotówki
5	Stopa zwrotu z kapitału własnego	15	Rotacja aktywów obrotowych
6	Stopa zwrotu z aktywów	16	Rotacja aktywów
7	Wskaźnik płynności bieżącej	17	Wskaźnik pokrycia majątku
8	Wskaźnik płynności szybkiej	18	Stopa zadłużenia
9	Wskaźnik podwyższonej płynności	19	Wskaźnik obsługi zadłużenia
10	Rotacja należności	20	Dług/EBITDA

Źródło: opracowanie własne.

Do każdej upadłej spółki dobrano zdrowe przedsiębiorstwo, kierując się kryteriami:

– ta sama branża,

– jak najbardziej zbliżony poziom aktywów,
– jak najbardziej zbliżona liczba zatrudnionych pracowników.

W ramach przeprowadzonej analizy natknięto się na następujące rodzaje braków danych:

– brak wartości pojedynczych wskaźników finansowych na dany okres sprawozdawczy (imputacja pozycyjna) – w zbiorach danych wykorzystanych do prognozowania upadłości z rocznym oraz dwuletnim wyprzedzeniem czasowym tego typu braki danych stanowiły odpowiednio 9% oraz 4% całego zbioru danych,
– brak całego sprawozdania finansowego na dany okres sprawozdawczy (imputacja kompleksowa) – problem dotyczył dwóch przedsiębiorstw w analizie z dwuletnim wyprzedzeniem czasowym,
– niezgodność okresu sprawozdawczego danego przedsiębiorstwa z pozostałymi spółkami – problem ten dotyczył jednego przedsiębiorstwa (zarówno w analizie z rocznym, jak i z dwuletnim wyprzedzeniem czasowym).

W sytuacji braków pojedynczych wartości wskaźników finansowych, zarówno w przypadku zdrowych, jak i upadłych przedsiębiorstw, porównaniu poddano dwa sposoby imputacji. W pierwszym przypadku braki danych uzupełniano medianami obliczonymi dla wszystkich obserwacji (dla wszystkich przedsiębiorstw łącznie[2] – tzw. imputacja z wykorzystaniem mediany), zaś w drugim – obliczano mediany osobno dla zdrowych i upadłych podmiotów[3] (tzw. imputacja medianą warunkową). W przypadku zdrowych przedsiębiorstw zastosowano również trzecie podejście – ponowny dobór zdrowych przedsiębiorstw w taki sposób, by wyeliminować brakujące dane. W podejściu tym zrezygnowano z kryterium jak najlepszego dopasowania zdrowych przedsiębiorstw do bankrutów na rzecz eliminacji braków danych.

Spółki, dla których brakowało całego sprawozdania finansowego, pomijano w dalszej części analizy. Jednak w sytuacji, gdy brak sprawozdania dotyczył spółki zdrowej i nie wynikał z niedopełnienia przez nią obowiązku, a z faktu, że spółka ta przyjmuje inny okres sprawozdawczy (inny niż jak standardowo przyjęło się stosować – koniec roku kalendarzowego), porównano dwa podejścia. Zastosowano dwa podejścia. W pierwszym przypadku sprawdzono, czy należy pominąć w analizie upadłe przedsiębiorstwo, do którego dobrano zdrowy podmiot wykazujący się jednak brakiem sprawozdania. Drugie podejście opierało się na stworzeniu tzw. wirtualnego przed-

[2] W dalszej części pracy nazywana medianą dla ogółu.
[3] W dalszej części pracy nazywana medianą grup.

siębiorstwa, przyjmując za wartości wskaźników odpowiednio mediany grup czy mediany ogółu obserwacji.

W niniejszej pracy dobór zmiennych objaśniających do modeli logitowych oparto na metodzie kroczącej analizy logitowej. W badaniu wykorzystano dwie odmiany metody kroczącej – postępującą oraz wsteczną. W pierwszej z nich do wyboru zmiennych o statystycznie istotnym wpływie używana jest statystyka punktowa, zaś metoda krocząca wsteczna oparta jest na statystyce Walda, która podlega rozkładowi Chi-kwadrat. Mówiąc bardziej szczegółowo, początkowo w metodzie kroczącej postępującej w modelu znajduje się zaledwie wyraz wolny. Dana zmienna objaśniająca zostaje uwzględniona w modelu, gdy jej wkład w prognozę (istotność statystyczna określona przez *p*-value) jest wyższy od ustalonej granicznej wartości $p1$ wprowadzania ($p < p1$ wprowadzania). Z kolei w metodzie wstecznej model zawiera początkowo wszystkie potencjalne zmienne objaśniające, a następnie są one sukcesywnie usuwane na podstawie oceny statystycznej istotności ich wpływu. Jeżeli wkład prognostyczny danego wskaźnika jest gorszy niż graniczna wartość $p2$ usuwania ($p > p2$ usuwania), zostaje on wyeliminowany z modelu. W badaniu wartości $p1$ wprowadzania oraz $p2$ usuwania zostały ustalone na domyślnym poziomie, równym 0,05.

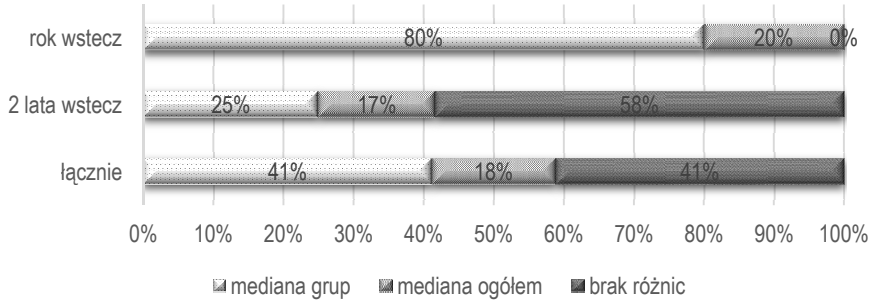
Jako dodatkowy wariant analizy, w celu jej pogłębienia, dokonano budowy modeli po wcześniejszej selekcji zmiennych objaśniających. W pierwszym przypadku uwzględniano wyniki analizy wariancji opartej na teście *F* Fishera–Snedecora. Pomijano te wskaźniki finansowe, których średnie wartości w grupie bankrutów i w grupie zdrowych przedsiębiorstw były równe, a zatem nie różnicowały obu grup (więcej o analizie wariancji w: Górecki, 2011; Aczel, 2009; Koronacki, Mielniczuk, 2006). W drugim przypadku w grupie zmiennych objaśniających pomijano wskaźniki: cykl operacyjny oraz cykl konwersji gotówki, gdyż stanowią one liniową kombinację innych wskaźników zawartych w pierwotnej grupie zmiennych. W trzecim przypadku zastosowano równocześnie obie opisane powyżej modyfikacje.

4. WYNIKI BADANIA

Łącznie otrzymano 74 poprawne modele[4] – 26 do prognozowania z rocznym oraz 48 z dwuletnim wyprzedzeniem czasowym. Porównując je pod względem skuteczności klasyfikacyjnej, kierowano się w pierwszej kolejności zasadą, że poprawność klasyfikacji upadłych przedsiębiorstw jest

[4] Model uznaje się za poprawny, gdy analiza doboru zmiennych objaśniających do modelu wykazała statystyczną istotność wpływu co najmniej dwóch zmiennych objaśniających.

ważniejsza niż poprawność klasyfikacji zdrowych podmiotów. Wynika to z faktu, że uznanie zdrowego przedsiębiorstwa za upadłe pociąga za sobą niższe koszty niż sytuacja, kiedy upadłość ogłasza przedsiębiorstwo, w którego przypadku model prognozuje dobrą kondycję finansową w okresie roku czy dwóch najbliższych lat[5]. Warto również zaznaczyć, że każdorazowo porównywano pary analogicznych modeli[6].



Wykres 1. Odsetek modeli o większej skuteczności klasyfikacji w zależności od uzupełniania braków danych – medianą grup lub medianą ogółu obserwacji

Źródło: opracowanie własne.

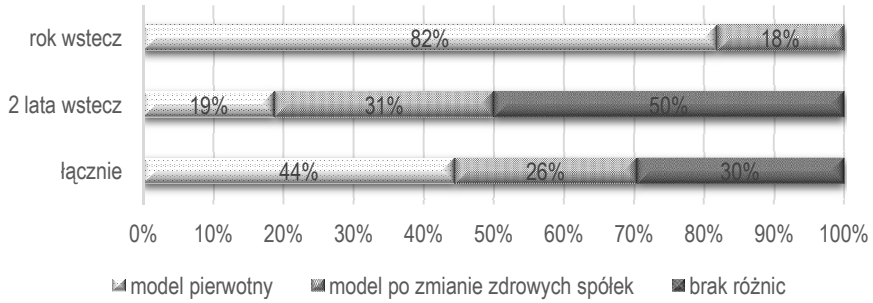
Porównując poprawność klasyfikacyjną modeli oszacowanych na danych, których braki uzupełniano medianami grup oraz medianami ogółu obserwacji, lepiej wypadły modele bazujące na pierwszej z koncepcji. Dla analizy z dwuletnim wyprzedzeniem czasowym różnice pomiędzy modelami nie były znaczące, gdyż tylko w 25% przypadków metoda oparta na medianach grup obserwacji wypadła korzystniej, a aż w 58% porównywanych parach modeli nie zaobserwowano różnic pomiędzy modelami (wykres 1). Niemniej jednak, dla roku przed ogłoszeniem bankructwa, różnice są znaczące. Ponieważ w 80% przypadków lepsze wyniki uzyskano przy zastosowaniu mediany grup, a tylko w 20% przypadków – mediany ogółu obserwacji.

[5] Tzn. w przypadku modeli, w których klasyfikacje zarówno upadłych, jak i zdrowych przedsiębiorstw różniły się, za lepszy model uznawany był ten o większej skuteczności klasyfikacji bankrutów, nawet jeśli pod względem zdrowych przedsiębiorstw wypadał gorzej.

[6] Pary analogicznych modeli to dwa modele, które różnią się pod względem tylko jednej cechy. Przykładowo, przy analizie metody wstecznej i postępującej porównywano modele, które pod innymi względami (tzn. sposób uzupełniania danych, okres analizy, rodzaj ewentualnej modyfikacji zbioru zmiennych objaśniających, rodzaj ewentualnej modyfikacji doboru zdrowych przedsiębiorstw, czy ewentualne usunięcie podmiotów) nie różniły się między sobą.

Przeciętna poprawność klasyfikacyjna modeli uzupełnianych medianami grup obserwacji w analizie z rocznym wyprzedzeniem czasowym wyniosła ok. 85%, a zdrowe przedsiębiorstwa zostały poprawnie rozpoznawane przez te modele średnio z ok. 87% poprawnością. W przypadku analizy z dwuletnim wyprzedzeniem czasowym średnie wartości poprawności klasyfikacyjnych wynosiły odpowiednio 71% dla upadłych oraz 95% dla zdrowych przedsiębiorstw.

Zastosowanie mediany wyznaczonej dla ogółu obserwacji kilkukrotnie skutkowało tym, iż nie otrzymano modelu, gdyż wyniki analizy doboru zmiennych objaśniających wskazywały istotność wpływu tylko jednej zmiennej. W przypadku stosowania mediany wyznaczonej osobno dla grupy zdrowych oraz upadłych przedsiębiorstw nie spotkano się z tego typu problemem, co jest dużą zaletą tej metody.



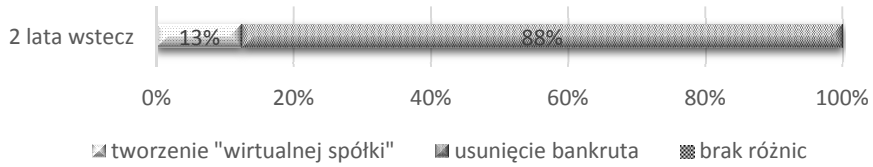
Wykres 2. Odsetek modeli o większej skuteczności klasyfikacji w zależności czy dokonano zmiany zdrowych spółek, czy nie

Źródło: opracowanie własne.

Dokonując porównania modeli, można również zaobserwować, iż podczas doboru zdrowych spółek nie warto kierować się liczbą braków danych. Porównano bowiem modele oszacowane na pierwotnych danych, które zawierały spółki dobrane na podstawie głównych kryteriów parowania (tj. branża, poziom aktywów oraz liczba zatrudnionych pracowników), z modelami, które zostały oszacowane na zbiorze danych, w którym zdrowe przedsiębiorstwa o największej ilości braków danych zostały zastąpione innymi spółkami (będącymi w drugiej kolejności pod względem wspomnianych kryteriów doboru). Analiza ta wykazała, iż zmiana zdrowych spółek nie wpłynęła na zwiększenie skuteczności klasyfikacji upadłych przedsiębiorstw, w 30% modeli dała takie same wyniki, a w 44% modeli – gorsze (wykres 2). Jednakże zauważono różnice w wynikach pomiędzy analizą dokonaną dla roku oraz 2 lat wstecz. Wśród modeli prognozujących upadłość

z rocznym wyprzedzeniem czasowym niemal jednoznacznie (w 80% przypadków) można stwierdzić lepszą skuteczność klasyfikacyjną modeli pierwotnych. Z kolei w przypadku modeli prognozujących bankructwo z dwuletnim wyprzedzeniem czasowym, w jednym na pięć przypadków porównywanych par modeli, model pierwotny okazał się lepszy od modeli opartych na danych z mniejszą ilością braków danych, a w co drugiej parze porównywanych modeli skuteczności klasyfikacyjne nie różniły się.

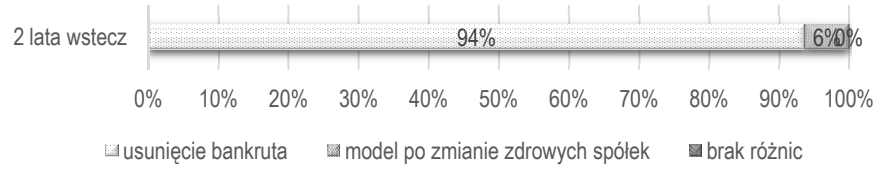
Warto jednak podkreślić, że różnice w poprawności klasyfikacyjnej porównywanych modeli nie były znaczące. Modele oparte na pierwotnych danych rozpoznały poprawnie średnio 84% (dla analizy z rocznym wyprzedzeniem czasowym) oraz 72% (dla analizy z dwuletnim wyprzedzeniem czasowym) upadłych przedsiębiorstw, podczas gdy średnia poprawność klasyfikacyjna modeli oszacowanych po ponownym doborze zdrowych spółek była w obu przypadkach niższa o 2 punkty procentowe. Z kolei pod względem średniej poprawności klasyfikacyjnej zdrowych spółek porównywane grupy modeli różniły się o 1 punkt procentowy dla roku wstecz oraz nie różniły się w przypadku analizy z dwuletnim wyprzedzeniem czasowym.



Wykres 3. Odsetek modeli o większej skuteczności klasyfikacji spółek w zależności od redukcji zbioru danych lub tworzenia „wirtualnej spółki"

Źródło: opracowanie własne.

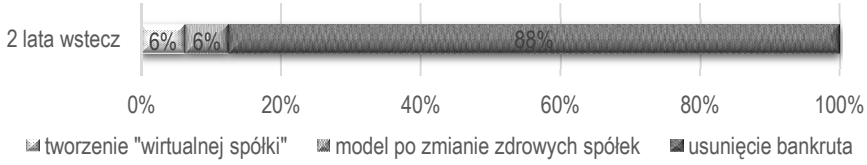
W przypadku braku całego sprawozdania finansowego zdrowej spółki, potencjalnej do tworzenia pary z danym upadłym przedsiębiorstwem, nie warto jest kreować „wirtualnego sprawozdania", zastępując wszystkie wskaźniki finansowe medianami. Lepszym rozwiązaniem jest eliminacja upadłego podmiotu ze zbioru danych. Wniosek ten wynika z faktu, iż porównując poprawność klasyfikacji upadłych przedsiębiorstw modeli zawierających „wirtualne sprawozdanie finansowe", z wynikami klasyfikacji modeli oszacowanych po wykluczeniu ze zbioru danych bankrutów „bez pary", aż w 88% przypadków otrzymano gorsze wyniki (wykres 3). Pomimo mniejszej ilości obserwacji poprawność klasyfikacyjna kształtowała się średnio na poziomie 73% dla upadłych i 80% dla zdrowych przedsiębiorstw.



Wykres 4. Odsetek modeli o większej skuteczności klasyfikacji w zależności od redukcji zbioru danych lub zmiany zdrowych spółek

Źródło: opracowanie własne.

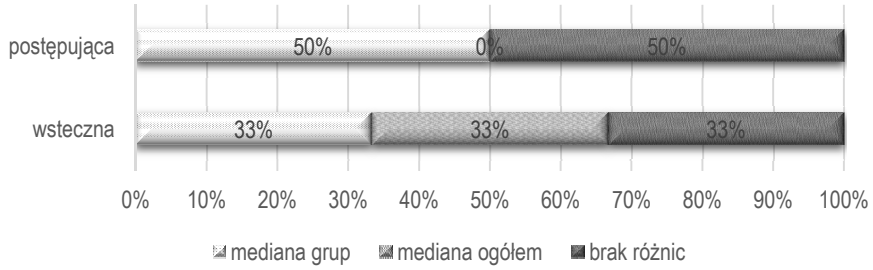
Porównując podejście opierające się na redukcji zbioru obserwacji o przedsiębiorstwa, których problem braków danych dotyczył w największym stopniu, z podejściem bazującym na zmianie zdrowych spółek, niemal za każdym razem lepszy okazał się model bazujący na mniejszej ilości obserwacji (wykres 4). Można zatem wnioskować, że modele są bardziej odporne na mniejszą ilość danych niż na mniejsze dopasowanie profilem zdrowych spółek do bankrutów. Jednakże, podobnie jak w poprzednich przypadkach, średnie wartości poprawności klasyfikacyjnej grup modeli nie różniły się znacząco między sobą. Zarówno w przypadku zdrowych, jak i upadłych przedsiębiorstw była to różnica 2 punktów procentowych na korzyść modeli szacowanych po uprzednim usunięciu bankrutów, generujących problem niekompletności danych.



Wykres 5. Odsetek modeli o większej skuteczności klasyfikacji modeli w zależności od tworzenia „wirtualnej spółki” lub zmiany zdrowych spółek

Źródło: opracowanie własne.

Porównując zatem trzy rozpatrywane powyżej podejścia: tworzenie „wirtualnej spółki”, zmiana zdrowych spółek oraz usunięcie bankrutów, w 88% przypadków najlepszym rozwiązaniem było podejście trzecie (wykres 5).



Wykres 6. Odsetek modeli o większej skuteczności klasyfikacji modeli w zależności od uzupełniania braków danych – medianą grup lub medianą ogółu obserwacji oraz przyjętej metody doboru zmiennych

Źródło: opracowanie własne.

Warto również zaznaczyć, iż porównując skuteczności klasyfikacyjne modeli w zależności od sposobu doboru zmiennych objaśniających, zarówno w przypadku metody krokowej postępującej, jak i metody krokowej wstecznej, trudno jest jednoznacznie stwierdzić, czy lepszym podejściem jest zastosowanie mediany grup, czy mediany ogółu obserwacji. W przypadku metody postępującej w połowie przypadków lepszym podejściem okazało się uzupełnianie braków medianami grup, aczkolwiek w pozostałych 50% porównywanych modeli nie zauważono różnic w poprawności klasyfikacyjnej (wykres 6). Z kolei w przypadku metody wstecznej taki sam odsetek modeli przemawiał za medianą grup, jak za medianą ogółem obserwacji (33%). Pozostałe 33% nie wykazały różnic w metodach.

PODSUMOWANIE

Podsumowując, modele logitowe, wykorzystane do prognozowania upadłości przedsiębiorstw, są bardziej odporne na większe rozbieżności w profilach zdrowego i upadłego przedsiębiorstwa. Tym samym, w przypadku imputacji pozycyjnej, nie warto modyfikować zbioru obserwacji w celu ograniczenia niekompletności danych. Jednakże w przypadku braku zdrowej spółki, która mogłaby posłużyć jako zdrowy odpowiednik dla danego bankruta, lepszym podejściem od kreowania „wirtualnej zdrowej spółki" jest pominięcie w badaniu owego upadłego przedsiębiorstwa, co świadczy o odporności modelu logitowego na mniejszą ilość danych. Rozpatrując łącznie wyniki analizy z rocznym oraz dwuletnim wyprzedzeniem czasowym, można wnioskować, że lepszym podejściem jest zastosowanie imputacji medianą warunkową (grup obserwacji) niż medianą dla ogółu obserwacji.

Rozważając analizę dla różnych okresów, w przypadku analizy dla danych rok przed ogłoszeniem upadłości lepszym podejściem jest zastosowanie mediany grup niż mediany ogółu obserwacji. Modele oparte na pierwotnym zbiorze danych (danych niekompletnych) cechują się lepszą poprawnością klasyfikacyjną niż modele oszacowane na zmodyfikowanym zbiorze danych, w celu redukcji problemu niekompletnych danych.

W analizie prognozowania upadłości z dwuletnim wyprzedzeniem czasowym ponownie lepszym podejściem jest zastosowanie mediany grup obserwacji, niż mediany wyznaczonej na całym zbiorze danych. Warto też dokonać zmiany zdrowych spółek w celu ograniczenia brakujących danych. W przypadku imputacji kompleksowej lepszym rozwiązaniem jest usunięcie upadłego podmiotu generującego problem niekompletności danych niż kreowanie „wirtualnego zdrowego przedsiębiorstwa" czy zmiana zdrowych spółek.

LITERATURA

Aczel A. D. (2009), *Complete business statistics*, McGraw-Hill/Irwin, New York.

Balicki A. (2004), *Metody imputacji brakujących danych w badaniach statystycznych*, „Wiadomości Statystyczne", 9.

Beręsewicz M. (2010), *Imputacja jako sposób rozwiązywania problemów braków danych*, e-wydawnictwo 2011.

Bragg S. M. (2010), *Wskaźniki w analizie działalności przedsiębiorstwa*, Oficyna, Warszawa.

Górecki T. (2011), *Podstawy statystyki z przykładami w R*, Wydawnictwo BTC, Legionowo.

Gruszczyński M., Kuszewski T., Podgórska M. (2009), *Ekonometria i badania operacyjne*, PWN, Warszawa.

Hołda A. (2006), *Zasada kontynuacji działalności i prognozowanie upadłości w polskich realiach gospodarczych*, Wydawnictwo Akademii Ekonomicznej w Krakowie.

Jerzemowska M. (2013), *Analiza ekonomiczna w przedsiębiorstwie*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Kalton G., Kasprzyk D. (1982), *Imputing for Missing Survey Responses*, *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Koop G. (2011), *Wprowadzenie do ekonometrii*, Oficyna, Warszawa.

Korol T. (2010), *Systemy ostrzegania przedsiębiorstw przed ryzykiem upadłości*, Oficyna, Warszawa.

Korol T. (2013), *Nowe podejście do analizy wskaźnikowej w przedsiębiorstwie*, Oficyna, Warszawa.

Koronacki J., Mielniczuk J. (2006), *Statystyka dla studentów kierunków technicznych i przyrodniczych*, Wydawnictwo Naukowo-Techniczne, Warszawa.

Larose D. T. (2008), *Metody i modele eksploracji danych*, PWN, Warszawa.

Lewandowski R., Wołowski P. (2011), *Prawo upadłościowe i naprawcze*, Wydawnictwo C. H. Beck, Warszawa.

Longford N. T. (2005), *Missing Data and Small Area Estimation*, Springer Science + Business Media, Inc.

Młodak A. (2010), *Imputacja danych w spisach powszechnych*, „Wiadomości Statystyczne", 8.

Piasecki T. (2014), *Metody imputacji w badaniach gospodarstw domowych*, „Wiadomości Statystyczne", nr 9.

Piasecki T. (2013), *Imputacja dochodów w badaniach statystyki publicznej dotyczącej gospodarstw domowych*, GUS.

Podstawka M. (2010), *Finanse. Instytucje, instrumenty, podmioty, rynku, regulacje*, PWN, Warszawa.

Prusak B. (2011), *Ekonomiczna analiza upadłości przedsiębiorstw. Ujęcie międzynarodowe*, CeDeWu.pl.

Rozporządzenie Ministra Finansów z dnia 19 lutego 2009 r. w sprawie informacji bieżących i okresowych przekazywanych przez emitentów papierów wartościowych oraz warunków uznawania za równoważne informacji wymaganych przepisami prawa państwa niebędącego państwem członkowskim, DzU 2009, Nr 33, poz. 259.

Rószkiewicz M. (2002), *Metody ilościowe w badaniach marketingowych*, PWN, Warszawa.

Ustawa z dnia z dnia 29 września 1994 r. o rachunkowości, DzU 1974, Nr 121, poz. 591 z późn. zm.

Ustawa z dnia 15 września 2000 r. Kodeks spółek handlowych, DzU 2000, Nr 94, poz. 1037 z późn. zm.

Ustawa z dnia 29 lipca 2005 r. o ofercie publicznej i warunkach wprowadzania instrumentów finansowych do zorganizowanego systemu obrotu oraz o spółkach publicznych, DzU 2005, Nr 184, poz. 1539 z późn. zm.

Ustawa z dnia 29 lipca 2005 r. o obrocie instrumentami finansowymi, DzU 2005, Nr 183, poz. 1538.

Wędzki D. (2009), *Analiza wskaźnikowa sprawozdania finansowego*, t. 2, Oficyna, Kraków.

Zeliaś A., Pawełek B., Wanat S. (2003), *Prognozowanie ekonomiczne. Teoria, przykłady, zadania*, PWN, Warszawa.

THE INFLUENCE OF DATA IMPUTATION METHODS ON THE CLASSIFICATION EFFICIENCY OF THE LOGIT MODEL USED FOR FORECASTING THE BANKRUPTCY OF COMPANIES

A b s t r a c t. Forecasting the bankruptcy of companies exposes the missing data problem, which applies chiefly to entities having financial problems, who wish to conceal thereby their bad situation. One of the methods of making up incomplete data is imputation. The aim of the paper is to present different data imputation variants and to compare their influence on the classification efficiency of one of the statistical bankruptcy forecasting methods – the logit model. The results have shown that the best approach is to use the median as determined separately for healthy and bankrupt companies.

K e y w o r d s: bankruptcy, forecasting bankruptcy, logit model, imputation, missing data estimation, model classification efficiency.