Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 395 · 2020

Andrzej Paliński

AGH w Krakowie
Wydział Zarządzania
Katedra Informatyki Stosowanej
palinski@zarz.agh.edu.pl

WYKRYWANIE ZAGROŻENIA UPADŁOŚCIĄ JAKO PROBLEM KLASYFIKACJI DANYCH NIEZBALANSOWANYCH

Streszczenie: W artykule wykorzystano wybrane algorytmy uczenia maszynowego oraz techniki przygotowania danych (*preprocessing*) stosowane w klasyfikacji na zbiorach niezbalansowanych w celu oceny ich skuteczności w prognozowaniu upadłości z użyciem danych zawierających wskaźniki finansowe podmiotów gospodarczych. Trafność prognoz upadłości na pierwotnym niezbalansowanym zbiorze danych o przeważającym udziale podmiotów prowadzących działalności nad upadłymi była bliska zero. Trafność prognozowania upadłości klasyfikatorów utworzonych na zbiorach zbalansowanych była odwrotnie proporcjonalna do całkowitej trafności klasyfikacji i wahała się od 10% – dla całkowitej trafności klasyfikacji wynoszącej 93%, do 77% – dla całkowitej trafności klasyfikacji równej 49%. Lepsze wyniki klasyfikacji osiągały algorytmy *gradient boosting* i drzewo klasyfikacyjne w stosunku do sztucznej sieci neuronowej. W problemie klasyfikacji na zbiorach niezbalansowanych wystąpił efekt wymiany – albo możliwe jest zwiększenie trafności klasyfikacji upadłości kosztem nadmiarowości obiektów klasyfikowanych jako upadłe, albo – zwiększenie trafności klasyfikacji całkowitej algorytmu kosztem zmniejszenia trafności klasyfikacji samej upadłości.

Słowa kluczowe: upadłość, zbiór niezbalansowany, uczenie maszynowe, klasyfikacja, preprocessing.

JEL Classification: C53, C55, G33.

Wprowadzenie

Predykcja niewypłacalności i upadłości od dawna jest zagadnieniem badawczym ważnym z punktu widzenia inwestorów i wierzycieli. Prognozowanie upadłości podmiotów gospodarczych metodami statystycznymi zostało zapoczątkowane przez Beavera [1966], a popularność zyskało dzięki modelowi Alt-

mana [1968]. W kolejnych latach powstało wiele modeli przewidywania upadłości. Najczęściej wykorzystywano analizę dyskryminacyjną, modele logitowe, sztuczne sieci neuronowe i drzewa klasyfikacyjne [Pociecha, red., 2014; Gissel, Giacomino, Akers, 2007]. Podobnie jak w przypadku Altmana, liniowa funkcja dyskryminacyjna była dla przykładu wykorzystywana w pracach Wilcoxa [1973] czy Laitinena [1991]. Równolegle prowadzono badania z użyciem uogólnionej metody regresji i modelu logitowego lub probitowego [np. Zmijewski, 1984; Li, Miu, 2010]. W polskich realiach badania nad prognozowaniem upadłości były prowadzone m.in. przez Mączyńską [1994] i Wędzkiego [2005].

Budowa modeli predykcji upadłości napotyka dwa główne problemy. Pierwszy problem to dobór wskaźników uwzględnianych w analizie – wykorzystuje się zwykle wiedzę ekspercką, ale nie ma metody, która by jednoznacznie uzasadniła wstępny dobór wskaźników do analizy. Drugi problem to niezbalansowany zbiór danych historycznych służących budowie modeli – w próbce zdecydowanie przeważają firmy niezagrożone upadłością. Tym samym modele lepiej prognozują brak zagrożenia upadłością niż samą upadłość.

Wraz z gwałtownym rozwojem technologii informatycznych w latach dziewięćdziesiątych ubiegłego wieku rozpoczęto intensywne badania nad wykorzystaniem technik sztucznej inteligencji do prognozowania upadłości. Klasyczne modele ekonometryczne zaczęto zastępować sztucznymi sieciami neuronowymi i drzewami klasyfikacyjnymi [m.in. Serrano-Cinca, 1996; Geng, Bose, Chen, 2015]. Narzędzia uczenia maszynowego lepiej radzą sobie z szybko rosnącymi zbiorami nieoczyszczonych danych, tzn. zawierającymi błędne lub niekompletne dane.

Systemy uczące samodzielnie mogą dokonywać doboru wskaźników predykcji upadłości spośród dowolnie dużego zbioru potencjalnych zmiennych, co częściowo likwiduje problem arbitralności w doborze wskaźników. Drugi problem – niezbalansowanego zbioru danych − jest znacznie trudniejszy do rozwiązania. Wypracowano kilka grup metod radzenia sobie z tym problemem. Dzielą się one na metody, w których modyfikuje się dane w celu zbalansowania zbioru danych, metody, w których modyfikuje się algorytm, wzmacniając klasyfikację klasy mniejszościowej, oraz mieszane, tzw. wrażliwe na koszt [He, Garcia, 2009; Galar i in., 2012]. Skuteczność pierwszej grupy metod zostanie przetestowana na zbiorze podmiotów gospodarczych zagrożonych upadłością w dalszej części niniejszego artykułu. Jako algorytmy uczenia maszynowego wykorzystane zostaną drzewa klasyfikacyjne: proste oraz wzmocnione, a także sieć neuronowa.

Celem artykułu jest ocena skuteczności metod uczenia maszynowego i metod przygotowania danych (preprocessingu) stosowanych w klasyfikacji na zbiorach niezbalansowanych w odniesieniu do prognozowania upadłości podmiotów gospodarczych. Dalszy układ opracowania jest następujący: w pierwszej części przedstawiono charakterystykę zbiorów danych, w drugiej zaprezentowano charakterystykę metod klasyfikacji w zbiorach niezbalansowanych. W kolejnej części zawarto wyniki badań prognozowania upadłości metodami uczenia maszynowego. Artykuł kończy podsumowanie zawierające ocenę wyników.

1. Klasyfikacja na podstawie niezbalansowanych danych

Zbiór niezbalansowany to taki zbiór, w którym klasa mniejszościowa zawiera znacznie mniej przykładów niż pozostałe klasy. Zazwyczaj głównym celem jest poprawne rozpoznawanie przykładów z klasy mniejszościowej, np. niewypłacalność kredytobiorcy, upadłość, oszustwa ubezpieczeniowe, podatkowe itp. Klasyfikacja i uczenie maszynowe na zbiorze niezbalansowanym stanowi duży problem, gdyż algorytmy, optymalizując funkcję celu, poprawiają trafność klasyfikacji, nie biorą pod uwagę, do jakiej klasy należą poszczególne przykłady. Tym samym klasa mniejszościowa traci na znaczeniu. Główne trudności w fazie uczenia na zbiorach niezbalansowanych wynikają z tego, że:

- algorytmy uczące zakładają zrównoważone dane,
- strategie klasyfikacji sprzyjają klasom większościowym,
- występuje trudność w odróżnieniu błędnych (zabrudzonych) danych od przykładów z klasy mniejszościowej.

Grupy metod rozwiązujących problem niezbalansowanych danych spotykane w literaturze to [He, Garcia, 2009; Galar i in., 2012; Mahani, Ali, 2020; Ziemba, 2013]:

- Metody modyfikacji danych, tzw. podejście zewnętrzne, w którym dane przetwarza się przed zastosowaniem klasyfikatorów. Dane są niezależne od wybranego algorytmu uczenia klasyfikatora.
- Metody modyfikacji algorytmów, tzw. podejście wewnętrzne, w którym klasyczne algorytmy wzbogaca się o mechanizmy uwzględniające dysproporcję klas. W ramach tego podejścia stosuje się indukcyjne ukierunkowanie (*inductive bias*) oraz uczenie, w którym bierze się pod uwagę tylko przykłady z klasy mniejszościowej, omijając przykłady z pozostałych klas.
- Transformacje do zadania wrażliwego na koszt (*cost-sensitive learning*), stanowiące kombinację dwóch wcześniejszych metod. Dane wejściowe są mo-

dyfikowane przez nadanie im różnych wag (kosztów), a algorytmy uczenia są wzbogacane o mechanizmy uwzględniające różne wagi nadane obserwacjom. Metoda ta jest stosowana w przypadkach, w których występują znaczne różnice w kosztach dotyczących błędnych decyzji.

W dalszej części artykułu zostanie zastosowane podejście zewnętrzne polegające na preprocessingu (modyfikacji) niezbalansowanych danych. Możliwych jest kilka metod modyfikacji niezbalansowanych zbiorów danych [Maalouf, Trafalis, 2011; Mahani i Ali, 2020; Ziemba, 2013], z których najczęściej stosuje się:

- losową eliminację (*random-undersampling*) obiektów z klasy większościowej; wadą tej metody jest groźba odrzucenia potencjalnie ważnych danych;
- eliminację świadomą (*Neighbour Cleaning Rule*) wykorzystującą algorytm K-NN najbliższych sąsiadów; dla każdego przykładu ze zbioru danych jest znajdowanych trzech najbliższych sąsiadów (3NN); jeśli przykład należy do klasy większościowej, a 3NN wskazuje na klasę mniejszościową, to taki przykład jest usuwany; jeśli przykład należy do klasy mniejszościowej i algorytm 3NN błędnie go sklasyfikuje, to 3 sąsiednie obserwacje są usuwane;
- próbkowanie losowe (*random-oversampling*), które polega na replikacji obserwacji z klasy mniejszościowej poprzez losowanie ze zwracaniem; wadą tej metody jest wzrost prawdopodobieństwa przeuczenia, ponieważ tworzy się dokładne kopie istniejących przykładów;
- inteligentne próbkowanie poprzez generowanie syntetycznych obserwacji z wykorzystaniem przykładów z klasy zdominowanej; jedną z najpopularniejszych metod jest algorytm SMOTE (*Synthetic Minority Oversampling Technique*), w którym dla każdej próbki z klasy mniejszościowej jest wprowadzany przykład syntetyczny wzdłuż odcinków łączących najbliższych dwóch sąsiadów (2NN) z klasy mniejszościowej.

W przypadku zbiorów niezbalansowanych konieczne jest także przyjęcie bardziej szczegółowych miar trafności klasyfikacji niż tylko trafność całkowita. Ocena trafności klasyfikacji opiera się na tzw. macierzy pomyłek (tabela 1).

Tabela 1. Macierz pomyłek

		Rzeczywiste		
		0 – klasa negatywna	**1 – klasa pozytywna**	Suma
Przewidywane	**0 – klasa negatywna**	prawdziwie negatywne (TN)	fałszywie negatywne (FN)	TN + FN
	1 – klasa pozytywna	fałszywie pozytywne (FP)	prawdziwie pozytywne (TP)	FP + TP
Suma		TN + FP	FN + TP	

Źródło: Opracowanie własne na podstawie: Lantz [2015].

Na podstawie macierzy pomyłek wprowadza się poniższe wskaźniki trafności klasyfikacji [Lantz, 2015].

1. Dokładność (*accuracy*) – stosunek poprawnie zakwalifikowanych obserwacji do całości próby:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2. Precyzja (*precision*) – proporcja obserwacji pozytywnych, które zostały zakwalifikowane prawidłowo:

$$precision = \frac{TP}{TP + FP} \tag{2}$$

3. Czułość (*sensitivity*, *recall*, *true positive rate*) – relacja poprawnie sklasyfikowanych klasyfikacji pozytywnych do ogólnej liczby obserwacji pozytywnych wprowadzonych do modelu. Wysoka wartość tej miary wskazuje na skuteczne wykrywanie przypadków klasy pozytywnej:

$$sensitivity = \frac{TP}{TP + FN} \tag{3}$$

4. Swoistość (*specificity*, *true negative rate*) – stosunek prawidłowo sklasyfikowanych obserwacji klasy negatywnej do wszystkich negatywnych predykcji:

$$specificity = \frac{TN}{TN + FP} \tag{4}$$

W dalszej części zostaną wykorzystane dwa wskaźniki: *accuracy* mierzący ogólną trafność klasyfikacji oraz *sensitivity* mierzący poprawność klasyfikacji upadłości.

2. Zbiór danych wykorzystany w badaniach

W artykule wykorzystano ogólnie dostępny zbiór danych umieszczony w repozytorium danych uczenia maszynowego (Machine Learning Depository, [www 1]). Dane te posłużyły pierwotnie do testowania trafności prognostycznej nowego podejścia w uczeniu maszynowym zaprezentowanego w pracy Zięby, Tomczaka i Tomczaka [2016]. Dane zostały pobrane z serwisu EMIS (*Emerging Markets Information Service*) i obejmują okres 2000-2012.

Tabela 2. Wskaźniki finansowe wykorzystane w analizie

Lp.	Opis	Lp.	Opis
Atr1	Zysk netto/aktywa ogółem	Atr33	Koszty operacyjne/zobowiązania krótkoterminowe
Atr2	Zobowiązania ogółem/aktywa ogółem	Atr34	Koszty operacyjne/zobowiązania ogółem
Atr3	Kapitał obrotowy/aktywa ogółem	Atr35	Zysk ze sprzedaży/aktywa ogółem
Atr4	Majątek obrotowy/zobowiązania krótkoterminowe	Atr36	Sprzedaż/aktywa ogółem
Atr5	[(Należności + inwestycje krótkoterminowe – zobowiązania krótkoterminowe)/(koszty operacyjne – amortyzacja)] * 365	Atr37	(Majątek obrotowy – zapasy)/zobowiązania długoterminowe
Atr6	Zyski zatrzymane/aktywa ogółem	Atr38	Kapitał stały/aktywa ogółem
Atr7	EBIT/aktywa ogółem	Atr39	Zysk ze sprzedaży/sprzedaż
Atr8	Księgowa wartość kapitału/zobowiązania ogółem	Atr40	(Majątek obrotowy – zapasy – należności)/zobowiązania krótkoterminowe
Atr9	Sprzedaż/aktywa ogółem	Atr41	Zobowiązania ogółem/(zysk operacyjny + amortyzacja) * 12/365
Atr10	Kapitał własny/aktywa ogółem	Atr42	Zysk operacyjny/sprzedaż
Atr11	(zysk brutto + przychody nadzwyczajne + koszty finansowe)/aktywa ogółem	Atr43	Rotacja należności + rotacja zapasów w dniach
Atr12	Zysk brutto/zobowiązania krótkoterminowe	Atr44	(Należności * 365)/sprzedaż
Atr13	(Zysk brutto + amortyzacja)/sprzedaż	Atr45	Zysk netto/zapasy
Atr14	(Zysk brutto + odsetki)/aktywa ogółem	Atr46	(Majątek obrotowy – zapasy)/zobowiązania krótkoterminowe
Atr15	(Zobowiązania ogółem * 365)/(zysk brutto + amortyzacja)	Atr47	(Zapasy * 365)/koszty wyrobów sprzedanych
Atr16	(zysk brutto + amortyzacja)/aktywa ogółem	Atr48	(Zysk operacyjny – amortyzacja)/aktywa ogółem
Atr17	Aktywa ogółem/zobowiązania ogółem	Atr49	(Zysk operacyjny – amortyzacja)/sprzedaż
Atr18	Zysk brutto/aktywa ogółem	Atr50	Majątek obrotowy/zobowiązania ogółem
Atr19	Zysk brutto/sprzedaż	Atr51	Zobowiązania krótkoterminowe/zobowiązania ogółem
Atr20	(Zapasy * 365)/sprzedaż	Atr52	(Zobowiązania krótkoterminowe * 365)/koszty wyrobów sprzedanych
Atr21	Sprzedaż(n)/sprzedaż(n-1)	Atr53	Kapitał własny/aktywa trwałe
Atr22	Zysk operacyjny/aktywa ogółem	Atr54	Kapitał stały/aktywa trwałe
Atr23	Zysk netto/sprzedaż	Atr55	Kapitał obrotowy
Atr24	Zysk brutto (z 3 lat)/aktywa ogółem	Atr56	(Sprzedaż – koszty wyrobów sprzedanych)/sprzedaż
Atr25	Kapitał własny – kapitał akcyjny/aktywa ogółem	Atr57	(Majątek obrotowy – zapasy – zobowiązania krótkoterminowe)/(sprzedaż – zysk brutto – amortyzacja)
Atr26	(Zysk netto + amortyzacja)/aktywa ogółem	Atr58	Koszty ogółem/sprzedaż ogółem
Atr27	Zysk operacyjny/koszty finansowe	Atr59	Długoterminowe zobowiązania/kapitał własny
Atr28	Kapitał obrotowy/aktywa trwałe	Atr60	Sprzedaż/zapasy
Atr29	Logarytm aktywów ogółem	Atr61	Sprzedaż/należności
Atr30	(Zobowiązania ogółem – gotówka)/sprzedaż	Atr62	(Zobowiązania krótkoterminowe * 365)/ sprzedaż
Atr31	(Zysk brutto + odsetki)/sprzedaż	Atr63	Sprzedaż/zobowiązania krótkoterminowe
Atr32	(Zobowiązania bieżące * 365)/koszty wyrobów sprzedanych	Atr64	Sprzedaż/aktywa trwałe

Źródło: Opracowanie na podstawie: Zięba, Tomczak, Tomczak [2016].

W badaniu wzięto pod uwagę dane obejmujące 10 007 podmiotów prowadzących działalność oraz 495 takich, które ogłosiły upadłość po 3 latach od momentu zgromadzenia danych (10 502 sprawozdania finansowe). Dane obejmują 64 wskaźniki finansowe dla każdej z firm oraz informację, czy podmiot ogłosił upadłość. Opis poszczególnych wskaźników umieszczono w tabeli 2.

3. Wyniki klasyfikacji podmiotów zagrożonych upadłością

W pierwszym kroku przeprowadzono klasyfikację, opierając się na pierwotnym niezbalansowanym zbiorze zawierającym brakujące dane. Spośród 10 502 podmiotów 495 ogłosiło upadłość, co stanowi 4,7% całości zbioru danych. Wykorzystano dwa algorytmy uczenia maszynowego: drzewo klasyfikacyjne i sztuczną sieć neuronową stworzone przez Microsoft [www 5]. Całkowita trafność klasyfikacji mierzona wskaźnikiem *accuracy* wyniosła dla poszczególnych algorytmów odpowiednio: 95,1% oraz 95,0%. Niestety żaden z algorytmów nie zakwalifikował prawidłowo ani jednego przypadku upadłości, co oznacza, że wskaźnik *sensitivity* wyniósł zero. Szczegółowe dane dotyczące klasyfikacji na podstawie zbioru niezbalansowanego zawarto w tabeli 3.

Tabela 3. Podsumowanie wyników prognozowania upadłości dla zbioru niezbalansowanego (zbiór testowy zawiera 1000 obserwacji, 1 oznacza upadłość)

	Rzeczywiste	
Prognozowane	0	1
Drzewo klasyfikacyjne		
0	951	49
1	0	0
Sieć neuronowa		
0	950	49
1	1	0

Źródło: Opracowanie własne.

W kolejnym kroku zastosowano *undersampling*, tworząc zbiór z losowo usuniętą częścią danych z klasy „zdrowych" przedsiębiorstw. Otrzymano zbiór liczący 995 podmiotów, w tym 495 tych, które ogłosiły upadłość, co stanowi 49,7% całości tego zbioru. Całkowita trafność klasyfikacji mierzona wskaźnikiem *accuracy* wyniosła dla poszczególnych algorytmów odpowiednio: 85,9% oraz 63,1%. Jeżeli chodzi o trafność klasyfikacji upadłości, to zgodnie ze wskaźnikiem *sensitivity* przypadki upadłości zostały prawidłowo sklasyfikowa-

ne przez drzewo klasyfikacyjne w 79,3% przypadków, a przez sieć neuronową tylko w 28,0% przypadków. Szczegółowe dane dotyczące klasyfikacji z wykorzystaniem zbioru zbalansowanego w wyniku *undersamplingu* są zawarte w tabeli 4.

Tabela 4. Podsumowanie wyników prognozowania upadłości dla zbioru zbalansowanego w wyniku losowego usunięcia części danych z klasy większościowej – *undersamplingu* (zbiór testowy zawiera 30% obserwacji, 1 oznacza upadłość)

	Rzeczywiste	
Prognozowane	0	1
Drzewo klasyfikacyjne		
0	137	31
1	11	119
Sieć neuronowa		
0	146	108
1	2	42

Źródło: Opracowanie własne.

Lepiej dopasowany algorytm drzewa klasyfikacyjnego użyto następnie do klasyfikacji na pierwotnym niezbalansowanym zbiorze danych. W tym przypadku całkowita zdolność klasyfikacji mierzona wskaźnikiem *accuracy* wyniosła jedynie 51,8%, podczas gdy wskaźnik *sensitivity* okazał się zadowalający i wyniósł 77,0%.

Biorąc pod uwagę dosyć wysoką wartość wskaźnika *sensitivity*, w sytuacji zakwalifikowania blisko połowy przypadków jako zagrożonych upadłością, dokonano ponownego uczenia drzewa klasyfikacyjnego na zbiorze zawierającym tylko przypadki sklasyfikowane jako zagrożone upadłością (5332 podmioty). Uzyskane rezultaty są jednak niezadowalające, gdyż co prawda wskaźnik *accuracy* wyniósł teraz 93,2%, ale wskaźnik *sensitivity* – zaledwie 23,9%.

W następnym kroku zastosowano *oversampling*, tworząc zbiór z kilkukrotnie zreplikowanymi 495 przypadkami upadłości. Otrzymano zbiór liczący 13 967 podmiotów, w tym 3960 tych, które ogłosiły upadłość, co stanowi 28,4% całości tego zbioru. Całkowita trafność klasyfikacji mierzona wskaźnikiem *accuracy* wyniosła dla poszczególnych algorytmów odpowiednio: 82,9% oraz 73,1%. Trafność klasyfikacji upadłości mierzona wskaźnikiem *sensitivity* wyniosła dla drzewa klasyfikacyjnego 57,4%, a sieci neuronowej – zaledwie 4,0%. Szczegółowe dane dotyczące klasyfikacji na podstawie zbioru zbalansowanego w wyniku *oversamplingu* zawarto w tabeli 5.

Tabela 5. Podsumowanie wyników prognozowania upadłości dla zbioru zbalansowanego w wyniku replikowania danych z klasy mniejszościowej – *oversamplingu* (zbiór testowy zawiera 1000 obserwacji, 1 oznacza upadłość)

	Rzeczywiste	
Prognozowane	0	1
Drzewo klasyfikacyjne		
0	674	115
1	56	155
Sieć neuronowa		
0	719	258
1	11	12

Źródło: Opracowanie własne.

Algorytm drzewa klasyfikacyjnego użyto następnie do klasyfikacji na pierwotnym niezbalansowanym zbiorze danych. W tym przypadku całkowita zdolność klasyfikacji mierzona wskaźnikiem *accuracy* wyniosła 90,8%, podczas gdy wskaźnik *sensitivity* wyniósł 52,5%.

W kolejnym kroku zastosowano bardziej zaawansowaną metodę replikacji klasy mniejszościowej – algorytm SMOTE. W tym celu wykorzystano bibliotekę *DMwR* języka R [www 3]. Otrzymano zbiór liczący 14 102 podmiotów, w tym 4 095 tych, które ogłosiły upadłość, co stanowi 29,0% całości tego zbioru. Całkowita trafność klasyfikacji mierzona wskaźnikiem *accuracy* wyniosła dla poszczególnych algorytmów odpowiednio: 81,0% oraz 70,8%. Jeżeli chodzi o trafność klasyfikacji upadłości, to zgodnie ze wskaźnikiem *sensitivity* przypadki upadłości zostały prawidłowo sklasyfikowane przez drzewo klasyfikacyjne w 51,5% przypadków, a przez sieć neuronową jedynie w 2,7% przypadków.

Algorytm drzewa klasyfikacyjnego użyto następnie do klasyfikacji na pierwotnym niezbalansowanym zbiorze danych. W tym przypadku całkowita zdolność klasyfikacji mierzona wskaźnikiem *accuracy* wyniosła 91,0%, podczas gdy wskaźnik *sensitivity* wyniósł 50,3%. Algorytm SMOTE nie przyczynił się do zwiększenia trafności klasyfikacji w stosunku do prostego *oversamplingu*, a uzyskane wyniki pozostały zbliżone.

Dotychczas wykorzystywany zbiór danych zawierał liczne braki danych, w szczególności dotyczące wskaźnika *atr*37, a także *atr*21 i *atr*27. Charakterystyka brakujących danych jest następująca:

- łączna liczba brakujących wartości w całym zbiorze wynosi 10 567,
- 44 atrybuty zawierają brakujące wartości,
- najwięcej brakujących wartości zawiera zmienna *atr*37, która posiada 4914 pustych wartości (45% wszystkich wartości tej zmiennej),
- stwierdzono brak przynajmniej jednej wartości w 5945 obserwacjach (co stanowi 54,5% wszystkich obserwacji).

W związku z powyższym dokonano uzupełnienia brakujących wartości metodą imputacji [Longford, 2005]. W tym celu wykorzystano bibliotekę *mice* języka R [www 4]. Jako metodę imputacji zastosowano średnią ruchomą, a jako zbiór wyjściowy wybrano zbiór uzyskany w wyniku *oversamplingu*. W odniesieniu do tak przygotowanego zbioru danych ponownie przeprowadzono klasyfikację z użyciem drzewa klasyfikacyjnego oraz sieci neuronowej. Całkowita trafność klasyfikacji mierzona wskaźnikiem *accuracy* wyniosła dla poszczególnych algorytmów odpowiednio: 69,6% oraz 67,3%. Jeżeli chodzi o trafność klasyfikacji upadłości, to wskaźnik *sensitivity* wyniósł dla drzewa klasyfikacyjnego zaledwie 16,2%, a dla sieci neuronowej – zaledwie 11,3%.

Algorytm drzewa klasyfikacyjnego użyto następnie do klasyfikacji na pierwotnym niezbalansowanym zbiorze danych. W tym przypadku całkowita zdolność klasyfikacji mierzona wskaźnikiem *accuracy* wyniosła 92,7%, podczas gdy wskaźnik *sensitivity* wyniósł jedynie 10,3%. Uzupełnienie brakujących danych nie przyczyniło się do zwiększenia trafności klasyfikacji. W odniesieniu do klasyfikacji upadłości doszło nawet do pogorszenia trafności, co może wskazywać na to, że braki danych nie są w pełni losowe.

Łączenie możliwości predykcyjnych wielu modeli jest uznawane za jedną z najbardziej skutecznych metod w zagadnieniach klasyfikacji. Metody zespołowe polegają na tworzeniu zespołu wielu tzw. słabych klasyfikatorów w jeden mocny klasyfikator cechujący się większą skutecznością niż w przypadku niezależnego działania każdego z nich [Lantz, 2015]. W przypadku klasyfikacji na zbiorach niezbalansowanych za najlepsze podejście jest uznawana metoda wzmocnienia (*boosting*) [Zięba, Tomczak, Tomczak, 2016], w przypadku której większe wagi w kolejnych iteracjach procesu uczenia są nadawane przypadkom źle sklasyfikowanym z klasy mniejszościowej. Jednym z takich algorytmów jest *gradient boosting* opracowany przez Friedmana [2001], który do klasyfikacji wykorzystuje zespół drzew klasyfikacyjnych zmodyfikowanych przez techniki agregacji i wzmacniania.

Przy użyciu biblioteki *caret* w języku R [www 2] zbudowano dwa modele *gradient boosting*. Pierwszy – na podstawie wcześniej wykorzystywanego zbioru z losowo usuniętą częścią danych z klasy większościowej (*undersampling*), w którym dodatkowo uzupełniono brakujące dane metodą średniej ruchomej. Drugi – na podstawie wcześniej używanego zbioru utworzonego w wyniku replikowania danych z klasy mniejszościowej z użyciem algorytmu SMOTE, w którym również zastosowano imputację metodą średniej ruchomej.

Algorytm *gradient boosting* uczony na zbiorze z losowo usuniętą częścią danych użyto następnie do klasyfikacji na pierwotnym niezbalansowanym zbiorze danych. W tym przypadku całkowita zdolność klasyfikacji mierzona wskaźnikiem *accuracy* wyniosła zaledwie 48,9%, podczas gdy wskaźnik *sensitivity* wyniósł aż 92,9%. Wynik można uznać za znakomity jeżeli chodzi o wykrywanie zagrożenia upadłością, problemem jest jednak to, że blisko połowa przypadków zdrowych podmiotów również została sklasyfikowana jako zagrożona upadłością.

Tabela 6. Podsumowanie wyników prognozowania upadłości na pierwotnym niezbalansowanym zbiorze danych w wyniku uczenia na zbiorach balansowanych różnymi metodami

Rodzaj zbioru uczącego	*Accuracy*	*Sensitivity*
Drzewo klasyfikacyjne – zbiór zbalansowany z losowo usuniętymi danymi z klasy większościowej (*undersampling*)	51,8%	77,0%
Drzewo klasyfikacyjne – zbiór złożony z obiektów sklasyfikowanych jako bankrut uczony ponownie	93,2%	23,9%
Drzewo klasyfikacyjne – zbiór zbalansowany z replikowanymi danymi z klasy większościowej (*oversampling*)	90,8%	52,5%
Drzewo klasyfikacyjne – zbiór zbalansowany z replikowanymi danymi z klasy większościowej metodą SMOTE	91,0%	50,3%
Drzewo klasyfikacyjne – zbiór zbalansowany z replikowanymi danymi z klasy większościowej (*oversampling*) z uzupełnionymi danymi metodą imputacji	92,7%	10,3%
Gradient boosting – zbiór zbalansowany z losowo usuniętymi danymi z klasy większościowej (*undersampling*)	48,9%	92,9%
Gradient boosting – zbiór zbalansowany z replikowanymi danymi z klasy większościowej metodą SMOTE oraz imputowanymi brakującymi danymi	95,5%	43,6%

Źródło: Opracowanie własne.

Algorytm *gradient boosting* uczony na zbiorze z replikowanymi danymi metodą SMOTE użyto także do klasyfikacji na pierwotnym niezbalansowanym zbiorze danych. W tym przypadku całkowita zdolność klasyfikacji mierzona wskaźnikiem *accuracy* wyniosła 95,5%, podczas gdy wskaźnik *sensitivity* wyniósł 43,6%. Rezultaty te można uznać za zbliżone do tych, które uzyskano dla prostego drzewa klasyfikacyjnego uczonego na zbiorach utworzonych w wyniku *undersamplingu* i *oversamplingu*. Podsumowanie wyników trafności predykcji algorytmów uczonych różnymi metodami na zbiorach zbalansowanych zawarto w tabeli 6.

Podsumowanie

W zagadnieniu klasyfikacji zagrożenia upadłością na zbiorze niezbalansowanym algorytmy drzewa klasyfikacyjnego i wzmocnionych drzew klasyfikacyjnych (*gradient boosting*) wykazały się zdecydowanie wyższą trafnością niż sieć neuronowa.

W przypadku uczenia algorytmu drzewa klasyfikacyjnego na zbiorze zbalansowanym utworzonym z losowo usuniętą częścią przypadków z klasy większościowej (*undersampling*) zapewnia się dużą trafność klasyfikacji upadłości przy niskiej trafności całkowitej. Z kolei ucząc algorytm na zbiorze zbalansowanym otrzymanym w wyniku replikacji przypadków z klasy mniejszościowej (*oversampling*), zapewnia się dużą trafność ogólną algorytmu przy obniżeniu trafności klasyfikacji przypadków upadłości.

Zastosowanie wyrafinowanych metod replikacji danych (SMOTE) oraz udoskonalonych algorytmów wzmocnionego uczenia maszynowego (*gradient boosting*) nie przyczyniło się do poprawy trafności klasyfikacji w stosunku do prostych metod replikacji danych oraz klasycznego drzewa klasyfikacyjnego.

W zagadnieniu klasyfikacji na zbiorach niezbalansowanych wystąpiła swego rodzaju substytucja – w wyniku działania klasyfikatora otrzymuje się albo zbiór wynikowy zawierający prawie wszystkie przypadki upadłości wraz ze znacznym nadmiarem podmiotów o dobrej kondycji finansowej, albo dobry klasyfikator ogólny, który jednak nie wykrywa dużej liczby podmiotów zagrożonych upadłością. Przy umiarkowanie dużej liczbie podmiotów można zbudować system wykrywający nadmiarowo przypadki upadłości, które następnie zostaną przeanalizowane przez specjalistę. W przypadku bardzo dużego zbioru danych nie będzie to jednak możliwe.

W odniesieniu do zagrożenia upadłością dużą rolę odgrywają czynniki jakościowe, takie jak: perspektywy rozwoju branży, produkt, struktura i rodzaj wierzycieli, kadra kierownicza przedsiębiorstwa i wiele innych [Paliński, 1999]. Prognozowanie upadłości bez uwzględnienia zmiennych jakościowych nie jest w stanie przynieść zadowalających rezultatów i to jest zapewne istotną przyczyną przeciętnych wyników klasyfikacji upadłości jedynie na podstawie wskaźników finansowych.

Literatura

Altman E.I. (1968), *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy*, "The Journal of Finance", Vol. 23, No. 4, s. 589-609.

Beaver W.H. (1966), *Financial Ratios as Predictors of Failure*, "Journal of Accounting Research", Vol. 4, s. 71-111.

Friedman J.H. (2001), *Greedy Function Approximation: A Gradient Boosting Machine*, "The Annals of Statistics", Vol. 29, No. 5, s. 1189-1232.

Galar M., Fernández A., Barrenechea E., Bustince H., Herrera F. (2012), *A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches*, "IEEE Systems, Man, and Cybernetics Society", Vol. 42(4), s. 3358-3378.

Geng R., Bose I., Chen X. (2015), *Prediction of Financial Distress: An Empirical Study of Listed Chinese Companies Using Data Mining*, „European Journal of Operational Research", Vol. 241, Iss. 1, s. 236-247.

Gissel J., Giacomino D., Akers M. (2007), *A Review of Bankruptcy Prediction Studies: 1930-Present*, "Journal of Financial Education", Vol. 33, s. 1-42.

He H., Garcia E.A. (2009), *Learning from Imbalanced Data*, "IEEE Transactions on Knowledge and Data Engineering", Vol. 21(9), s. 1263-1284.

Laitinen E.K. (1991), *Financial Ratios and Different Failure Processes*, "Journal of Business Finance & Accounting", Vol. 18, s. 649-673.

Lantz B. (2015), *Machine Learning with R – Second Edition*, Packt Publishing, Birmingham.

Li M.Y.L., Miu P. (2010), *A Hybrid Bankruptcy Prediction Model with Dynamic Loadings on Accounting-ratio-based and Market-based Information: A Binary Quantile Regression Approach*, "Journal of Empirical Finance", Vol. 17, s. 818-833.

Longford N.T. (2005), *Missing Data and Small-Area Estimation*, Springer, New York.

Maalouf M., Trafalis T. (2011), *Rare Events and Imbalanced Datasets: An Overview*, "International Journal of Data Mining, Modelling and Management", Vol. 3(4), s. 375-388.

Mahani A., Ali A. (2020), *Classification Problem in Imbalanced Datasets* [w:] A. Sadollah (ed.), *Recent Trends in Computational Intelligence*, IntechOpen, London.

Mączyńska E. (1994), *Ocena kondycji przedsiębiorstwa (Uproszczona metoda)*, „Życie Gospodarcze", nr 38, s. 42-45.

Paliński A. (1999), *Ocena procesu restrukturyzacji trudnych kredytów bankowych w latach 1992-1998 dla wybranych największych polskich banków*, „Banki i Kredyt", nr 12, s. 51-69.

Pociecha J., red. (2014), *Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej*, Fundacja Uniwersytetu Ekonomicznego, Kraków.

Serrano-Cinca S. (1996), *Self Organizing Neural Networks for Financial Diagnosis*, "Decision Support Systems", Vol. 17, Iss. 3, s. 227-238.

Wędzki D. (2005), *A Bankruptcy Logit Model for the Polish Economy*, "Argumenta Oeconomica Cracoviensia", No. 3, s. 49-70.

Wilcox J.W. (1973), *A Prediction of Business Failure Using Accounting Data*, "Journal of Accounting Research", Vol. 11, s. 163-179.

Zięba M. (2013), *Zespoły klasyfikatorów SVM dla danych niezbalansowanych*, Rozprawa doktorska zrealizowana pod kierunkiem naukowym J. Świątka na Politechnice Wrocławskiej (niepublikowana).

Zięba M., Tomczak S., Tomczak J. (2016), *Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction*, "Expert Systems with Applications", Vol. 58, s. 93-101.

Zmijewski M.E. (1984), *Methodological Issues Related to the Estimation of Financial Distress Prediction Models*, "Journal of Accounting Research", Vol. 22, s. 59-82.

[www 1] http://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data (dostęp: 10.04.2018).

[www 2] https://cran.r-project.org/package=caret (dostęp: 10.02.2019).

[www 3] https://cran.r-project.org/package=DMwR (dostęp: 10.02.2019).

[www 4] https://cran.r-project.org/package=mice (dostęp: 10.02.2019).

[www 5] https://docs.microsoft.com/en-us/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-2017 (dostęp: 10.02.2019).

BANKRUPTCY PREDICTION AS IMBALANCED CLASSIFICATION PROBLEM

Summary: Selected machine learning algorithms and data preprocessing techniques were used in the article to predict bankruptcy on an unbalanced data set containing financial ratios. The accuracy of bankruptcy forecasts on the original unbalanced data set of the prevailing share of entities still operating over the bankrupt ones was close to zero. The accuracy of bankruptcy forecasting classifiers created on balanced sets ranged from 10% to 77%, but was inversely proportional to the total accuracy of the classification, which ranged from 93% to 49%. Better classification results were achieved by the classification trees algorithms in relation to the artificial neural network. In the problem of classification in unbalanced data sets the effect of substitution occurred – or it is possible to increase the accuracy of classification of bankruptcy at the expense of redundancy of objects classified as bankrupt, or – to increase the accuracy of the overall classification of the algorithm at the expense of decreasing the classification of the bankruptcy itself.

Keywords: bankruptcy, imbalanced dataset, machine learning, classification, preprocessing.