Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 358 · 2018

Andrzej Paliński

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

METODY UCZENIA MASZYNOWEGO W PROGNOZOWANIU NIEWYPŁACALNOŚCI

Streszczenie: W artykule zastosowano wybrane algorytmy uczenia maszynowego na zbiorach danych zawierających wskaźniki finansowe w celu sprawdzenia skuteczności prognozowania upadłości. Trafność prognoz upadłości na zbiorach niezbilansowanych o przeważającym udziale firm prowadzących działalności nad upadłymi wyniosła jedynie 37%. Trafność prognozowania upadłości na zbiorach zbilansowanych wyniosła 60%. Dla porównania, uproszczone podejście eksperckie wyłoniło 76% spośród upadłych podmiotów, ale znacząco zawyżyło zbiór firm zagrożonych upadłością. Metody uczenia maszynowego okazują się skuteczne dla dużych zbiorów danych, które są zbyt liczne do analizy przez człowieka.

Słowa kluczowe: upadłość, prognozowanie, uczenie maszynowe, drzewo klasyfikacyjne.

JEL Classification: C53, C55, G33.

Wprowadzenie

Przewidywanie niewypłacalności i upadłości podmiotów gospodarczych od dziesiątek lat jest zagadnieniem badawczym istotnym z punktu widzenia inwestorów i wierzycieli firm. Prognozowanie upadłości podmiotów gospodarczych metodami statystycznymi zostało zapoczątkowane przez Beavera [1966], a popularność zyskało dzięki modelowi Altmana [1968]. W kolejnych latach powstało wiele modeli przewidywania upadłości różniących się doborem wskaźników finansowych, krajem pochodzenia podmiotów i metodami statystycznymi stosowanymi do budowy modeli. Najczęściej wykorzystywano analizę dyskryminacyjną, modele logitowe, sztuczne sieci neuronowe i drzewa klasyfikacyjne[1].

[1] Szeroki przegląd zagadnień prognozowania upadłości można znaleźć w pracy: [Pociecha (red.), 2014].

Podobnie jak w przypadku Altmana, liniowa funkcja dyskryminacyjna była wykorzystywana przykładowo w pracach Wilcoxa [1973], czy Laitinena [1991]. Równolegle prowadzono badania z użyciem uogólnionej metody regresji i modelu logitowego lub probitowego. Były to prace np.: Zmijewskiego [1984] oraz Li i Miu [2010]. W polskich realiach badania nad prognozowaniem upadłości były prowadzone m.in. przez Mączyńską [1994] i Wędzkiego [2009].

Budowa modeli predykcji upadłości napotyka dwa główne problemy. Pierwszy z nich 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 niezbilansowany zbiór danych historycznych służących budowie modeli – w próbce zdecydowanie przeważają firmy niezagrożone upadłością. Tym samym modele lepiej prognozują brak zagrożenia upadłością niż upadłość.

Wraz z gwałtownym rozwojem technologii informatycznych w latach 90. 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. Do narzędzi tych, oprócz dwóch wymienionych, można także zaliczyć: klasyfikatory Bayesa, systemy regułowe, systemy rozmyte i algorytmy ewolucyjne. W niniejszej pracy wykorzystane zostaną: drzewo klasyfikacyjne i sieci neuronowe wbudowane w oprogramowanie komercyjne.

Pierwszym celem artykułu jest ocena skuteczności typowych narzędzi uczenia maszynowego w prognozowaniu upadłości podmiotów gospodarczych. Drugim celem pracy jest porównanie trafności prognoz uzyskanych maszynowo z wiedzą ekspercką z zakresu oceny kondycji finansowej firm. Układ opracowania jest następujący: w pierwszej części przedstawiona jest charakterystyka zbiorów danych. W kolejnej zawarto wyniki badań prognozowania upadłości metodami uczenia maszynowego. W następnej dokonano próby wyodrębnienia podmiotów zagrożonych upadłością prostą metodą oceny wskaźników finansowych. Praca zakończona jest podsumowaniem zawierającym ocenę wyników.

1. Zbiór danych wykorzystany w badaniach

W pracy 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 2007-2012 dla podmiotów, które ogłosiły upadłość, oraz 2000-2012 dla podmiotów nadal prowadzących działalność. W sumie w zbiorze danych jest 700 upadłych przedsiębiorstw (2400 sprawozdań finansowych) i 10000 działających przedsiębiorstw (65000 sprawozdań). Dane zgromadzone są w pięciu następujących zbiorach danych obejmujących 5 kolejnych lat:

- Rok1 – 6756 podmiotów prowadzących działalność oraz 271 takich, które ogłosiły upadłość po 5 latach (7027 sprawozdań finansowych),
- Rok2 – 9773 podmiotów prowadzących działalność oraz 400 takich, które ogłosiły upadłość po 4 latach (10173 sprawozdania finansowe),
- Rok3 – 10008 podmiotów prowadzących działalność oraz 495 takich, które ogłosiły upadłość po 3 latach (10503 sprawozdania finansowe),
- Rok4 – 9277 podmiotów prowadzących działalność oraz 515 takich, które ogłosiły upadłość po 2 latach (9792 sprawozdania finansowe),
- Rok5 – 5500 podmiotów prowadzących działalność oraz 410 takich, które ogłosiły upadłość po 1 roku (5910 sprawozdań finansowych).

Dane załadowano do bazy danych utworzonej przy pomocy programu Microsoft SQL Server 2017 Enterprise®. W każdej z pięciu tabel reprezentujących kolejne lata umieszczono 64 wskaźniki finansowe dla każdej z firm. Opis poszczególnych wskaźników umieszczono w tabeli 1.

Tabela 1. Wskaźniki finansowe wykorzystane w analizie

Lp.	Opis	Lp.	Opis
1	*2*	*3*	*4*
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

cd. tabeli 1

1	*2*	*3*	*4*
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/aktyw 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].

2. Wyniki budowy modeli prognozowania upadłości

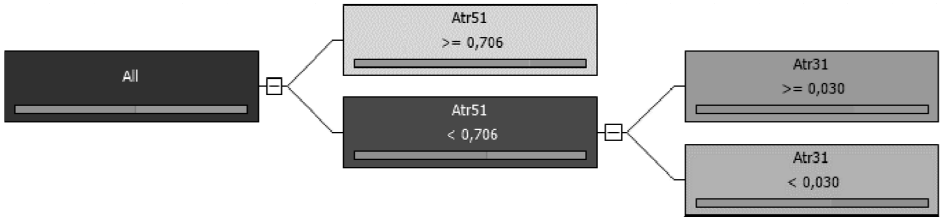
W pierwszej kolejności wykorzystano całe niezbilansowane zbiory danych, gdzie znacząco przeważały firmy, które nie ogłosiły upadłości – ponad 95% w każdym roku analizy. Zastosowano 2 narzędzia uczenia maszynowego: drzewo klasyfikacyjne i sztuczne sieci neuronowe, zaimplementowane w oprogramowaniu SQL Server. Naiwny klasyfikator Bayesa został pominięty, gdyż nie wspierał ciągłych wartości atrybutów (wskaźników finansowych).

Trafność prognoz dla poszczególnych lat w przypadku drzewa klasyfikacyjnego wahała się od 97% dla 1. roku do 94% dla 5. roku, czyli jednego roku przed ogłoszeniem upadłości. Pozornie bardzo wysoka skuteczność modeli wynikała z dużej zdolności modeli do przewidywania kontynuacji działalności. Niestety trafność prognozowania upadłości wynosiła od 24% dla 3. roku do 49% dla 1. roku (średnio 37% dla wszystkich lat), co wynikało z przeważającego udziału w próbie spółek od dobrej kondycji finansowej. O ile w przypadku sieci neuronowych trafność ogólna prognoz była nieznacznie niższa (od 96% do 92%), niż w przypadku drzew klasyfikacyjnych, o tyle sieci neuronowe prawie nie prognozowały przypadków upadłości.

Ze względu na niewielką skuteczność prognozowania upadłości przez modele uczone na pełnych zbiorach danych, co powinno być głównym zastosowaniem modeli tego typu, utworzono zbiory zbilansowane, które składają się ze wszystkich przypadków upadłości oraz równolicznych, losowo wybranych próbek przedsiębiorstw, które nie ogłosiły upadłości. Utworzono tym samym 5 zbiorów o licznościach wahających się od niecałych 600 do ponad 1000 przedsiębiorstw. Ponadto usunięto z modeli wskaźniki: Atr27 i Atr37 ze względu na zbyt duży udział brakujących danych oraz zrezygnowano z użycia sieci neuronowych.

Na rys. 1 przedstawiono przykładową strukturę drzewa klasyfikacyjnego dla zbioru Rok2. Podsumowanie wyników prognozowania przez modele na próbach testowych utworzonych automatycznie przez program z 30% danych z każdego ze 5. zbiorów umieszczono w tabeli 2.

Dane na przekątnych od 0-0 do 1-1 w tabeli 2 informują o liczbie przypadków poprawnej predykcji. Trafność prognoz wszystkich modeli nie jest wysoka i waha się od 56% do 71%. Jest też wyraźnie niższa w porównaniu z trafnością prognoz dla całych niezbilansowanych zbiorów danych. Z kolei trafność prognoz upadłości jest wyższa niż dla pierwotnych zbiorów danych, ale nadal niezbyt wysoka: Rok1 – 56%, Rok2 – 69%, Rok3 – 51%, Rok4 – 71% i Rok5 – 53%. Jedną z przyczyn takiego stanu rzeczy mogą być brakujące dane dotyczące niektórych wskaźników w zbiorze uczącym.



Rys. 1. Drzewo klasyfikacyjne dla zbioru Rok2

Modele utworzone dla danych zbilansowanych zostały następnie użyte do prognozowania upadłości dla pierwotnych niezbilansowanych dużych zbiorów danych. Ogólna trafność modeli jest także niezbyt wysoka (od 41% do 81%) i przeważnie odwrotnie proporcjonalna do trafności modeli w odniesieniu do prognozowania upadłości wynoszącej od 42% (1 rok przed upadłością) do 78% (2 lata przed upadłością). Średnia trafność przewidywania upadłości wyniosła 60%.

Tabela 2. Podsumowanie wyników prognozowania upadłości dla zbiorów zbilansowanych (zbiory testowe stanowią 30% danych ze zbiorów zbilansowanych, 1 oznacza upadłość)

	Rzeczywiste	
Prognozowane	0	1
Rok1		
0	45	37
1	36	47
Rok2		
0	61	40
1	49	90
Rok3		
0	133	73
1	15	77
Rok4		
0	76	44
1	84	106
Rok5		
0	102	64
1	10	73

Źródło: Opracowanie własne.

3. Uproszczone podejście eksperckie w prognozowaniu upadłości

Ze względu na to, że modele uzyskane w procesie uczenia maszynowego okazały się umiarkowanie skuteczne w prognozowaniu upadłości, spróbowano wyłonić z posiadanych zbiorów danych podmioty potencjalnie zagrożone upadłością na podstawie kilku podstawowych wskaźników finansowych. Klasyczna

analiza finansowa posługuje się pięcioma podstawowymi grupami wskaźników (płynności, rentowności, zadłużenia, sprawności działania i rynku kapitałowego), jednakże w odniesieniu do zagrożenia upadłością największe znaczenie ma zwykle relacja długu do kapitału – zbyt duży udział długu w finansowaniu majątku bardzo często prowadzi do upadłości, ponadto bardzo niska lub ujemna wartość kapitału obrotowego netto, straty finansowe powtarzające się w kolejnych latach działalności i ewentualnie długi cykl spłaty zobowiązań są także symptomami zagrożenia upadłością [Paliński, 1998]. W modelach predykcji upadłości pojawiają się najczęściej wskaźniki niosące podobną treść informacyjną: zysk netto/aktywa, wskaźnik bieżącej płynności, kapitał obrotowy/aktywa i zysk zatrzymany/aktywa [Gissel, Giacomino, Akers, 2007].

Imitując zachowanie analityka finansowego, który na podstawie niekorzystnych wartości pojedynczych wskaźników rozpoczęłby głębszą analizę kondycji finansowej podmiotu, wybrano arbitralnie 5 wskaźników o następujących wartościach: Atr1 < 0, Atr6 < 0, Atr2 > 0,7, Atr4 < 0,9 i Atr32 > 180. W wyniku przeszukania zbilansowanych zbiorów danych wyodrębniono firmy, które przekroczyły chociaż jedną wartość dowolnego z wybranych wskaźników. Rezultaty okazały się interesujące, gdyż zastosowanie bardzo prostego podejścia pozwoliło na znalezienie od 67% (dla 5 lat przed upadłością) do 84% (na rok przed upadłością) firm spośród tych, które ogłosiły upadłość. Średnia trafność takiej naiwnej prognozy upadłości na zbiorze zbilansowanym wyniosła 76%, co jest wartością lepszą niż wyniki algorytmów wykorzystujących metody uczenia maszynowego.

Słabą stroną takiego prostego podejścia jest jednak fakt, że w wynikach wyszukiwania otrzymuje się zbyt duże zbiory firm, potencjalnie zagrożonych upadłością. Około 40% firm znalezionych jako zagrożone upadłością nie było podmiotami, które ogłosiły upadłość. Zastosowanie tego samego podejścia w stosunku do liczniejszych, niezbilansowanych zbiorów danych dało jeszcze gorszy wynik – blisko 50% znalezionych firm z niekorzystną wartością jednego ze wskaźników finansowych nie upadło. Niemniej jednak, w odniesieniu do niewielkiej liczby firm, analityk finansowy po analizie szerszego zestawu wskaźników nie miałby trudności ze zidentyfikowaniem podmiotów wysoce zagrożonych upadłością ze wstępnie wyłonionego zbioru danych.

Podsumowanie

Przeprowadzone badania wykazały, że trafność prognoz upadłości metodami ucznia maszynowego wynosi średnio 60%. Najlepsze wyniki uzyskano przy użyciu drzewa klasyfikacyjnego. Wyniki przewidywania upadłości metodami

uczenia maszynowego nie są w pełni zadowalające. Za przyczynę umiarkowanej skuteczności tych metod można uznać:

- niezbyt dużą liczbę przypadków upadłości w zbiorach danych, w których z naturalnych przyczyn przeważają znacząco podmioty kontynuujące działalność,
- brak w posiadanym zbiorze, poza wskaźnikami finansowymi, danych makroekonomicznych i branżowych oraz danych jakościowych dotyczących kadry zarządzającej, sytuacji organizacyjnej w firmie, współpracy z wierzycielami i innych,
- trudność w prognozowaniu upadłości wynikającą z przyczyn pozafinansowych – decyzji wierzycieli dotyczących ugody lub układu z dłużnikiem, lub działań zarządu lub właścicieli firm świadomie doprowadzających do upadłości.

Na podstawie przeprowadzonej prostej próby wyłonienia podmiotów zagrożonych upadłością metodą ekspercką, która okazała się umiarkowanie skuteczna, można wysunąć ogólny wniosek: dla niewielkiej liczby przypadków ekspert może lepiej przewidywać upadłość. Dla dużych zbiorów danych automatyczne algorytmy będą skuteczniejsze, zwłaszcza jeżeli zbiór zmiennych objaśniających uzupełni się o dane jakościowe i makroekonomiczne.

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.

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.

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

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.

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

Paliński A. (1998), *Efektywność procesu restrukturyzacji trudnych kredytów bankowych*, praca doktorska, Akademia Ekonomiczna im. K. Adamieckiego w Katowicach.

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

Serrano-Cinca S. (1996), *Selforganizing 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., 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.

[www1] http://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data (dostęp: 10.04.2018).

MACHINE LEARNING METHODS IN BANKRUPTCY PREDICTION

Summary: The article uses selected machine learning algorithms on datasets containing financial ratios to check the effectiveness of bankruptcy prediction. The accuracy of bankruptcy forecasts for unbalanced dataset with the prevalence of companies still operating over bankrupts was only 37%. The accuracy of bankruptcy forecasting on a balanced dataset was 60%. The simplified expert approach selected 76% of bankrupt entities, but significantly overstated the set of companies exposed on bankruptcy. Machine learning methods are effective for large data sets that are too numerous for human analysis.

Keywords: bankruptcy, forecasting, machine learning, decision tree.