Uniwersytet Warszawski

Wydział Nauk Ekonomicznych

Krzysztof Szafrański

Nr albumu: 386 118

Porównanie metod stosowanych w predykcji bankructwa

Warszawa, czerwiec 2022

Wstęp

Od ponad 50 lat zagadnienie predykcji bankructwa cieszy się rosnącym zainteresowaniem badaczy, którzy do analizy tego tematu wykorzystują wiele różnych modeli. Począwszy od pracy Altmana (1968) najpopularniejsza była analiza dyskryminacyjna. Z kolei w latach 80 powszechnie wykorzystywano regresję logistyczną, a później do głosu doszły modele sieci neuronowych i inne techniki uczenia maszynowego (Bellovary i Giacomino, 2007).

Celem niniejszej pracy jest porównanie wyników modelu sieci neuronowych z bardziej klasycznymi metodami - regresją logistyczną i analizą dyskryminacyjną. Analiza zostanie przeprowadzona na zbiorze danych zawierającym informacje o wskaźnikach finansowych dla polskich przedsiębiorstw, z których część zbankrutowała. Na podstawie literatury należy się spodziewać, że model sieci neuronowych powinien prowadzić do najlepszych predykcji.

W pierwszej części pracy dokonano krótkiego przeglądu literatury. Następnie opisano wykorzystane dane, a na końcu przedstawiono wyniki przeprowadzonej analizy.

Przegląd literatury

Jedną z najbardziej znanych prac z obszaru predykcji bankructwa jest artykuł Edwarda Altmana (1968). Autor wykorzystuje w nim analizę dyskryminacyjną, która służy do klasyfikacji obserwacji do określonych grup, w oparciu o indywidualne charakterystyki. Jego model w 95% poprawnie grupował przedsiębiorstwa na te, które zbankrutowały i te, które wciąż funkcjonowały. Jednak trafność predykcji była wysoka tylko na rok i na 2 lata przed bankructwem, natomiast zdecydowanie malała dla wcześniejszych lat.

Po początkowej popularności analizy dyskryminacyjnej, autorzy zaczęli częściej wykorzystywać regresję logistyczną. Shi i Li (2019) przeprowadzili przegląd literatury, z którego wynikało, że model logit był używany w prawie 40% z przeanalizowanych przez nich artykułów dotyczących predykcji bankructwa.

Back i in. (1996) sprawdzali jak różne modele wpływają na selekcję zmiennych niezależnych oraz prowadzą do innej dokładności predykcyjnej. Każda z porównywanych metod różniła się pod względem wybranych wskaźników finansowych. Ponadto metoda sieci neuronowych prowadziła do wyraźnie lepszych predykcji niż analiza dyskryminacyjna i logistyczna. W przypadku tej metody ponad 97% obserwacji było klasyfikowanych poprawnie.

Mimo dużej popularności sieci neuronowych niektóre prace wskazywały, że inne metody uczenia maszynowego mogą się spisywać jeszcze lepiej. Na przykład badanie Barbozy, Kimury i Altmana (2017) wskazywało na wysoką skuteczność trzech metod - bagging, boosting i lasów losowych - które prowadziły do o około 10% dokładniejszych predykcji w porównaniu do przedstawionych wcześniej modeli.

Opis danych

Dane wykorzystane w analizie zaczerpnięto ze strony: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data. Zawierają one informacje o wskaźnikach finansowych dla 5910 polskich przedsiębiorstw na rok przed ewaluacją. Wśród analizowanych firm 410 ogłosiło upadłość. Baza zawiera aż 64 wskaźniki finansowe, jednak zdecydowano się wykorzystać tylko 15 z nich, które były najistotniejsze w kontekście predykcji bankructwa zdaniem autorów bazy (Zięba, Tomczak i Tomczak, 2016). Poniżej przedstawiono ich opis:

- X5 - [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
- X9 - sales / total assets
- X13 - (gross profit + depreciation) / sales
- X15 - (total liabilities * 365) / (gross profit + depreciation)
- X22 - profit on operating activities / total assets
- X25 - (equity - share capital) / total assets
- X27 - profit on operating activities / financial expenses
- X31 - (gross profit + interest) / sales
- X36 - total sales / total assets
- X40 - (current assets - inventory - receivables) / short-term liabilities
- X42 - profit on operating activities / sales
- X48 - EBITDA (profit on operating activities - depreciation) / total assets
- X52 - (short-term liabilities * 365) / cost of products sold)
- X58 - total costs /total sales

Baza danych jest zdecydowanie niezbilansowana - przedsiębiorstwa, które zbankrutowały stanowią mniej niż 7% wszystkich obserwacji. Problem ten może negatywnie wpływać na dokładność predykcji modeli, dlatego zdecydowano się przeprowadzić oversampling, tak aby wyrównać liczebność dwóch klas przedsiębiorstw. Veganzones i Severin (2018) porównywali skuteczność różnych metod oversamplingu i wskazywali, że to algorytm SMOTE prowadzi do największej poprawy dokładności predykcji. Dokonano również standaryzacji zmiennych dla łatwiejszej interpretacji wyników. Zbiór danych

podzielono na część treningową (70%), na której będą estymowane modele oraz na część testową (30%), na której sprawdzona zostanie dokładność predykcji modeli.

Wyniki

Poniżej przedstawiono wyniki liniowej analizy dyskryminacyjnej (Tabela 1) oraz regresji logistycznej (Tabela 2). W analizie dyskryminacyjnej największy wpływ na różnicowanie przedsiębiorstw mają zmienne V9, V22, V36 i V48. Z kolei w regresji logistycznej istotne są zmienne V9, V22, V25, V36, V42, V48, V52 i V58.

Tabela 1. Oszacowania zmiennych w funkcji dyskryminacyjnej

```
Coefficients of linear discriminants:
             LD1
V5  -0.065775143
V9   1.321251676
V13 -0.129987113
V15  0.047330964
V22 -1.483283392
V25 -0.611338989
V27 -0.221411976
V31  0.037471965
V36 -1.157200674
V40  0.004321534
V42  0.154035986
V48  1.134399137
V52  0.106098392
V58  0.321838930
```

Źródło: opracowanie własne.

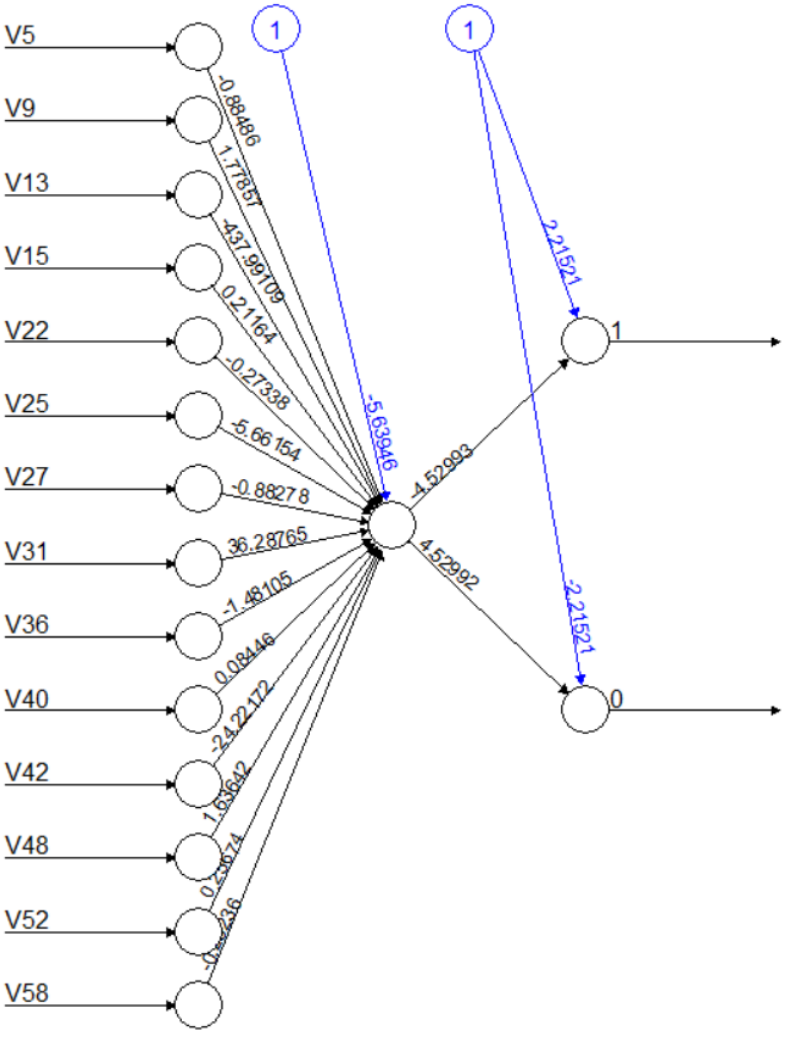
Tabela 2. Oszacowania parametrów w regresji logistycznej

```
Coefficients:
             Estimate Std. Error z value        Pr(>|z|)
(Intercept) -0.877678   0.125222  -7.009 0.0000000000024 ***
V5           0.006947   0.030278   0.229        0.818532
V9           1.424205   0.113678  12.528         < 2e-16 ***
V13         -9.711937   9.261471  -1.049        0.294344
V15          0.153575   0.095478   1.608        0.107727
V22         -4.526885   0.308192 -14.689         < 2e-16 ***
V25         -3.273920   0.172158 -19.017         < 2e-16 ***
V27         -0.615258   0.317013  -1.941        0.052283 .
V31          1.630221   1.566067   1.041        0.297892
V36         -1.109430   0.118396  -9.371         < 2e-16 ***
V40          0.038325   0.026870   1.426        0.153785
V42          0.610535   0.170294   3.585        0.000337 ***
V48          3.942815   0.336301  11.724         < 2e-16 ***
V52          0.182296   0.036722   4.964 0.0000006898111 ***
V58          0.474027   0.096650   4.905 0.0000009362883 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Źródło: opracowanie własne.

Poniżej przedstawiono graficzny schemat wykonanego modelu sieci neuronowych (Rys. 1). Zmienne są wprowadzane w tzw. warstwie wejściowej, a następnie są przekazywane do neuronu w warstwie ukrytej z różnymi wagami, gdzie wypracowuje się rozwiązania. W uproszczeniu wagi można interpretować podobnie jak oszacowania w regresji. Co ciekawe, największe wagi przypisano zmiennym V13, V31 i V42, które nie były najistotniejsze w poprzednich modelach. Z kolei niebieskie strzałki reprezentują tzw. *bias*, który ma spełniać podobną funkcję jak stała w normalnej regresji. Warstwa ukryta może też zawierać więcej niż jeden neuron, jednak dodanie kolejnych neuronów nie poprawiało dokładności predykcji.

Rys. 1. Graficzny schemat modelu sieci neuronowych



Źródło: opracowanie własne.

Poniżej przedstawiono macierze błędów klasyfikacji dla trzech oszacowanych modeli (Tabela 3). Najlepsze predykcje uzyskuje model logitowy oraz model sieci neuronowych. Pierwszy poprawnie klasyfikuje 78,15% obserwacji, a drugi - 77,57%. Nieco gorszą dokładność ma analiza dyskryminacyjna, która poprawnie zaklasyfikowała 73,93% obserwacji.

Tabela 3. Macierze błędów klasyfikacji dla trzech oszacowanych modeli

LAD

Prediction	Reference 0	1
0	1206	362
1	447	1088

LOGIT

Prediction	Reference 0	1
0	1238	330
1	348	1187

SIECI NEURONOWE

Prediction	Reference 0	1
0	1227	341
1	355	1180

Źródło: opracowanie własne.

Podsumowanie

Celem niniejszej pracy było porównanie trzech popularnych metod wykorzystywanych w badanich nad predykcją bankructwa - analizy dyskryminacyjnej, regresji logistycznej oraz metody sieci neuronowych. Wyniki analizy wskazują, że modele różnią się pod względem zmiennych, które mają największy wpływ na klasyfikację, co jest spójne z wynikami jednego z badań (Back i in., 1996). Jednak na podstawie uzyskanych wyników nie można powiedzieć, że model sieci neuronowych prowadzi do najlepszych predykcji, ponieważ nieco lepszą dokładność uzyskano wykorzystując regresję logistyczną.

\

Literatura

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589-609.

Back, B., Laitinen, T., Sere, K., & van Wezel, M. (1996). Choosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms. *Turku Centre for Computer Science Technical Report*, *40*(2), 1-18.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405-417.

Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial education*, 1-42.

Shi, Y., & Li, X. (2019). An overview of bankruptcy prediction models for corporate firms: A systematic literature review. *Intangible Capital*, *15*(2), 114-127.

Veganzones, D., & Séverin, E. (2018). An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems*, *112*, 111-124.

Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert systems with applications*, *58*, 93-101.