UNIWERSYTET
EKONOMICZNY
W POZNANIU

Bartłomiej Lach

Predicting bankruptcy of enterprises in Poland using the methods of combining and selecting classifiers

THE SUMMARY OF DOCTORAL DISSERTATION

Supervisor:

Dr hab. Barbara Będowska-Sójka,
Prof. UEP

Auxiliary supervisor:

Dr Sergiusz Herman

Poznań 2020

With dynamic changes in the economic and legal environment, taking optimal business decisions and retaining the ability to quickly and adequately react to threats to business activity become increasingly difficult. For this reason, among business managers there is a growing need for employing tools which would monitor the financial condition of enterprises as well as signal in advance a potential risk of bankruptcy. Corporate bankruptcy is a common economic phenomenon which is impossible to be eliminated. Early detection of the risk of bankruptcy is aimed at protecting the entities concerned against its negative effects. Bankruptcy prediction gives managers in enterprises time to react and take suitable decisions that would allow avoiding or minimising negative consequences of this form of closing business activity. Prediction of bankruptcy does not have to relate only to monitoring of the financial condition of a business. In today's economic reality, security of a business does not depend only on a company's own financial standing but also on business partners' stability and reliability. Therefore, detection and monitoring of the risk of corporate bankruptcy constitute an important area of interest among business practitioners.

Corporate bankruptcy prediction is also a subject of many academic studies. It is frequently presented as posing a problem of classification which consists in determining the membership of an object, i.e an enterprise, in one of the two possible populations: *entities with good financial standing or entities exposed to the risk of bankruptcy*.

The literature presents some attempts to create corporate bankruptcy prediction tools with the use of various classification methods. One of the first and best known bankruptcy prediction models was the Z-score model proposed by Altman [1968]. The model was developed based on data from the American market and its construction uses linear discriminant analysis. One of the first bankruptcy prediction models in Poland was proposed by Mączyńska [1994]. Later studies employed diverse classification methods for multidimensional objects, in an effort to explain the causes of corporate bankruptcy and obtain possibly the most effective prediction tools for this phenomenon.

As found by the review of research studies dedicated to corporate bankruptcy prediction in Poland the most popular classification methods used in these studies include: linear discriminant analysis, logistic regression, classification trees and artificial neural networks [Hadasik 1998, Gruszczyński 2003, Hamrol et.al. 2004, Wędzki 2005, Korol and Prusak 2005, Hołda 2006, Hołda and Pociecha 2009, Korol 2010, Pociecha et al. 2014, Paliński 2018]. The above-mentioned methods are often used to predict corporate bankruptcy

also for other markets, as mentioned by the authors who reviewed research studies in this field [Bellovary et al. 2007, Prusak 2019].

It is difficult to determine which of the classification methods applied in the literature of the subject is clearly better. One of the ways of improving the quality of object classification is applying ensemble methods. The use of ensemble classifiers consists in taking final classification decisions about membership of an object by indicating many individual classifiers.

The basic division of ensemble classifiers is connected with the number of classification methods applied to create them. In a situation when a set of classifiers is built on the basis of one classification method (e.g. a classification tree), it is called a homogeneous ensemble. A classifier based on the use of one classification method is therefore strengthened by many classifiers of the same type [Krzyśko et al, 2008].

In another approach, a set of classifiers is formed by means of a bigger number of various classification methods. In this case, an ensemble classifier is called heterogeneous or a multiple classifier system [Ćwiklińska-Jurkowska, 2013]. This group of ensemble classifiers is built by the methods of combining or selecting individual classifiers, which constitutes the Author's main scope of interest in this dissertation. Classifier combination and selection methods were developed as part pattern recognition studies, which are a component of machine learning. The attempt of simultaneous use of individual classifiers is a natural development direction of classification methods. Kuncheva [2004] emphasises growing interest in the ensemble approach for classifying multidimensional objects and the need to verify the effectiveness of this approach through its practical use in empirical research.

The conducted research, whose development and results are described in this dissertation, is an effort to compare various classification methods for business bankruptcy prediction in Poland. It is intended to answer the questions if and when it is advisable to apply the ensemble approach based on classifier combination and selection to predict corporate bankruptcy. In the course of the research, results of numerous ensemble classifiers based on various methods of aggregation were compared. The empirical research allowed assessing and comparing the classification results, both for individual and ensemble classifiers, which offered a chance to obtain a fuller picture of the effectiveness of particular methods.

In the foreign literature an overwhelming majority of studies devoted to the use of ensemble methods for corporate bankruptcy prediction present an attempt at enhancing classifiers of the same type. The most common methods comprise: bagging, boosting and

random subspaces. The effect of applying these methods is homogeneous ensemble classifiers [Fedorova et al. 2013, Kim and Kang 2012]. Examples of using heterogeneous ensembles to predict corporate bankruptcy make up a significantly smaller proportion of all the research in this field [Xiao et al. 2012, Geng et al. 2015].

In the Polish literature there are few examples of using ensemble methods to predict corporate bankruptcy. In recent years several scientific papers have been published which tried to use homogeneous sets of classifiers for corporate bankruptcy prediction where the individual classifier was a classification tree [Pawełek and Grochowina 2017a, 2017b, Gąska 2016]. The use of heterogeneous ensembles of classifiers to predict corporate bankruptcy in Poland was described in an article written by the author of this dissertation [Lach, 2017]. The subject of this doctoral dissertation and the extensive research on the effectiveness of classifier combination and selection methods fill a gap in the studies on corporate bankruptcy prediction in Poland and contribute to the research in this field.

Considering the aforementioned remarks, the Author formulated the main goal of the dissertation and four specific goals. **The main goal of the dissertation** is to examine the effectiveness of combining and selecting classifiers in business bankruptcy prediction in Poland.

The conducted empirical research allowed verification of **the main hypothesis of the dissertation**, which reads as follows:

H1. Using classifier combination and selection methods improves the accuracy of object classification in business bankruptcy prediction.

In addition to the main hypothesis, the dissertation verified two **auxiliary hypotheses** :

H2. Ensemble classifiers based on combination methods are marked with a better ability to correctly predict business bankruptcy than ensemble classifiers based on selection methods.

H3. The effectiveness of classifier combination and selection methods depends on the number of exogenous variables considered during the formation of classifiers.

The doctoral dissertation consists of five chapters. The first three chapters are theoretical and the remaining chapters present the results of the conducted empirical research. The theoretical part defines the notion of corporate bankruptcy, presents its causes and the statistical data which describe this phenomenon in Poland, highlighting the changes which took place in the bankruptcy law in 2016. The subsequent part shows the problem of

classifying multidimensional objects and classification methods applied in corporate bankruptcy prediction. Moreover, it draws attention to the evaluation of classification quality of the models and reviews the studies on corporate bankruptcy prediction in the Polish and foreign literature. The last chapter of the theoretical part presents essential characteristics of the working of classifier combination and selection methods. In particular, this fragment discusses the methods used for enhancing classifiers of the same type (homogeneous ensembles) as well as combined aggregation methods for classifiers of various types (heterogeneous ensembles). The last fragment of this chapter also presents an overview of the Polish and foreign research on corporate bankruptcy prediction which adopted the ensemble approach.

In the context of the research goals and verification of the hypotheses, a very important role is played by the empirical part of the dissertation, included in the last two chapters. The empirical research used information on financial condition of 180 joint stock companies operating in Poland. These companies conducted business activity in one of the three following sectors: construction, wholesale and retail trade, and the processing industry. On the basis of the data from annual financial statements, 19 financial indicators were selected which would describe possibly in the most comprehensive way financial situation of each enterprise examined in the study.

The research applied four individual classifiers. These were: linear discriminant analysis, logistic regression, random forests and artificial neural networks. Based on them, nine ensemble heterogeneous classifiers were built (seven by combination methods and two by selection methods). The chosen combined classifiers were the following: a majority vote classifier, a summation classifier (including a weighted sum classifier), a multiplicative classifier and three classifiers based on positional statistics: the minimum classifier, the median classifier, and the maximum classifier). During the formation of combined ensemble classifiers the study used the architecture of common features of individual classifiers [Krzyśko et al., 2008]. To build selection classifiers the study adopted the approach proposed by Kunchevą [2000] – the Clustering and Selection method (CS). The CS method is based on determining static regions (observation clusters) within with the objects are similar in terms of certain defining features. For this purpose the method of iterative optimisation to divide sets of objects was used, i.e. k-means clustering. Each of these regions is assigned an individual classifier which in a given region obtained the highest power of correct classification at the stage of learning.

Most of the analyses performed as part of the conducted research were made with the use of a self-developed software application created in the R environment with the graphical user interface. The application enabled improving work on the effectiveness of classifier combination and selection methods. It can be used for any problem of binary classification of objects. The available software and popular statistical packages do not offer tools that could directly use classifier combination and selection methods, which was an important reason for undertaking work on the Author's IT solution.

To meet the goals of the dissertation, the fourth chapter of was dedicated to comparing the accuracy of predictions obtained for all individual and ensemble classifiers in the training and testing sets. The highest level of the global accuracy rate in the training set, indicating a proportion of correctly classified objects, was given to two ensemble selection classifiers. In the case of testing set objects, the highest prediction validity was also shown for two ensemble classifiers: the summation and weighted sum classifiers. Out of individual classifiers, the highest forecast accuracy was proved for the artificial neural network and random forest.

The subsequent fragment of the fourth chapter analyses the effect of training and testing set sampling on the validity of classification. Multiple draws as well as the formation of classifiers enabled comparing the forecast accuracy. Among the training set objects the highest forecast accuracy was again scored by the ensemble selection classifier. In the case of the testing set objects, the carried out analysis showed that for classifiers built on the basis of a selected set of financial indicators, the use of classifier combination and selection did not improve forecast accuracy. The highest global accuracy rate out of all the thirteen classifiers was obtained for the random forest, which in the study belonged to the group of individual classifiers.

Another stage of the study evaluates the prediction stability of particular classifiers and their ability to generalise results. For this purpose, standard deviations of the obtained global accuracy rates for 100 draws of the training and testing sets were determined. The ability to generalise classifier results was assessed on the basis of a proposed indicator measuring a mean deviation of global accuracy rates obtained for the testing set objects from the training set objects. The conducted analysis indicated that the highest prediction ability in the training and testing sets was obtained by the random forest classifier.

To find the answer to the question why only in some cases of the draws for training and testing sets the use of ensemble classifiers improved prediction accuracy, the last

fragment of the fourth chapter was dedicated to examining the effect of a degree of compliance of individual classifiers on the effectiveness of combination and selection methods. The results of the statistical test showed that a higher proportion of ambiguously classified objects by individual classifiers increases the chance of improved forecast accuracy through the use of combination or selection methods. The study assumed that an object is classified ambiguously if at least one of the individual classifiers indicates its membership in a different population than the other classifiers.

The last, fifth chapter of the dissertation examined the influence of diagnostic variable selection on the effectiveness of classifier combination and selection methods in corporate bankruptcy prediction. The analysis of the relationships between the number of exogenous variables taken into consideration while building classifiers and the effectiveness of combination and selection methods constituted one of the specific goal of the dissertation. As part of this stage an extensive simulation test was performed, encompassing almost 550 thousand individual and ensemble classifiers. The formation of classifiers adopted the random subspaces method consisting in multiple draws of various sets of exogenous variables for each possible number of variables from 2 to 18.

The results for the training set objects displayed the expected dependence in which with an increase in the number of variables, individual and ensemble classifiers obtained increasingly high average global accuracy rates. For each possible number of variables the highest forecast accuracy was presented by ensemble selection classifiers, which proved the results obtained earlier. The results for the testing set objects were equally interesting. The average global accuracy rate for all classifiers rose until the number of variables exceeded 6-8. When the number of variables still rose, most individual and ensemble classifiers were marked with an increasingly low ability to classify objects correctly. The only exception was the random forest classifier whose average global accuracy rate increased also in the case of higher numbers of variables. The random forest was definitely the most effective in corporate bankruptcy prediction when the number of variables exceeded 9.

Comparing the validity of predictions obtained for the examined ensemble classifiers in the testing set allowed **positive verification of the hypothesis which says that ensemble classifiers based on combination methods are marked with a better ability to correctly predict corporate bankruptcy than ensemble classifiers based on selection methods (H2)**. For each possible number of variables from 2 to 18, out of all ensemble classifiers, the best results were obtained by one of the combination classifiers. The highest scores of the

averaged global accuracy rate were obtained by the median, summation and weighted sum classifiers. Comparative analysis also showed that applying ensemble classifiers based on probability vectors *a posteriori* (e.g. summation or median classifier) represents a more effective method of aggregating classification results than the majority vote method, which relies only on indications of object membership in a population.

In the simulation test, ensemble combination classifiers achieved the highest average global accuracy rates out of all the analysed classifiers when they were formed of a smaller number of exogenous variables. The results obtained for the analysed method allow **positive verification of the hypothesis which assumed that the effectiveness of classifier combination and selection methods depends on the number of exogenous variables considered during classifier formation (H3)**. It is also confirmed by the frequency analysis for classifiers with the highest global accuracy rate in the testing set. In the case of classifiers with a small number of variables, for more than a half of the draws in exogenous variable sets, the most effective was one of the nine ensemble classifiers. The summation and median ensemble classifiers also obtained the highest global accuracy rates for the number of variables amounting to 6 and from 8 to 10. Therefore, it needs to be stated that **the main hypothesis H1, which assumed that classifier combination and selection methods improves object forecast accuracy in corporate bankruptcy prediction, was confirmed only for classifiers with a small number of exogenous variables.** Among the classifiers with the number of variables exceeding 8, distinctly the most effective classifier was the random forest.

The final part of the fifth chapter examines to to what was the impact of particular financial indicators affect correct classification of enterprises into the population of companies with good financial standing or the one exposed to bankruptcy. For this purpose the study proposed to create a measure of discriminant power, relying on the results of the simulation test. The measure determined for a given financial indicator corresponded to the average global accuracy rate for all classifiers with two exogenous variables, one of them being that indicator. The highest discriminant power was obtained by the following indicators: equity capital / total liabilities, total liabilities / total assets, equity capital / total assets and the current ratio.

Literature (selected)

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

Bellovary, J. L., Giacomino, D. E., Akers, M. D., 2007, A Review of Bankruptcy Prediction Studies: 1930 to Present, Journal of Financial Education, Vol. 33.

Ćwiklińska-Jurkowska, M., 2013, Performance of Resampling Methods Based on Decision Trees, Parametric and Nonparametric Bayesian Classifiers for Three Medical Datasets, Studies in Logic, Grammar and Rhetoric 35 (1), pp. 71-86.

Fedorova, E., Gilenko, E., Dovzhenko, S., 2013, Bankruptcy prediction for Russian companies: Application of combined classifiers, "Expert Systems with Applications", Vol. 40, No. 18.

Gąska, D., 2016, Przewidywanie bankructwa przedsiębiorstw za pomocą metod uczenia, praca doktorska, promotor pracy: dr hab. Joanna Dębicka, Uniwersytet Ekonomiczny we Wrocławiu.

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, pp. 236–247.

Gruszczyński, M., 2003 Modele mikroekonometrii w analizie i prognozowaniu zagrożenia finansowego przedsiębiorstw, Zeszyty PAN, nr 34, Warszawa.

Hadasik, D., 1998, Upadłość przedsiębiorstw w Polsce i metody jej prognozowania, Zeszyty naukowe – seria II, Prace habilitacyjne, Zeszyt 153, Akademia Ekonomiczna w Poznaniu, Poznań.

Hamrol, M., Czajka B., Piechocki M., 2004 Upadłość przedsiębiorstwa – model analizy dyskryminacyjnej, „Przegląd Organizacji" nr 6.

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

Hołda, A., Pociecha, J., 2009, Probabilistyczne metody badania sprawozdań finansowych, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Kim, M.J., Kang, D. K., 2012, Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction, Expert Syst. Appl. 39, pp. 9308–9314.

Korol, T., 2010, Systemy ostrzegania przedsiębiorstw przed ryzykiem upadłości, Wolters Kluwer, Warszawa.

Korol, T., Prusak, B., 2005, Upadłość przedsiębiorstw a wykorzystanie sztucznej inteligencji, CeDeWu, Warszawa.

Krzyśko, M., Wołyński, W., Górecki, T., Skorzybut, M., 2008, Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości, Wydawnictwo Naukowo-Techniczne w Warszawie, Warszawa.

Kuncheva, L., 2000, Cluster-and-Selection Method for Classifier Combination, 4th International Conferenceon Knowledge-Based Intelligent Engineering Systems & Allied Technologies, Brighton, UK.

Kuncheva, L., 2004, Combining Pattern Classifiers: Methods and Algorithms, Wisley.

Lach, B., 2017, Metody łączenia i selekcji klasyfikatorów w prognozowaniu upadłości przedsiębiorstw, Przegląd Statystyczny, R. LXIV – ZESZYT 2, pp. 177-191.

Mączyńska, E., 1994, Ocena kondycji przedsiębiorstwa (uproszczone metody), Życie Gospodarcze, 38, pp. 42-45.

Paliński, A., 2018, Metody uczenia maszynowego w prognozowaniu niewypłacalności, Studia Ekonomiczne, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, Nr 358, pp. 173-181.

Pawełek, B., Grochowina, D., 2017a, Forecasting the Bankruptcy of Companies: the Study on the Usefulness of the Random Subspaces and Random Forests Methods, The 11th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena, Cracow University of Economics.

Pawełek, B., Grochowina, D., 2017b, Podejście wielomodelowe w prognozowaniu zagrożenia przedsiębiorstw upadłością w Polsce. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, pp. 171-179.

Pociecha, J., Pawełek, B., Baryła, M., Augustyn, S., 2014, Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej, Fundacja Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Prusak., B., 2019, Corporate bankruptcy prediction in Poland against the background of foreign epperience, Financial Internet Quartetly “e-Finanse”, vol. 15/no. 1, pp. 10-19.

Wędzki, D., 2005, Bankruptcy Logit Model for Polish Economy, Argumenta Oeconomia Cracoviensia, nr 3 , Kraków, pp. 49-70.

Xiao, Z., Yang, X., Pang, Y., Dang, X., 2012, The prediction for listed companies’ financial distress by using multiple prediction methods with rough set and Dempster-Shafer evidence theory, Knowl.-Based Syst. 26, pp. 196–206.