A COMPARISON OF VARIABLES SELECTION
METHODS AND THEIR SEQUENTIAL APPLICATION:
A CASE STUDY OF THE BANKRUPTCY OF POLISH COMPANIES

Mikhail Zanka, MSc

University of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Institute of Informatics
Banacha 2, 02-097 Warsaw, Poland
e-mail: raznem@gmail.com
ORCID: 0000-0002-7327-8842

Abstract

Research background: Even though in recent decades, a lot of new techniques were developed, there is still a lack of studies aimed at comparing the performance of variable selection methods. Bankruptcy prediction is an excellent example of the conservative research field with the tendency to use classical approaches. Although the results of studies in this field are directly applied in banks and other financial institutions, variables selected for these models can be biased by the author's preference for one technique.
Purpose: This work aims to compare different variable selection approaches and introduce a new methodology of sequential variable selection that can be applied when the low-dimensional model is preferred.
Research methodology: This study has been conducted on Polish companies' insolvency data from the period of 2007–2013. The risk has been modeled with logistic regression; hence variables have been selected with approaches suitable for linear models.
Results: The one-step methods did not lead to sufficient dimensionality reduction, while the sequential approach provided compact models keeping the high-performance level. Also, this method allowed us to identify the main financial determinants of insolvency for studied companies, which are the volume of total assets and the ratio of profit to total assets.
Novelty: This paper compares different variable selection methods and demonstrates the effectiveness of their sequential application for dimensionality reduction.

Keywords: Variable Selection, Ensemble Models, Bayesian Model Averaging, LASSO, Bankruptcy Prediction

JEL classification: C51, C52, C53, G33, M21

Introduction

Research of bankruptcy estimators do not have a clearly defined standard of variable selection. Some authors use a pool of variables that are selected and based on theoretical studies, which can cause a lack of empirical quality of the model. In other studies, authors start from a large number of variables and decrease the dimensionality of the model using only one preferred technique. However, this approach can cause suboptimal model specification. This paper investigates the performance of different variable selection methods in case of modeling the risk of companies' bankruptcy.

Model predictive power is strongly dependent on data availability. The most complete source of data available for researchers is the financial statements of joint-stock companies. The obligation to publish statements allows getting relatively large sets of diverse financial data. It leads to a high number of potential determinants and results in the problem of variable selection. Among variable selection methods, logistic regression and discriminative analysis are two of the most commonly used and both require independence of explanatory variables. Otherwise, there is no theoretical guarantee that estimators will be unbiased. However, most financial indicators are highly correlated which creates an additional challenge in using these two methods.

This study was made on the data containing financial indicators of Polish companies. Even in the case of only Polish data, the studies conducted over the last 20 years differ in terms of the variables used, and hence it is hard to select a specific set of determinants based on them. In addition to variable selection based on theoretical literature, there are two common quantitative approaches: selection with the t-statistic and stepwise regression, which were compared in this study with other approaches. In practice, the specifications obtained by researchers may vary even in the case of models estimated with an identical approach. As a result, it is hard to predict which set of variables will form the optimal specification. However, some typical variables appear in different models, therefore this work compares all obtained models as well as the results of other authors.

Modern studies of bankruptcy risk estimation vary considerably in terms of different estimation approaches. Some researchers still use the discriminative analysis proposed by E.I. Altman (1968), while in recent works, experiments with such new modeling methods as neural networks (Iturriaga, Sanz, 2015) and random forest (RF) classifiers (Barboza, Kimura, Altman, 2017) appeared. They are characterized by better predictive power than linear or decision tree models. However, higher complexity leads to difficulties in the inference of the

effects of explanatory variables, which directly influences the possibility to analyze obtained specifications from a theoretical perspective. In this work, we focus on the theoretical and quantitative interpretation of the results. Therefore, variables were selected for the logistic regression model, as it provides an interpretable estimation of variables effect for a binary target. An alternative approach is decision trees (DTs) based methods. For example, in the work of L. Obermann and S. Waack (2015), the authors used DTs to demonstrate that easy interpretable methods are comparable to more complex black-box methods for insolvency prediction. Since we decided to model insolvency with logistic regression, methods based on more complex models, like random forests or neural networks, will be omitted as they consider non-linear dependencies, which are not possible to fit with logistic regression without additional feature engineering. Therefore, while selecting different approaches, we had to choose the ones suitable for linear models. This requirement leads to the selection of methods that are well-suited for simpler models like linear or logistic regression. Hence, methods that are based on more complex models, like random forests or neural networks, consider non-linear dependencies, which are not possible to fit with logistic regression without additional feature engineering. Therefore, while selecting different approaches, we had to choose the ones suitable for linear models.

The main purpose of this paper is to empirically compare prevalent variable selection methods with newer and less common ones. The choice of the bankruptcy prediction is motivated by the fact that this is a well-studied problem both from a theoretical and modeling perspective. This work is structured as follows: section 2 presents the results of the studies of bankruptcy risk estimation, section 3 contains the results of the empirical study, while section 4 concludes these results.

1. Bankruptcy risk estimation studies

The fact that the difference in financial indicators characterized the threat of bankruptcy was noticed in R. Smith and A. Winakor (1935) and C. Merwin (1942) studies. This notice became a fundamental theoretical basis for modeling the risk of bankruptcy. Next, the Altman's Z-score and J. Ohlson's (1980) O-score formed one of the standard approaches for bankruptcy risk measurement, proved that financial data contains this information. According to the J. Sun (2014), methods of feature selection used in these works form a classical approach which, together with a quantitative and qualitative selection, can be used to define 3 ways of variable selection observed in the literature.

The authors of classic works on bankruptcy prediction choose variables based on their popularity in the literature and their own professional judgment. The modern successor of this approach is a qualitative selection when an author selects features based on theoretical conclusions without using the data in this process. This approach includes both selections of variables from the previous studies and a theoretical analysis of indicators made by the researcher.

The quantitative variables selection could be combined with qualitative methods. Therefore, the researcher starts from an initial variable set which can consist of all available variables or some qualitatively selected subset. For instance, classic t-test, correlation analysis or stepwise regression are one of the most popular approaches including recent studies that H.A. Alaka (2018) demonstrates in his review of bankruptcy prediction models. J. Pereira, M. Basto and A. Ferreira (2016) compared different methods of the stepwise approach, together with LASSO and ridge regression. S. Tian (2015) investigated the importance of variables widely used in the literature by applying the LASSO model to them. The author compared the results achieved with the quantitative approach, the purely theoretical choice of variables and R. Merton's (1974) structural model which is a popular bankruptcy risk measurement among practitioners. J. Traczynski (2017) used the Bayesian model averaging (BMA) as a basic approach to select the main predictors of the companies' default, which allowed to highlight the two most important indicators from an initial set of variables. H. Höglund (2017) successfully used a genetic algorithm for variable selection to build a model out of 36 and 72 variable sets. In this study, the main focus is the quantitative selection, so all these techniques together with simulated annealing (Rutenbar, 1989) and a genetic algorithm (Shin, Lee, 2002) were compared using a single data set.

Most of the variables in bankruptcy prediction models are related to liquidity, debt service capability, and profitability. For example, we can compare the classical Altman's model, the first insolvency logit model created by Ohlson and Hołda's logit model created for Polish companies. All of them use profit and volume of assets related variables. Both of the logit models have total liabilities in their specifications. In his empirical study, J. Traczynski (2017) concluded using the BMA approach that the ratio of total liabilities to total assets and the volatility of market returns are only significant variables in all industry groups and overall samples. Even though the author used a non-classical approach, this result is consistent with classical models. However, these models do not have exactly the same variables – each author represents the same indicators differently. It leads to a situation when the authors of empirical studies need to create and select variables in their work. This study is dedicated to the variable selection part of this process, especially to the case with the availability of a lot of financial indicators when an author needs to select the best predictors out of them.

As was shown, researchers of insolvency use various methods that lead to the following research question: What is the comparative performance of these methods in the case of the same data set? Also, taking into account the fact that different studies have variables describing the same features of the company, we can hypothesize that it is possible to highlight several most important determinants of insolvency which will be consistent with previous studies. The second hypothesis is that the sequential application of different variable selection methods can allow getting a higher reduction in dimensionality than a single model approach. A reduced number of variables also will be helpful in highlighting key indicators of bankruptcy risk.

2. Empirical study

2.1. Data pre-processing

The original dataset contains the financial indicators of companies from 2000–2012 and target values from 2007 to 2013 provided on the UCI Machine Learning repository by M. Zieba, S.K. Tomczak, and J.M. Tomczak (2016). The dataset consists of 5 different subsets with different bankruptcy terms. In this work, the only subset with a target variable indicating the fact of bankruptcy in the next year was used. All financial indicators with data gaps when they occurred for more than 5% of observations were dropped from the original variable set. Other indicators do not have a lot of gaps, so observations with undefined values were removed. After the data cleaning process 56 financial indicators and 5,538 companies left, 388 of which declared bankruptcy. This data was stratified sampled with a 0.7 train-test split ratio, a separate test subset allowed to estimate the predictive power of the model on new data.

2.2. Removal of correlated variables

First, the removal of correlated variables was applied to the dataset as one of the variable selection methods. Pairs of variables with a correlation higher than 0.7 were selected and variables with a higher average correlation in the correlation matrix were removed. This approach allowed reducing dimensionality from 56 to 23 variables. Logistic regression with these variables was compared with the model containing a full set of them. We can notice the Area Under Receiver Operating Characteristic Curve (ROC-AUC hereinafter AUC) drop from 0.638 to 0.5167 for the new model. However, the t-test shows the statistical significance of all coefficients. The new model's AUC is close to 0.5 which indicates the poor quality of this variable selection approach. At the same time, dimensionality reduction is not sufficient to distinguish the main bankruptcy

determinants. Therefore, of these two models, the one containing all indicators will be used as a baseline at the first stage of variable selection.

2.3. First stage of variable selection

At the first stage, the following variable selection approaches were used: LASSO, elastic-net, stepwise regression, BMA, GA, and SA. Also, ridge regression and regression on variables selected with principal component analysis (PCA) were used to illustrate a performance achievable by less interpretable linear models (Jolliffe, 2011). For the second model, 16 principal components were selected and based on the 90% of explained variance threshold. Then statistically non-significant variables were rejected which led to a model with 3 principal components.

Specifications gained by different algorithms were compared with the AUC and Akaike information criteria (AIC) (Friedman, Hastie, Tibshirani, 2013). Performance metrics and a number of variables for models obtained in this stage are shown in (Table 1). The number of variables for models built on principal components is marked with an asterisk, to highlight that these models require information from all original variables. Also, AIC values cells for models with regularization are empty as information criteria do not have an interpretation for them.

Table 1. Models performance at the first stage of variable selection

Model	Number of variables	AUC, training data	AUC, testing data	AIC
Logistic regression	56	0.6380997	0.5970997	–
Logistic regression, correlated variables dropped	23	0.5167212	0.5288804	–
Logistic regression on principal components	3*	0.8163424	0.7985610	1,753.3
LASSO regression	18	0.8156513	0.7960104	–
LASSO regression on principal components	10*	0.8087156	0.7943732	1,754.3
Ridge regression	56	0.8246938	0.8062405	–
Elastic net regression, alpha = 0.5	26	0.8163566	0.7967901	–
Elastic net regression, alpha = 0.5, principal components	12*	0.8087156	0.8000256	–
Backward stepwise regression	33	0.8817738	0.8269569	1,425.8
Forward stepwise regression	14	0.8412288	0.7834680	1,536.9
Bayesian model averaging	7	0.8254806	0.7980208	1,605.5
Genetic algorithm	44	0.5749300	0.5677849	19,629.0
Simulated annealing	21	0.8347237	0.8031526	1,576.0

* Number of variables for models built on PCA (also mentioned in the text)

Source: own elaboration.

The results of a models performance shows that the only genetic algorithm converged to the model with training AUC (0.5749) lower than the baseline, whereas it took the most computational time. All other approaches allowed to get models with the AUC higher than 0.8. Among them, backward stepwise regression has the highest values of performance metrics (train AUC – 0.8818). However, it converged to the largest specification with 33 variables which is not a sufficient reduction of dimensionality to be able to separate the main determinants and make the model easily interpretable. From that perspective, BMA results are the most interesting as they provide the highest dimensionality reduction without a significant performance loss with training AUC equals to 0.8256. The rest of the variable selection methods have a larger number of variables and average performance. The last of the variable selection techniques is forward stepwise regression which demonstrates good performance on training data (AUC – 0.8412) and a relatively small number of variables. Nevertheless, performance on the test data is below average which means that the model has the highest over fitting.
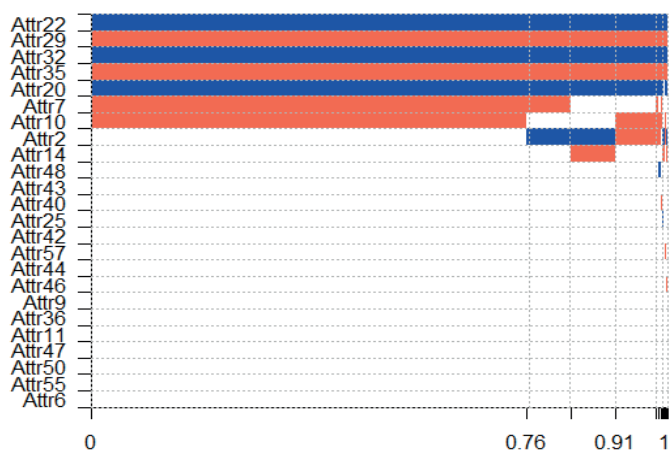

Figure 1. First stage BMA cumulative model probabilities

Source: own elaboration.

BMA results can be considered as optimal from the perspective of further theoretical interpretation. Figure 1 illustrates the probability of 35 the most likely specifications, where on the vertical axis are variables which appeared in them and on the horizontal axis is the cumulative probabilities of these variable sets. Cell color indicates a sign of parameter: red is positive, blue is negative. The figure shows the same number of variables for the 4 most probable models, so BMA clearly indicates a smaller dimensionality than the other approaches. Also, the

posterior probability of the selected model is equal to 76%, which indicates a confident choice of this specification over others.

2.4. Second stage of variable selection

At the next stage, only the BMA and LASSO regression were used. Variable selection was evaluated on 5 specifications from the first stage which were obtained by LASSO regression, GA, SA, forward and backward stepwise regressions. New models were compared with the following performance metrics: AIC, McFadden pseudo-R^2, AUC on training and testing data. Among these models, LASSO regression evaluated on a variable set from backward stepwise regression converge to one of the smallest specifications with 3 variables. Even though this model has the worst performance with both AUC around 0.797, it is close to the average values from the previous stage models. So, this specification can be treated as an extreme case of model dimensionality reduction providing a very small set of strong predictors. These predictors are the following indicators: *log(total assets), working capital / total assets and profit from sales / total assets.* This simple variable set covers all the most important indicators of a company's default risk: liquidity, debt service capability, and profitability.

LASSO regression on the forward stepwise selection provided a smaller specification containing only 2 variables. This specification differs from the first model of this stage by removing one variable from it. However, a lack of these variables causes a significant performance drop (train AUC – 0.7777, test AUC – 0.7632, pseudo-R^2 – 0.102). The next model is the LASSO regression on the genetic algorithm set which converged to the same specification as LASSO regression at the first stage. Therefore, these two models will not be considered in the further part of this work. The last LASSO regression was performed on a simulated annealing variables set and reduced dimensionality from 21 variables to 10 which makes it one of the biggest specifications. However, the performance of this model is below the average level with the lowest testing AUC (0.7902) among all of the compared models.

The next half of this stage is the BMA approach for variable selection with the same order that the first stage models used. In the case of BMA on backward stepwise regression results, the model converged to specification with 8 variables. In Figure 2 we can see the posterior probabilities of the two most probable models which have close values and differ only by one variable (*Attr14* and *Attr7*). Similar probability values can be explained by the high correlation between them which almost equals 1, whereas *Attr14* is *(gross profit + interest) / total assets* and *Attr7* is *gross profit / total assets*. So, the influence of the interest on the difference between these indicators is not significant and makes them interchangeable from the perspective of the

bankruptcy prediction model. Similar behavior can be observed for BMA on the simulated annealing variables set. The resulting model contains 7 variables and has an average performance according to all metrics. The posterior probability of this model is 0.25, whereas the second most probable model has 0.23. These models also differ by two variables, which also are highly correlated and represent the proportion of operating profit divided by sales revenues, but in a less probable case the operational depreciation of fixed assets was added to the numerator.
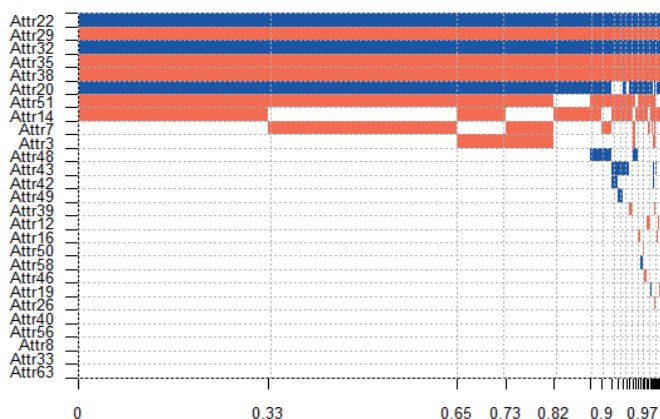

Figure 2. Cumulative probabilities of BMA on the backward stepwise regression variables set
Source: own elaboration.

At the second stage, in addition to previous performance metrics, the McFadden pseudo-R^2 (McFadden 1974) was used. These values, as well as BMA results from the previous stage, are shown in Table 2). The results indicate that BMA leads to the selection of models with better performance than LASSO regression. The model with the best fitting to the train set as well as to the test set was obtained by the Bayesian approach. However, LASSO regression converged to a much smaller specification than others provided a good performing model with only 3 variables. Also, it is worth noting that the first stage BMA allowed getting the model with the performance and variable reduction to a level of the second stage.

Table 2. Performance of the best models from the second stage of variable selection

Model	Number of variables	AUC, training data	AUC, testing data	AIC	Pseudo-R^2
1	2	3	4	5	6
Bayesian model averaging (previous stage results)	7	0.8254806	0.7980208	1,606	0.2350
LASSO regression (backward stepwise regression)	3	0.7967585	0.7970518	1,740	0.1500

1	2	3	4	5	6
LASSO regression (simulated annealing)	10	0.8086200	0.7902076	1,608	0.1955
Bayesian model averaging (backward stepwise regression)	8	0.8211599	0.7984051	1,594	0.2410
Bayesian model averaging (forward stepwise regression)	7	0.8345242	0.8078834	1,571	0.2550
Bayesian model averaging (LASSO regression)	7	0.8334707	0.8133465	1,578	0.2510
Bayesian model averaging (genetic algorithm)	5	0.8141387	0.7992181	1,631	0.2170
Bayesian model averaging (simulated annealing)	7	0.8263224	0.8045142	1,599	0.1968

Source: own elaboration.

Table 3. Variables occurrences frequency in second stage models comparison

Variable	Formula	Sign	Frequency
Attr35	Profit on sales / total assets	−	8
Attr29	log(Total assets)	−	7
Attr20	(Inventories × 365) / sales revenue	+	6
Attr10	Equity / total assets	−	4
Attr32	(Current liabilities × 365) / production cost of products sold	+	4
Attr48	(Operating profit – depreciation) / total assets	+	4
Attr22	Operating profit / total assets	+	3

Source: own elaboration.

Table 3 presents the financial indicators occurring 3 or more times in models shown in (Table 2). Signs of the parameters are also presented in this table and they all are identical among models. All but the last two shown variables are consistent with intuition and theory. The logarithm of total assets occurs in all models except BMA estimated on GA specification. 4 out of the 6 remaining variables are normalized by the total assets value, for example, *Attr35*, which is *profit on sales / total assets*, occurs in all second stage models. Variables from this table can be considered as the best determinants of bankruptcy prediction among all 56 financial indicators from the initial variables set. All these determinants except *Attr48* appear in the BMA from the first stage, which indicates the good performance of the Bayesian approach.

The *Attr48* and *Attr22* have positive signs indicating that their increase causes higher bankruptcy risk. Both of them represent operational profit values, which is a very common variable in other bankruptcy risk studies. For example, studies of Polish companies' bankruptcy determinants by B. Prusak (2005), E. Mączyńska and M. Zawadzki (2006) and M. Gruszczyński (2003) contain operational profit, in all of them. However, the sign of this variable is opposite to

our results. It could be explained by the fact that together with this variable *profit on sales / total assets* (*Attr35*) occurs in all second stage models. According to the ceteris paribus assumption, the probability of a company's bankruptcy increases together with an increase in operating profit and no change in profit from sales. Such change occurs when a company receives a higher profit not by increasing sales but by reducing administrative costs, which is the typical behavior of management in difficult financial situations.

Conclusions

This work compares different variable selection approaches including non-classical methods like BMA, GA or SA. The best set of variables from the perspective of performance were obtained by the backward stepwise regression. However, this model contains 33 variables which is one of the largest specifications in this study. The multistage approach used in this study consisted of two stages. The results of the first stage indicated the lack of sufficient dimensionality reduction among the most used models, whereas the second stage provided more compact models without a significant performance drop. At the first stage, BMA was the best solution for a single variable selection method which allowed getting a model with 7 variables.

Our results demonstrate the efficiency of a multi-stage approach, the second step of which was the estimation of the LASSO and BMA models on variable sets from the previous stage. As expected, it allowed further dimensionality reduction. For instance, the specification obtained with the LASSO regression estimated on the stepwise regression set from the previous stage contained 3 variables with a relatively good performance level. Other models have higher dimensionality; however, the performance is better than the BMA model from the first stage. The best-fitted specification was obtained by the application of BMA to the backward stepwise specification, which performed the best at the first stage. Therefore, the sequential application of different variable selection methods allowed achieving better results from both perspectives. Specifically, the application of BMA to the best performing model from the first step.

The frequency analysis of the second stage models variables shows the tendency to converge to financial variables describing the same aspects of companies' condition. These aspects are profit on sales, total assets, inventory, equity, current liabilities, and operating profit. The pool of the variables, as well as their signs, are consistent with the results of other empirical and theoretical studies. The volume of total assets is present in 7 out of 8 models and all of them have a company's profit normalized by total assets as a variable. This result indicates that these two indicators are the main financial variables in the case of insolvency prediction.

References

Alaka, H.A., Oyedele, O.L. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, *94*, 164–184. DOI: 10.1016/j.eswa.2017.10.040.

Altman, E.I. (1968). Financial Ratios, Discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, *23* (4).

Barboza, F., Kimura, H., Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417. DOI: 10.1016/j.eswa.2017.04.006.

Fleiss, J.L., Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613–619.

Friedman, J., Hastie, T., Tibshirani, R. (2013). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *The Mathematical Intelligencer*, *27* (2), 83–85. DOI: 10.1007/BF02985802.

Greenberg, E. (2012). *Introduction to Bayesian Econometrics*. New York, NY: Cambridge University Press. DOI: 10.1017/CBO9781139058414.

Gruszczyński, M. (ed.) (2012). *Mikroekonometria. Modele i metody analizy danych indywidualnych*. Warszawa: Wolters Kluwer Polska.

Gruszczyński, M. (2003). Modele mikroekonometrii w analizie i prognozowaniu zagrożenia finansowego przedsiębiorstw. *Zeszyty Polskiej Akademii Nauk*, *34*.

Hołda, A. (2000). *Optymalizacja i model zastosowania procedur analitycznych w rewizji sprawozdań finansowych*. Kraków: Akademia Ekonomiczna.

Höglund, H. (2017). Tax payment default prediction using genetic algorithm-based variable selection. *Expert Systems with Applications*, *88*, 368–375. DOI: 10.1016/j.eswa.2017.07.027.

Iturriaga, F.L., Sanz, I.P. (2015). Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks. *Expert Systems with Applications*, *42* (6), 2857–2869. DOI: 10.1016/j.eswa.2014.11.025.

Jolliffe, I. (2011). Principal Component Analysis. In: M., Lovric (ed.), *International Encyclopedia of Statistical Science*. Berlin–Heidelberg: Springer. DOI: 10.1007/978-3-642-04898-2_455.

Karr, C. (1995). Adaptive control of an exothermic chemical reaction system using fuzzy logic and genetic algorithms. In: L.R., Medsker (ed.), *Hybrid intelligent systems* (pp. 187–203). New York, NY: Springer US.

Mączyńska, E., Zawadzki, M. (2006). Dyskryminacyjne modele predykcji bankructwa przedsiębiorstw. *Ekonomista*, *2*, 205–235.

McFadden, D.L. (1974). *Conditional logit analysis of qualitative choice behaviour*. *Frontiers in econometrics* (pp. 105–142). New York: Academic Press.

Merton, R. (1974). On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *Journal of Finance*, *29*, 449–470.

Merwin, C. (1942). *Financing small corporations in five manufacturing industries, 1926–1936*. New York: National Bureau of Economic Research.

Obermann, L., Waack, S. (2015). Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks. *Expert Systems with Applications*, *42* (23), 9117–9128. DOI: 10.1016/j.eswa.2015.08.009.

Ohlson, J. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, *18* (1), 109–131. DOI: 10.2307/2490395.

Pereira, J., Basto, M., Ferreira, A. (2016). The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. *Procedia Economics and Finance*, *39*, 634–641. DOI: 10.1016/S2212-5671(16)30310-0.

Prusak, B. (2005). *Nowoczesne metody prognozowania zagrożenia finansowego przedsiębiorstwa*. Warszawa: Difin.

Rutenbar, R.A. (1989). Simulated annealing algorithms: an overview. *IEEE Circuits and Devices Magazine*, *5* (1), 19–26. DOI: 10.1109/101.17235.

Smith, R., Winakor, A. (1935). Changes in financial structure of unsuccessful industrial corporations. *The Accounting Review*, *10* (2), 174–184.

Shin, K., Lee, Y. (2002). A genetic algorithm application in bankruptcy prediction modelling. *Expert Systems with Applications*, *23* (3), 321–328. DOI: 10.1016/S0957-4174(02)00051-9.

Sun, J., Hui, L. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, *57*, 41–56. DOI: 10.1016/j.knosys.2013.12.006.

Tian, S. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, *52*, 89–100. DOI: 10.1016/j.jbankfin.2014.12.003.

Traczynski, J. (2017). Firm Default Prediction: A Bayesian Model-Averaging Approach. *Journal of Financial and Quantitative Analysis*, *52* (3), 1211–1245. DOI: 10.1017/S002210901700031X.

Zieba, 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. DOI: 10.1016/j.eswa.2016.04.001.