The impact of Data structure on classification ability of financial failure prediction model

Lucia Svabova[1,*], and *Lucia* Michalkova[1]

[1]Department of Economics, Faculty of Operation and Economics of Transport and Communications, University of Zilina, Univerzitna 1, 010 26 Zilina, Slovakia

Abstract. The creation of prediction models to reveal the threat of financial difficulties of the companies is realized by the application of various multivariate statistical methods. From a global perspective, prediction models serve to classify a company into a group of prosperous or non-prosperous companies, or to quantify the probability of financial difficulties in the company. In many countries around the world, real financial data about the companies are used in developing these prediction models. In Slovakia, standard data from the financial statements and annual reports of Slovak companies are used for the creation of the company's failure model. Since in this case there are generally large data files, it is necessary to pre-process the data by the selected methods before the prediction model is constructed. A database of the companies needs to be prepared for the subsequent application of statistical methods, and it is also highly appropriate to focus globally on the detection of potential extreme and remote observations. Therefore, the article will focus on quantifying the impact of the data structure detected, for example, the occurrence of extreme and remote observations in the data set, on the resulting overall classification of the prediction ability of the models created.

1 Introduction

Initial data processing for further statistical and econometric analyzes is a very important part of the analyst's work. This preparation of data requires a lot of time, analyst experience, knowledge of the data and the situation we are trying to analyze. Such data preprocessing is also needed when analyzing the issue of predicting the financial difficulties of businesses, a topic that is current in recent years.

The issue of predicting the financial situation of companies is relatively young field of economic research. Its origin dates back to the 30s´ of the 20th century, but the first prediction models have appeared only in the 60s´ of the 20th century [1]. As a first study focused on finding the main differences between companies with and without financial problems, based on the analysis of the financial ratios, can be considered the work of Fitzpatrick from 1932 [2]. Since then, prediction financial analysis has undergone significant developments, from one-dimensional and multidimensional discriminant

* Corresponding author: lucia.svabova@fpedas.uniza.sk

analysis, through logistic regression to artificial intelligence. At present, experts' views on various methods of predicting the financial situation of the companies differ. Some authors deal with the possibility of using models developed in the last century for predicting bankruptcy of existing companies at present. This results in different adjustments and recalculations in the original models. Other authors focus on creating new models using new ratios and new methods [3]. As a result of the development of artificial intelligence, new methods such as machine learning techniques, neural networks and genetic algorithms are being introduced into prediction financial analysis. Given the different opinions of experts on various prediction methods, it can be argued that every method has its advantages and disadvantages, and also limitations of its use [4]. But the constant research in this area proves currentness of this topic even today. In any case, the issue of predicting the financial situation of a company is still up to date due to its great importance not only for the company itself but also for all stakeholders [5].

The created prediction models are evaluated in terms of their success in the correct classification of companies, especially in the correct prediction of financial difficulties. In this article we will focus on the analysis of the impact of data preparation in the process of developing prediction models on the results of the correct classification. The aim of the paper is to find out whether the identification and removal of remote and extreme values in the data set results in an improvement of the classification ability of the model created by discriminant analysis, logistic regression and classification tree method (namely CART). The contribution of the paper is a new view of the prediction ability of the created models, where the emphasis is on thorough data preparation.

Our study is organized in four chapters. The first one provide the introduction to the issue of bankruptcy prediction and highlight the current state of the issue in the form of literature review. The second chapter describes briefly the methods and the data used in this study. The third chapter provide the results of the analysis of the impact of outlier occurrence on the classification ability of the prediction model created by three selected methods. Discussion compares the results of the methods used in this study and discuss the weaknesses and next direction of the study.

1.1 Literature review

The prediction of bankruptcy is a topic that has been in recent years dealt with by economists in many countries of the world. As a first study focused on creation of prediction model based on the analysis of the differences of financial ratios, is considered the study of Fitzpatrick from 1932 [6]. Since then, prediction financial analysis has undergone significant developments, from one-dimensional and multidimensional discriminant analysis, through logistic regression to artificial intelligence. The method of discriminant analysis was used for the first time by Beaver in 1966, who also formed the basis for prediction models. Based on his research, in 1968 Altman used multivariate discriminant analysis to develop the probably most famous bankruptcy prediction model [2]. Ohlson's study from 1980 was the first who used the method of logistic regression for creating the model to predict the probability of company failure [7].

At present, authors used different methods of creating of the prediction models: from the older methods of discriminant analysis and logistic regression, to more modern methods of neural networks, genetic algorithms, classification trees, and random forests [8]. Several prediction models were in the last few years also created in Slovakia. As the best known and frequently used we can call models by Gurcik from 2002 and Chrastinova from 1998 [5]. In recent years, several authors created new prediction models in the conditions of Slovakia. Kovacova and Kliestik in [9] developed models for bankruptcy prediction of Slovak companies by using logit and probit method and provide the comparison of overall

prediction power of the two developed models. Gavurova et al. in the study [10] analyzed the impact of trend variables on the prediction power of the models constructed using discriminant analysis and decision trees. They developed a new model for Slovak companies by using the decision tree technique. Mihalovic in [11] also dedicated his study to development of bankruptcy prediction models in Slovak Republic, the first one is estimated via discriminant analysis, and another is based on a logistic regression. Other authors in Slovakia deal with the application of existing models to predict the financial difficulties of companies in Slovakia, for example [12], [13].

Several authors have also dealt with the occurrence of outliers in data used for bankruptcy prediction models in recent years. They mostly examined the impact of outliers on the resulting prediction power of the created models. For example, Tsai and Cheng in [14] studied bankruptcy prediction performance achieved after removal of different outlier volumes from datasets. Linares-Mustaros et al. in [15] dealt with problems of the occurrence of outliers in using cluster analysis to classify firms according to their financial structures. Alrawashdeh et al. in [16] tried to eliminate the problem of high sensitivity of linear discriminant analysis to the occurrence of outliers in data and to improve the classification ability of created models also in bankruptcy prediction. Figini et al. in their study [17] describes novel approaches to predict default for SMEs by detecting multivariate outliers. Pawelek et al. in [18] made an empirical study about the influence of detecting and eliminating outliers on the effectiveness of the bankruptcy prediction logit model for Polish companies. A similar issue is addressed in their subsequent studies [19] and [20].

2 Materials and Methods

Outlying and extreme observations are observations in the statistical set that are significantly smaller or larger than other values. These values can occur in the data file for various reasons. They can occur as an error in records, most often caused by a human factor, for example when manually rewriting records into electronic form. Outlier may also appear in the file as a measurement that is actually significantly different from the others [18].

Outlying and extreme observations may signal various anomalies in the data that need to be addressed in the pre-data phase to further apply more advanced statistical methods. Some methods are very sensitive to the occurrence of such values in the file. In general, it is recommended to first detect extreme (but also outlying) observations and then analyze them and consider removing them from the data file [20]. It is advisable to remove those which, from an expert point of view, represent problematic points and misrepresent the parameters of the regression function. The solution of course depends on the specific application and the analyst's decision.

A special group of outliers observed are the so-called multivariate outliers. Multidimensional observations can become outliers if their values of multiple variables are some unique combination, different from the combination of variable values for other units in the set. A suitable metric to identify outlying multivariate outliers is the Mahalanobis distance. This metric measures the multidimensional distance of each observation from the group centroid. In this paper, we focus on detecting multivariate outliers using Mahalanobis distance according to [21].

Subsequently, we will focus on comparing the predictive ability of models detecting the financial difficulties of companies in Slovakia. We will compare models created using discriminant analysis, logistic regression and the CART binomial tree method, both models that originate from a data file from which multivariate outlying values have been removed, and also from the one in which they were left.

2.1 Data used in the study

In this study we use the database of financial indicators of Slovak enterprises from 2016 and 2017. In total, there are 45,458 enterprises in the database. Data are from Amadeus - A database of comparable financial information for public and private companies across Europe. The data contains values of 21 financial ratios of Slovak enterprises from 2016. At the same time, the database contains a variable identifying the financial difficulties of the enterprise in 2017. A more detailed description of the variables as well as the identification of non-prosperous enterprises is given in [22]. The following table lists the numbers of prosperous and non-prosperous enterprises in the database.

Table 1. The numbers of prosperous and non-prosperous companies in the database.

Prosperity		
	Frequency	Percent
prosperous	22778	50.1
non-prosperous	22680	49.9
Total	45458	100.0

Using the multivariate outliers identification methodology described in [21], we identified a total of 555 outlying companies in the dataset, see the following table.

Table 2. The numbers of outliers and non-outliers in the database.

outlier		
	Frequency	Percent
non-multivariate outlier	44903	98.8
multivariate outlier	555	1.2
Total	45458	100.0

Of these, 342 companies were prosperous and 213 were non-prosperous. On this sample of Slovak enterprises we created a model of prediction of non-prosperity by three methods: discriminant analysis, logistic regression and CART method of binomial trees. Then we analyze the sensitivity of the models prediction power to the presence of outliers in the data file. The prediction power of models is assessed on the basis of the classification table, mainly on the basis of the percentage of correctly identified non-prosperous enterprises. We also use the AUC value under the ROC curve.

3 Results

3.1 Models created by Discriminant analysis

In the first step, we created a model from a database that does not remove companies that were marked as multivariate outliers. The first model, created by discriminant analysis, achieved a total prediction ability of 70.4%. At the same time, 74.4% of non-prosperous enterprises were correctly classified. AUC of this model is 0.778.

We then removed companies that were identified as multivariate outliers from the database and created the same model using discriminant analysis without them. Table 3 shows a comparison of the classification of the created models. The overall classification capability of the model created improved by 1.2% when multivariate outliers were removed. However, for non-prosperous businesses, the correct classification has improved by up to 12%. The size of the AUC increased to 0.807.

Table 3. Classification results of models created by discriminant analysis.

Classification Results											
Original database						**Database without outliers**					
Y_2017			Predicted Group Membership		Total	Y_2017			Predicted Group Membership		Total
			0	1					0	1	
Original	Count	0	15151	7627	22778	Original	Count	0	12709	9727	22436
		1	5816	16864	22680			1	3045	19422	22467
	%	0	66.5	33.5	100.0		%	0	56.6	43.4	100.0
		1	25.6	74.4	100.0			1	13.6	86.4	100.0
a. 70.4% of original grouped cases correctly classified.						a. 71.6% of original grouped cases correctly classified.					

3.2 Models created by Logistic regression

The first logistic regression model, created from the original data set, achieved a total of 75% of the correct classification, while 83.4% of the non-prosperous enterprises were correctly classified. The AUC of this model was 0.841.

The second model, created from the dataset from which we removed multivariate outliers, achieved a better overall classification of 85.8%. The correctly classified non-prosperous enterprises in this case were 84.1%. The AUC of this model is 0.921.

The following table compares the classification of both logistic regression models.

Table 4. Classification results of models created by logistic regression.

Classification Results							
Original database				**Database without outliers**			
Observed	Predicted			Observed	Predicted		
	0	1	Percentage Correct		0	1	Percentage Correct
0	15200	7578	66,7	0	19625	2811	87,5
1	3771	18909	83,4	1	3576	18891	84,1
Overall Percentage			75,0	Overall Percentage			85,8
a. The cut value is ,500				a. The cut value is ,500			

3.3 Models created by CART tree

Using the binomial tree method, we also created two models to predict the financial difficulties of a company. The first model, created from the original database, achieved an overall correct classification of 89.1%. However, the correct classification of non-prosperous enterprises is only 88.1%. The AUC of this model is 0.911.

The second model, created after the removal of multivariate outliers, achieved an overall correct classification of 88.7%. This is even a little bit worse classification in the test sample than the original data set. In this case, non-prosperous businesses were correctly classified at 87.8%. The AUC of this model is 0.945.

The following table compares the classification capability of both CART models in a test sample that was 20% of the data set.

Table 5. Classification results of models created by CART.

Classification Results									
Original database					**Database without outliers**				
Sample		Predicted			Sample		Predicted		
		0	1	Percent Correct			0	1	Percent Correct
Test	0	4086	453	90,0%	Test	0	4073	475	89,6%
	1	540	4009	88,1%		1	551	3983	87,8%
	Overall Percentage	50,9%	49,1%	89,1%		Overall Percentage	50,9%	49,1%	88,7%
Growing Method: CRT Dependent Variable: Y_2017					Growing Method: CRT Dependent Variable: Y_2017				

4 Discussion

Multivariate outliers have been identified in the enterprise database as those enterprises that have a significantly different combination of financial ratios than other enterprises in the database. From the classification results in predicting financial difficulties, it can be concluded that removing multivariate outliers from the database improves the results achieved.

The discriminant analysis model improved by 1.2% in the overall classification, but up to 12.1% improved in the classification of non-prosperous enterprises. So that, removing outliers in the discriminant analysis significantly improved the percentage of correctly classified non-prosperous companies. The elimination of outliers therefore has a significant impact on the prediction of the company's non-prosperity. The logistic regression model has improved the overall percentage of correct business classification. The model improved by 10.7% in the overall classification and by 0.7% in the classification of non-prosperous enterprises.

On the model created by the binomial tree method CART, the elimination of outliers from the database does not have a significant impact on prediction power of the model. After removing the outliers, the CART model achieved almost the same classification results.

The weakness of this study can be considered that the used financial ratios of the companies have not been analyzed in terms of other assumptions that should be met in the methods. Therefore, we would see the further direction of this study in the analysis of the impact of multicolinearity among variables on the prediction ability of the created models.

5 Conclusion

In this paper, we focused on the prediction ability of prediction models of non-prosperity of Slovak companies. Models were created using three frequently used methods: discriminant analysis, logistic regression, and CART binomial tree method. We investigated the impact of identifying multivariate outliers in the enterprise database and removing them from the database on the resulting prediction power of the models. We assessed both the percentage of correct classification and the percentage of correctly classified non-prosperous enterprises. In summary, removing outliers from the database improves the classification ability of the generated discriminant model as well as the logistic regression model. Removing outlying companies from the database does not affect the classification of CART model.

This research was financially supported by the Slovak Research and Development Agency – Grant NO. APVV-17-0546: Variant complex model of Earnings management in conditions of Slovak republic as an essential tool of the market uncertainty.

References

1. M. Durica, I. Podhorska, P. Durana, Business failure prediction using cart-based model: A case of Slovak companies. *Ekonomicko-manazerske spektrum* **13**, 1, 51-61 (2019)
2. J. Salaga, V. Bartosova, E. Kicova, Economic value added as a measurement tool of financial performance. *Procedia Economics and Finance* **26**, 484-489 (2015)
3. K. Valaskova, K. Kramarova, B. Kollar, Theoretical aspects of a model of credit risk determination- Credit risk. *Advances in Education Research* **81**, 401-406 (2015)
4. Svabova, Durica, 2016
5. N. Shpak, O. Sorochak, M. Hvozd, W. Sroka, Risk evaluation of the reengineering projects: A Case Study Analysis. *Scientific Annals of Economics and Business* **65**, 2, 215-226 (2018)
6. P. Adamko, E. Spuchlakova, K. Valaskova, The history and ideas behind VaR. *Procedia Economics and Finance* **24**, 18-24 (2015)
7. J. Oláh, G. Karmazin, D. Máté, J.K. Grabara, J. Popp, The effect of acquisition moves on income, pre-tax profits and future strategy of logistics firms. *Journal of International Studies* **10**, 4, 233-245 (2017)
8. M. Dobrodolac, L. Svadlenka, M. Cubranic-Dobrodolac, S. Cicevic, B. Stanivukovic, A model for the comparison of business units. *Personnel Review* **47**, 1, 150-165 (2018)
9. M. Kovacova, T. Kliestik, Logit and Probit application for the prediction of bankruptcy in Slovak companies. *Equilibrium* **12**, 4, 2017.
10. B. Gavurova, F. Janke, M. Packova, M. Pridavok, Analysis of impact of using the trend variables on bankruptcy prediction models performance. *Ekonomicky casopis* **65**, 4, 2017.
11. M. Mihalovic, Performance comparison of multiple discriminant analysis and logit models in bankruptcy prediction. *Economics & Sociology* **9**, 4, 2016.
12. P. Adamko, L. Svabova, Prediction of the risk of bankruptcy of Slovak companies. In M. Culik (Ed.). *Proceedings of the 8th International scientific conference managing and modelling of financial risks*. Ostrava, Czech Republic, 2016.
13. K. Valaskova, L. Svabova, M. Durica, Verifikácia predikčných modelov v podmienkach Slovenského poľnohospodárskeho sektora. *Economics, Management, Innovation* **9**, 3, 30-38, 2017.
14. C.F. Tsai, K.C. Cheng, Simple instance selection for bankruptcy prediction. *Knowledge-based Systems* **27**, 2012.
15. S. Linares-Mustaros, G. Coenders, M. Vives-Mestres, Financial performance and distress profiles. From classification according to financial ratios to compositional classification. *Advances in Accounting* **40**, 2018.
16. M.J. Alrawashdeh, T.R. Radwan, K.A. Abunawas, Performance of linear discriminant analysis using different robust methods. *European Journal of Pure and Applied Mathematics* **11**, 1, 2018.
17. S. Figini, F. Bonelli, E. Giovannini, Solvency prediction for small and medium enterprises in banking. *Decision Support Systems* **102**, 2017.

18. B. Pawelek, K. Galuszka, J. Kostrzewska, M. Kostrzewski, Classification methods in the research on the financial standing of construction enterprises after bankruptcy in Poland. In F. Palumbo, A. Montanari, M. Vichi (Eds.). *Data science: Innovative developments in data analysis and clustering*. Springer International Pulishing, 2015.
19. J. Kostrzewska, M. Kostrzewski, B. Pawelek, K. Galuszka, The classical and Bayesian logistic regression in the research on the financial standing of enterprises after bankruptcy in Poland. In M. Papiez & S. Smiech (Eds.). *Proceedings of 10th professor Aleksander Zelias international conference on modelling and forecasting of socio-economic phenomena*. Cracow: Foundation of the Cracow University of Economics, 2016.
20. B. Pawelek, J. Kostrzewska, A. Lipieta, The problem of outliers in the research on the financial standing of construction enterprises in Poland. In M. Papiez & S. Smiech (Eds.). *Proceedings of 9th professor Aleksander Zelias international conference on modelling and forecasting of socio-economic phenomena*. Cracow: Foundation of the Cracow University of Economics, 2015.
21. B.G. Tabachnick, L.S. Fidell, *Using multivariate statistics*. Boston, USA: Pearson Education, 2013.
22. L. Svabova, M. Durica, I. Podhorska, Prediction of Default of Small Companies in the Slovak Republic. *Economics and Culture* **15**, 1, 88-95, 2018.