Handling outliers in bankruptcy prediction models based on logistic regression

Tünde Katalin Szántó[1]

Abstract

The primary tool for managing bank default risk is the credit rating of potential customers. The focus of the present study is on the logistic regression method used to construct 95% of the lender scorecards. The aim of the research is to determine how much the treatment of outliers improves the classification accuracy of the models when using a method that is highly sensitive to outliers, and which method of treating outliers results in the highest classification accuracy. Furthermore, what criteria should be used to determine the cut-off value of the models for a sample that does not contain solvent and insolvent businesses in equal proportions. The analysis was carried out on a sample of 1677 construction companies. The results show that the treatment of outliers significantly improves the predictive ability of the models, while the replacement of outliers with the closest non-outlier proved to be the most effective for treating outliers. When determining the cut-off, it is inappropriate to use the value that results in the highest classification accuracy, as this may lead to an increase in the proportion of first-order errors. The optimisation of this value may depend on the degree of credit risk taken by a given financial institution in its portfolio of loans.

KEYWORDS: bankruptcy prediction, default risk, logistic regression, outlier management
JEL-CODES: G33, C6, G17
DOI: https://doi.org/10.35551/PFQ_2023_3_5

[1] PhD student, University of Szeged, Faculty of Economics, szanto.tunde.katalin@o365.u-szeged.hu

Introduction

The importance of corporate bankruptcy prediction has increased in recent decades, with the bankruptcy of the borrower companies being the main cause of banking crises in the Japanese and Nordic banking systems in the early 2000s, highlighting the importance of assessing the survivability of customers when lending. Banks are the most important users of bankruptcy prediction models, but they can also be useful for audit firms and bond selling companies (Virág 2004). Currently, in the aftermath of the coronavirus epidemic and the war situation, the entire global economy is characterised by great uncertainty, which makes it even more important to effectively predict the bankruptcy of businesses. From a banking perspective, effective credit rating is particularly important, as the coronavirus epidemic has significantly increased banks' operating costs (Doma-Kozma 2022).

A record number of liquidation proceedings were opened in Hungary in the Q3 of 2022, with the number of liquidations opened in September four times up the number a year earlier (Opten 2022).[2] The subject of our study is bankruptcy prediction for Hungarian construction companies based on logistic regression. The reason for our choice is that the construction industry is a dominant sector in the Hungarian economy, in terms of gross value added and the increase in the number of businesses (KSH 2021/a). The use of logistic regression in bankruptcy prediction has been widespread since the 1970s, and 95% of bank lender scorecards are still based on this method, so banks prefer to use it in their lending decisions (Nyitrai-Virág 2017). Logistic regression is a very popular technique in bankruptcy prediction, as it does not require a normal distribution of variables, but it is very sensitive to outliers, which are mainly typical for bankrupt businesses, and therefore it is necessary to deal with these values when building models (Nyitrai 2017). Two of the most common practices for dealing with outliers are the replacement of outliers with the closest values which are no longer outliers, and the exclusion of outliers from the sample (Nyitrai-Virág 2017, Nyitrai 2017).

The sample is a large sample of 1,677 construction companies of which 1,293 are still in operation and 384 are currently in liquidation. The source of the reports is the Crefoport database. The sample does not include active and insolvent businesses in equal proportions, which is consistent with the fact that in reality, well-functioning, solvent businesses are present in the economy in a much larger proportion.

[2] Source: https://www.opten.hu/kozlemenyek/meglodultak-a-felszamolasi-eljarasok-fizeteskeptelenseget-jelez-vagy-covid-utohatas Downloaded on: 05.03.2023

THE HISTORY OF BANKRUPTCY PREDICTION

Bankruptcy in the legal sense means insolvency, bankruptcy being the event when a company is unable to meet its payment obligations on time. However, bankruptcy is not a sudden situation, but a longer process, a possible outcome of a period of financial difficulties. Financial difficulties arise in the life of a business when value destruction occurs due to inefficiency of assets or a poorly designed asset portfolio. The destruction of value reduces the market value of the asset portfolio, which in turn increases the financing leverage of the business. These effects lead to liquidity problems, which can result in insolvency, which in legal terms means bankruptcy (Pálinkó-Svoób 2016).

But financial difficulties and bankruptcy do not necessarily occur together. The management may recognise the value destruction in time and successfully prevent the insolvency by restructuring the company's operations. It is also possible, however, that an external shock can lead to bankruptcy without any warning. This is most often the case for companies with low capital and asset portfolio and which finance their operations with short-term funds. In a recession, it becomes more difficult to renew short-term funds, which can lead to a sudden liquidity shortage and insolvency (Pálinkó-Svoób 2016).

There are two types of insolvency proceedings in Hungary. Bankruptcy is a type of reorganisation procedure, the aim of which is for the debtor to reorganise its business after reaching an agreement with its creditors and to continue to operate thereafter. Liquidation proceedings, on the other hand, are a type of winding-up procedure, the aim of which is not to effectively reorganise the debtor's business but to wind up the debtor company without succession, seeking to satisfy creditors' claims as fully as possible (Piller 2013).

Most banks use some form of statistical scoring system in their lending practices to determine the probability of bankruptcy. These scoring systems can work with a wide range of inputs. General experience shows that variables that describe the behaviour of borrowers work best in determining the probability of bankruptcy, and these variables lead to higher accuracy than, for example, the use of financial ratios. Furthermore, a weakness of accounting data is that the balance sheet and income statement data can be manipulated in certain circumstances (Cziglerné 2020). The problem, however, is that there are no databases to monitor the behaviour of borrowers (Mikolasek 2018). For this reason, the use of financial ratios is common in the practice of bankruptcy forecasting, as the data of the annual accounts are publicly available to anyone. The scoring systems are based on objective factors and their analysis covers the whole area of the operation of businesses, thus providing a comprehensive picture of their management (Zéman-Hegedűs-Molnár 2018).

Before the last third of the 20th century, there was no adequate IT background or advanced statistical methods that would have allowed the creation of accurate bankruptcy prediction models, but even then, attempts were being made to find methods to predict the future of a company. At that time, they tried to compare different financial ratios of bankrupt companies and draw conclusions about the future of businesses based on those (Virág-Kristóf 2006). in 1930, for example, the Bureau of Business Research compared 24 sets of financial ratios of 29 firms to determine what similarities could be observed across bankrupt businesses (Bellovary-Giacomino-Akers 2007).

The first modern bankruptcy prediction model was created by Beaver in 1960. His work was based on a univariate discriminant analysis. The method consists of examining a financial ratio to decide whether a company can be classified as insolvent or as a survivor. His sample included 79 solvent and 79 insolvent companies. He examined 30 financial ratios and concluded that healthy and insolvent companies differ most in terms of Cash flow/Assets, Cash flow/Debt and Net sales/Debt (Virág 2004). The most reliable result was achieved by using the Cash flow/Total assets ratio, with which it became possible to predict bankruptcy with 90% accuracy one year before the insolvency occurred (Virág-Kristóf 2006). However, the disadvantage of the univariate discriminant analysis is that it often leads to inconsistent results, with one financial ratio being used to judge a company as a survivor and another as a defaulter (Virág 2004).

Edward I. Altman published his model in 1968, which was based on multivariate discriminant analysis. The sample he used consisted of 33 insolvent and 33 solvent small and medium-sized companies, as large companies rarely went bankrupt at that time. He examined a total of 22 financial indicators, five of which he eventually incorporated into his model, which is a linear function analysis, weighting the five variables by objective ratios, the sum of which gives a “Z” value. By comparing the “Z” value to a specified cut-off point it may be determined whether the company should be classified as a survivor or bankrupt (Virág 2004).

Although the use of multivariate discriminant analysis was a pioneer in the methodology, a problem in its application is that it requires that the variables be statistically independent, but there is often multicollinearity between financial indicators, which violates this requirement. In addition, it is important that the indicators follow a normal distribution. This constraint is overcome by logistic regression-based bankruptcy forecasting, which does not require the variables to be normally distributed. The method fits a logistic regression function to the observations using the maximum likelihood method (Virág-Kristóf 2006). A bankruptcy model based on logistic regression was first used by Ohlson to predict bankruptcy.

The application of probit regression in bankruptcy forecasting started in the 1980s, with Zmijewski being the first to build a model based on probit regression (Kristóf-Virág 2019). His sample included 800 solvent and 40 insolvent companies. The model uses three variables, the return on assets, the ratio of liabilities to assets and the liquidity ratio. It achieved an outstanding classification accuracy of 98% for the original sample (Zmijewski 1984).

Decision trees were first applied to bankruptcy prediction by Frydman, Altman and Kao in 1985. The use of decision trees is extremely popular, as it does not require the statistical conditions discussed earlier to be met (Kristóf-Virág 2019). A popular decision tree based method is the recursive parsing algorithm. The method works with univariate separation, splitting the data into two steps to form tree branches.

The initial data set is a sample in which it is known in advance which companies belong to the solvent and insolvent categories. The procedure examines the variables one by one, building the tree along the variables with the most separating values in order to create the most homogeneous classes possible. By grouping the data in terms of the dependent variable, the method aims to minimise variance within groups and maximise variance between groups (Virág-Kristóf 2006).

Another popular method based on decision trees is Chi-squared automatic interaction detection (CHAID). This procedure breaks down the set of values of the explanatory variable into intervals, then examines the class intervals by pair to determine whether the class intervals and the classification of the companies in them (bankrupt or solvent) are independent of each other, and if they are, the two class intervals are unified. The process continues as long as there remain only class intervals that are not statistically independent. As a result, the set of values of the explanatory variable is decomposed into class intervals (Nyitrai 2017).

In the 1990s, artificial intelligence, including the use of neural networks in bankruptcy prediction, came to the fore. Neural networks are parallel, distributed information processing devices that operate on a hardware or software basis. Networks are made up of several interconnected neurons and, unlike the methods discussed earlier, they acquire their abilities by learning. The way neurons are connected is different for each network. Neurons consist of three main layers, the input layer, the intermediate layer and the output layer. Neural networks learn through examples, and sufficiently trained networks can be used to make predictions on other data (Kristóf 2005). This method was first used by Odom and Sharda to predict corporate insolvency. Their three-layer neural network model outperformed the classification accuracy of models built using previous methods. Models based on neural networks have evolved continuously since then and are still a popular method today (Kristóf-Virág 2019). However, when used, the phenomenon of over-learning can be a problem. This means that the network does not learn the general problem during the learning process, but the specifics of the pattern on which it is built. In such cases, the model can no longer be used effectively on other databases. To avoid this problem, the database is usually divided into learning and testing databases. The net is trained and taught on the learning sample, and then tested to see how it performs on the previously unknown test sample. If the classification accuracy of the testing sample is similar to that of the learning sample, then the learning can be considered effective (Virág-Kristóf 2006).

Neuro-fuzzy systems started to be used in the early 2000s to predict corporate insolvency, with classification accuracy that surpassed traditional neural network models (Kristóf-Virág 2019). As data mining methods continue to improve, more accurate results are being achieved, but machine learning techniques have been hit by a number of hiccups. One of the criticisms is the black box problem, i.e. that only the inputs to the model and the outputs of the calculation are known at the time of modelling, but not the proportion of each variable in the model. Another problem is that the statistical significance of the variables cannot be tested (Nyitrai 2014).

It can therefore be seen that since the beginning of bankruptcy prediction, a wide variety of methods have been used in this area. Du Jardin (2010) estimates that in total more than fifty different methods have been published using more than five hundred different financial variables (Nyitrai 2017). The five most popular methods were the multivariate discriminant analysis, logistic regression, neural network, contingent claims analysis, and univariate analysis (Kiss-Kosztopulosz-Szládek 2021).

Due to the high number of techniques that can be applied, research is nowadays focused on improving existing methods (Nyitrai 2015). In line with this, one of the objectives of our research is to investigate which of several options for handling outliers results in higher classification accuracy.

The sample and methodology

The sample used for our study is a large sample of 1677 construction companies. Of the businesses in the sample, 1,293 are still active and 384 are in liquidation, showing that in reality solvent businesses represent a larger share of the economy than those that are going bankrupt. The data source is the Crefoport database.

The state of the construction industry in Hungary

The construction industry comprises three sectors: companies engaged in the construction of buildings (TEÁOR 41), construction of other structures (TEÁOR 42) and specialised construction (TEÁOR 43) can be classified as construction enterprises (KSH 2021/a).

Construction is a key industry in the national economy. In 2021, the three sectors accounted for 6.3% of gross value added. The importance of the industry is indicated by the fact that 7% of registered enterprises are construction enterprises and 8% of the employed work in the industry. The economic downturn caused by the coronavirus epidemic also had an impact on the construction sector, with production volumes in 2020 down significantly compared to 2019, but up 13% in 2021 compared to the previous year (KSH 2021/a).

The construction sector is considered particularly risky because of the high levels of chain debt in the industry (Limpek-Kosztopulosz-Balogh 2016). For this reason, the failure of one business can set off a chain reaction between business partners.

The sample includes enterprises that are still active, employed at least 5 people in the survey period and have been in operation since at least 2013, i.e. for at least 9 years. This is because well-run, solvent firms often have a similar financial structure to bankrupt firms in the first few years of operation, which can be distorting in research (du Jardin 2010). A total of 1,293 viable businesses were included in the sample.

The sample included 384 companies that are currently in liquidation. These are companies whose annual accounts were available for at least 3 years before the bankruptcy.

The sample was randomly divided into a learning sample and a test sample. We created the models using the learning sample. This database contained a total of 1188 active and 339 bankrupt businesses. The test sample was used to check the performance of the models for independent enterprises. This sample included 105 surviving and 45 bankrupt businesses.

Methodology

In our study we used the method of logistic regression. The reason is that some estimates suggest that 95% of lender scorecards are based on this method (Nyitrai-Virág 2017). Logistic regression is a good way of relating explanatory variables to the probability of a binary response. The outcome variable is a dummy variable representing the solvent or insolvent categories in the case of a corporate bankruptcy forecast. The process involves fitting a logistic regression function to the observations using the maximum likelihood method. By weighting the independent variables, we obtain a 'Z' value, which expresses the probability of companies going bankrupt (Virág-Kristóf 2006).

The logistic regression formula can be written as follows:

$$Pr\ (fizetőképes) = \frac{e^{Z}}{1+e^{Z}} = \frac{e^{\beta_0+\Sigma(\beta_j Z_j)}}{1+e^{\beta_0+\Sigma(\beta_j Z_j)}} \text{ (Virág–Kristóf 2006)}$$

The big advantage of the method is that it does not require a normal distribution of variables and matching covariance matrices in the two classes. Another advantage is that the heteroskedasticity between variables distorts the results only slightly, so the variables under study do not need to be subjected to complex mathematical transformations (Rácz-Tóth 2021). When using this method, it is important to reduce the number of variables in a reasonable way. This is most often done by backward elimination. The method leaves out the less significant variables of the model one by one. After dropping a variable, it always recalculates the regression coefficients and p-values until only sufficiently significant variables remain. The final model is constructed taking collinearity, significance and classification accuracy together into account. After defining the regression parameters, an important step is to determine the cut-off value. This is the value of the dependent variable of the function against which companies can be classified either as bankrupt or solvent (Virág-Kristóf 2006).

The disadvantage of the procedure is that it is sensitive to outliers, which are common among financial indicators, especially in the case of bankrupt companies. For this reason, it is important to manage database outliers prior to research (Nyitrai 2017).

Development of a model for predicting bankruptcy of construction companies in Hungary

One of the aims of our study is to compare the impact of several possible treatments of outliers on classification accuracy.

Variables tested during model building

We have included in our model the financial indicators most commonly found in the academic literature. Table 1 shows the indicators examined and how they are calculated.

table 1: The indicators examined and the way they are calculated

	Indicator	Method of calculation
X_1	Liquidity ratio	Current assets / Current liabilities
X_2	Liquidity quick ratio	(Current assets - Inventories) / Current liabilities
X_3	Cash flow / Liabilities	(Profit after tax + Depreciation) / Liabilities
X_4	Cash flow / Current liabilities	(Profit after tax + Depreciation and amortisation) / Current liabilities
X_5	Capital adequacy	(Fixed assets + Inventories) / Equity
X_6	Current assets ratio	Current assets / Balance sheet total
X_7	Turnover of assets	Net sales / Balance sheet total
X_8	Turnover of inventories	Net sales / Inventories
X_9	Turnover of receivables	Receivables / Net sales revenue
X_{10}	Indebtedness	Liabilities / Balance sheet total
X_{11}	Equity ratio	Equity / Balance sheet total
X_{12}	Return on assets	Profit after tax / Equity
X_{13}	Creditworthiness	Liabilities / Equity
X_{14}	Return on sales	Profit after tax / Net sales revenue
X_{15}	Return on assets	Profit after tax / Balance sheet total
X_{16}	Receivables / Current liabilities	Receivables / Current liabilities
X_{17}	Net working capital ratio	(Current assets - Current liabilities) / Balance sheet total
X_{18}	Company size	Natural logarithm of assets
X_{19}	Ratio of fixed assets covered by long-term liabilities	Long-term liabilities / Fixed assets

Source: own editing

When logistic regression is used, multicollinearity between variables can be a problem, so this needs to be investigated before setting up the model (Kristóf 2005). To filter out multicollinearity between variables, the variance inflation factor (VIF) was used. The VIF value of a variable is obtained from the corresponding diagonal value of the inverse of the correlation matrix, the index that estimates the extent to which the variance of the regression coefficients increases due to multicollinearity (Vörösmarty-Dobos 2020). There is no consensus in the literature as to the value of the VIF above which multicollinearity exists. The most commonly used cut-off value is 5, so no variable with a VIF value greater than 5 was included in the final model.

The completed models

In total, three models were created during the research. A model was constructed without handling the outliers, followed by the construction of the models with the outliers. There is no consensus inTHE academic literature as to what constitutes outlier data. Often, statistical rules of thumb are used to define outliers, where values that fall outside a range of standard deviations are considered outliers. The problem with this approach, however, is that the variance of the variables changes after the outliers have been treated, so indicators that were not outliers before are now considered outliers when the new variance is used. The check with the newly defined standard deviations should be continued until no new outliers are found after the change in standard deviation (Nyitrai-Virág 2017). For this reason, in our study, we used the built-in SPSS function to determine the outliers for the sample.

There is also no single agreed method for dealing with outliers among researchers. There are two procedures that are most commonly used. Outlier data can be treated by replacing the value that is considered an outlier with the closest value that is no longer an outlier (Nyitrai-Virág 2017). Another method is to leave out observations with outliers from the sample (Nyitrai 2017). In this study, we use both methods and compare their effectiveness.

The data in Table 2 show that 1 year before the bankruptcy, the model obtained by deleting the outliers in the learning sample classified the items in the sample with the highest classification accuracy. The model accurately classified 85.73% of surviving models and 82.32% of bankrupt firms, achieving a classification accuracy of 85.21% for the total sample. It can be seen that as the time horizon expands, all three models become less predictive of bankruptcy. 3 years before bankruptcy, the model without treatment of outliers has an accuracy of 67.42%, the model with replacement of outliers has an accuracy of 58.7% and the model with deletion of outliers has an accuracy of 49.75% among insolvent firms.

table 2: Classification accuracy of the models produced

		1 year before bankruptcy			2 years before bankruptcy			3 years before bankruptcy		
		Correctly classified (number)	Misclassified, (number)	Accuracy, %	Correctly classified (number)	Misclassified, (number)	Accuracy, %	Correctly classified (number)	Misclassified, (number)	Accuracy, %
Without handling outliers	**survivor**	907	281	76,34	874	313	73,63	801	387	67,42
	bankrupt	261	78	76,99	245	95	72,06	260	79	76,70
	Total	1168	359	76,48	1119	408	73,29	1026	461	69,48
Replacement of outliers	**survivor**	967	221	81,40	942	245	79,36	911	277	76,68
	bankrupt	265	74	78,17	226	114	66,47	199	140	58,70
	Total	1232	295	80,68	1168	359	76,49	1110	417	72,69
Deleting outliers	**survivor**	937	156	85,73	901	192	82,43	800	293	73,19
	bankrupt	163	35	82,32	140	58	70,71	99	100	49,75
		1100	191	85,21	1041	250	80,64	899	392	69,64

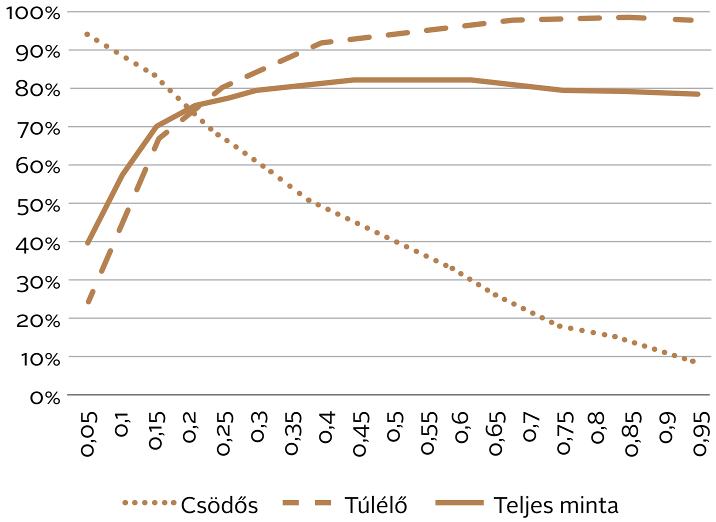
Source: own editing

Determining the cut-off values for the models

The sample did not contain solvent and insolvent businesses in equal proportions: of the companies surveyed, 1,188 are active and 339 are in liquidation. This reflects the fact that surviving businesses also make up a larger share of the economy than those that go bankrupt. When constructing bankruptcy prediction models, it is common practice to assign a cut-off value to the model that will result in the maximum classification accuracy. However, this procedure can only be used if the sample tested includes 50-50% of bankrupt and surviving businesses.

A good example of this is the model obtained without handling outliers. The maximum classification accuracy for the full sample would be obtained with a cut-off value of 0.55, in which case 86% of the companies in the sample would be correctly classified. The model correctly classifies 97% of the surviving businesses at this cut-off value. However, the proportion of correctly classified businesses among bankrupt enterprises is only 37.09% (Figure 1). When choosing the optimal cut-off values, it may not be appropriate to choose the value that maximizes the overall classification accuracy. If the primary objective is to minimise the overall error rate, a cut-off value is defined that classifies the active enterprises with the highest accuracy, but does not take into account the classification accuracy of insolvent enterprises with a lower proportion in the sample. This may lead to an increase in the proportion of first-order errors, so businesses that actually fail would be classified as survivors. As the primary users of bankruptcy prediction models are banks, this phenomenon can cause significant damage (Virág 2004). If a bank lends to a business that is assumed to be a survivor under the model, but is in fact about to go bankrupt, it may lose the capital it has lent and any interest income. Of course, the bank also suffers losses in the presence of second-order errors (i.e. when it misclassifies a firm that is actually surviving as bankrupt) because it loses profit, but higher rates of first-order errors cause greater disruption to bank operations (Zavgren 1985).

figure 1: Classification accuracy of the model without outliers with different cut-off values



Source: own editing

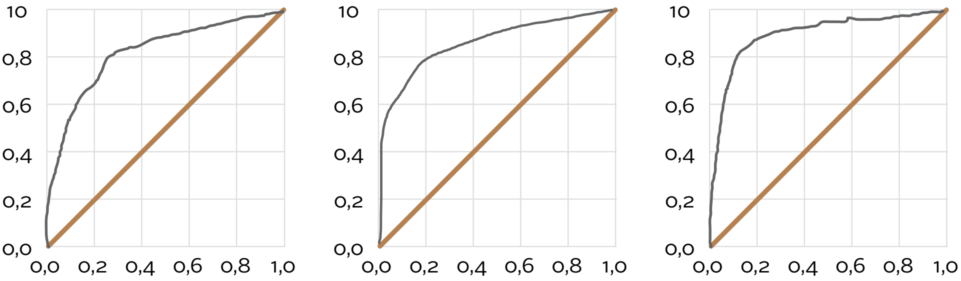
Nor can we rely on choosing a cut-off value that minimises first-order errors, since in this case the proportion of second-order errors may increase. For the model obtained without replacing outliers, the highest classification accuracy of 95.55% is achieved for bankrupt companies with a cut-off value of 0.05. In this case, however, the hit rate for surviving firms is only 25.34%.

When determining the optimal cut-off, a value should therefore be chosen that results in a sufficiently high classification accuracy, while keeping first-order errors low. For banks, this choice may depend on the level of default risk that the financial institution is willing to take in its loan portfolio.

Comparison of the completed models

The models were first compared on the basis of the cumulative rating accuracy curve (ROC). The ROC curve shows how well the classification values determined by the model correspond to the original classification at different cut-off values. The value of the curve should be compared to a 45° line, the more the curve diverges from the line, i.e. the larger the area under the curve, the more accurate the model (Virág-Kristóf 2009). The area under the ROC curve is 83.5% for the model without outliers, but higher at 85.2% for the model with the replacement of outliers. For the model with outliers removed, the area under the ROC is 90.8%, so the model classifies a randomly selected enterprise from the sample with an accuracy of 90.8% (Figure 2). Based on the ROC curves, the model obtained by removing the outliers seems to be the most appropriate.

figure 2: ROC curves of the models produced



Without handling the outliers

By replacing the outlier data

By deleting the outlier data

Source: own editing

For bankruptcy models, it is important that they not only work on the training sample with high classification accuracy, but are also suitable for predicting the bankruptcy of independent businesses, so the resulting models were also tested on an independent test sample. This sample includes 150 construction firms, of which 45 are in liquidation and 105 are operational and solvent.

Table 3 presents the classification accuracy of the obtained bankruptcy prediction models for the learning and test samples. The table shows that for the test sample, the model without handling outliers was the model with the lowest classification accuracy.

table 3: Accuracy of the built bankruptcy prediction models for the test and learning samples

		Classification of surviving businesses, %	Classification of bankrupt businesses, %	Accuracy of the total sample, %
Without handling the outliers	Learning sample	76,34	76,99	76,44
	Test sample	81,95	71,11	80,00
Replacing outliers	Learning sample	81,40	78,17	80,68
	Test sample	88,78	73,33	86,00
Deleting outliers	Learning sample	85,73	82,32	85,21
	Test sample	88,78	68,89	85,20

Source: own editing

The classification accuracy of the model obtained by excluding businesses with outlier data from the sample outperforms the predictive ability of the other two models on the learning sample. However, the test sample had lower accuracy for bankrupt businesses than the other two models, so the proportion of first-order errors is higher. This is because the outliers are mostly found in bankrupt businesses, with 42% of the insolvent businesses in the original learning sample having some kind of outlier and being excluded from the sample. The exclusion of these businesses from the sample caused a strong loss of information and data, which was not used in the model, and therefore, although the model classified the insolvent firms in the learning sample with an outstanding accuracy of 82.32%, it was much less efficient in the test sample: it classified the insolvent part of the test sample with a 68.89% hit rate for bankrupt firms.

All this confirms the claim that when building a bankruptcy prediction model based on logistic regression, it is important to prepare the data thoroughly and to handle outliers. The best model for predicting bankruptcy of construction firms in Hungary based on logistic regression is one that is constructed by replacing the outliers.

Summary

Financial institutions most often use the logistic regression method to produce credit scorecards. In our study, we used logistic regression to construct three bankruptcy prediction models to investigate how much the accuracy of the outlier-sensitive method is improved by treating the sample outliers before modelling, and whether the method of deleting outliers or replacing outliers with the closest values that are no longer outliers can achieve higher classification accuracy. In our study, we also examined the basis of the choice of a cut-off point for a sample that does not contain solvent and insolvent firms in equal proportions.

The completed models showed that the handling of outliers significantly improves the predictive ability of the models, and that replacing outliers with the closest non-outlier value proved to be more effective for handling outliers.

When determining cut-off values, the value that gives the highest classification accuracy is not the most appropriate, as it takes into account the hit rate of the surviving businesses included in the sample with a higher proportion, which can lead to a high first-order error rate. This is detrimental to the primary users of bankruptcy prediction models, i.e. banks, and it is therefore advisable to define a cut-off value that results in a sufficiently high classification accuracy while keeping the first-order errors low. The optimisation of this value may depend on the degree of credit risk taken by a given financial institution in its portfolio of loans. An interesting future research opportunity could be to examine the factors that banks use to determine the cut-off value in their portfolios. ■

References

1. 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, pp. 1-42.
2. Cziglerné Erb E. (2020). A reziduálisjövedelem-modell újbóli megjelenése a vállalatok és a beruházási projektek értékelésében, Pénzügyi Szemle, 3, pp. 430-442.
3. Doma, T., Kozma, N. (2022). A Covid 19-járvány hatása a magyarországi bankok működési kockázati veszteségeire, Gazdaság és Pénzügy, 9, pp 356-375. https://doi.org/10.33926/GP.2022.4.3
4. du Jardin, P. (2010). Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy, Neurocomputing, 73, pp. 2047-2060. https://doi.org/10.1016/j.neucom.2009.11.034
5. Kiss G., Kosztopulosz A., Szládek D. (2021). A magánkiadások hatása a hazai egészségügyi diagnosztikai szolgáltatók pénzügyi helyzetére, Köz-Gazdaság, 16, pp. 115-132. https://doi.org/10.14267/retp2021.04.08
6. Kristóf, T. (2005). A csődelőrejelzés sokváltozós statisztikai módszerei és empirikus vizsgálata, Statisztikai Szemle, 9, pp. 841-863.
7. Kristóf, T., Virág, M. (2019). A csődelőrejelzés fejlődéstörténete Magyarországon, Vezetéstudomány, 12, pp. 62-73.
8. KSH (2021). Helyzetkép, 2021 – Építőipar, Központi Statisztikai Hivatal, Budapest
9. Limpek, Á., Kosztopulosz, A., Balogh P. (2016). Késedelmes fizetés, tartozási láncok-A Dél-Alföld régió kis-és középvállalkozásainak pénzügyi kultúrája, Statisztikai Szemle, 94, pp. 365-387. https://doi.org/10.20311/stat2016.04.hu0365
10. Mikolasek, A. (2018). A hitelkockázati modellek alkalmazásának néhány problémája, Gazdaság és Pénzügy, 3, pp. 248-257.

11. Nyitrai, T. (2015). Hazai vállalkozások csődjének előrejelzése egy, két, illetve három évvel korábbi pénzügyi beszámolók adatai alapján, Vezetéstudomány, 5, pp. 55-65. ohttps://doi.org/10.14267/veztud.2015.05.06
12. Nyitrai, T. (2017). Stock és flow típusú számviteli adatok alkalmazása a csődelőrejelző modellekben, Vezetéstudomány, 48, pp. 68-77. https://doi.org/10.14267/veztud.2017.09.07
13. Nyitrai, T., Virág M. (2017). A pénzügyi mutatók időbeli tendenciájának figyelembevétele logisztikus regresszióra épülő csődelőrejelző modellekben, Statisztikai Szemle,1, pp. 5-28. https://doi.org/10.20311/stat2017.01.hu0005
14. Pálinkó, É., Svoób Á. (2016). A vállalati csőd bekövetkezésének fő okai és a csődhöz vezető folyamat, Pénzügyi Szemle, 4, pp. 528-543.
15. Piller, Zs., (2013). A fizetésképtelenségi eljárások illeszkedési módjai nemzetközi összehasonlításban, Pénzügyi Szemle, 2, pp. 151-164.
16. Rácz, T. A., Tóth, B. (2021). A hazai önkormányzatok pénzügyi zavarai az adósságkonszolidáció és az önkormányzati rendszer reorganizációjának tükrében, Pénzügyi Szemle, 1, pp. 88-108. https://doi.org/10.35551/psz_2021_1_5
17. Virág M, (2004). A csődmodellek jellegzetességei és története, Vezetéstudomány, 10, pp. 24-32.
18. Virág, M., Kristóf T. (2006). Iparági rátákon alapuló csődelőrejelzés sokváltozós statisztikai módszerekkel, Vezetéstudomány, 37, pp. 25-35. https://doi.org/10.14267/veztud.2006.01.04
19. Virág, M., Kristóf, T. (2009). Többdimenziós skálázás a csődmodellezésben, Vezetéstudomány, pp. 50-29. https://doi.org/10.14267/veztud.2009.01.05
20. Vörösmarty, Gy., Dobos, I. (2020). A vállalatméret hatása a zöldbeszerzési gyakorlatra, Statisztikai Szemle, 4, pp. 301-323. https://doi.org/10.20311/stat2020.4.hu0301
21. Zavgren, C. V. (1985). Assessing the vulnerability to failure of Americn industrial firms: a logistic analysis, Journal of Business Finance & Accounting, 12(1), pp. 19-45.
22. Zéman, Z., Hegedűs, Sz., Molnár, P. (2018). Az önkormányzati vállalkozások hitelképességének vizsgálata credit scoring módszerrel, Pénzügyi Szemle, 2, pp. 182-200.
23. Zmijewski, M. E. (1984). Methodological Issues Related tot he Estimation of Financial Distress Predection Models, Journal of Accounting Research, Vol. 22, pp. 59-82.