DOI: https://doi.org/10.17221/374/2015-AGRICECON

doi: 10.17221/374/2015-AGRICECON

Predicting financial distress of agriculture companies in EU

Václav KLEPAC*, David HAMPEL

Department of Statistics and Operation Analysis, Mendel University, Brno, Czech Republic

*Corresponding author: xklepac@node.mendelu.cz

Klepac V., Hampel D. (2017): **Predicting financial distress of agriculture companies in EU**. Agric. Econ. – Czech, 63: 347–355

Abstract: The objective of this paper is the prediction of financial distress (default of payment or insolvency) of 250 agriculture business companies in the EU from which 62 companies defaulted in 2014 with respect to lag of the used attributes. From many types of classification models, there was chosen the Logistic regression, the Support vector machines method with the RBF ANOVA kernel, the Decision Trees and the Adaptive Boosting based on the decision trees to acquire the best results. From the results, it is obvious that with the increasing distance to the bankruptcy, there decreases the average accuracy of the financial distress prediction and there is a greater difference between the active and distressed companies in terms of liquidity, rentability and debt ratios. The Decision trees and Adaptive Boosting offer a better accuracy for the distress prediction than the SVM and logit methods, what is comparable to the previous studies. From the total of 15 accounting variables, there were constructed classification trees by the Decision Trees with the inner feature selection method for the better visualization, what reduces the full data set only to 1 or 2 attributes: ROA and Long-term Debt to Total Assets Ratio in 2011, ROA and Current Ratio in 2012, ROA in 2013 for the discrimination of the distressed companies.

Keywords: agribusiness, classification, constrains, decision tree, default, nonlinear techniques, support vector machines

If the financial health of a company is weakened, the company deals with the financial distress which can develop into a financial crisis and finish as a default. The main instrument used to inform the people concerned about the standing of the company is the financial analysis based on accounting. The financial analysis, besides its basic functions, aims to identify scenarios leading to the financial health of the company and quality decisions of the management of the agribusiness companies. Based on the previous empirical studies, we are able to build prediction models with a high accuracy and effectiveness in the financial distress prediction.

The article deals with financial risks and the prediction of financial distress which is heavily connected to the detection of financial problems like the indebtness or liquidity issues across time. In the previous years, a large number of authors tried to improve the techniques for the classification of active and distressed (typically bankrupted) companies and the prediction possibility of these two states. This report is based on the classification of companies according to the dawning default or default management which represent standings with a high level of indebtedness or de facto a financial default.

Generally, the models of the financial distress or the bankruptcy prediction could be classified into three major groups: statistical techniques like the logistic regression (logit), probit or the multivariate discrimination analysis (MDA) models, the artificial intelligence or the data mining techniques, like the support vector machines (SVM), the neural networks (NN), decision trees (DT) and the theoretical models (expert evaluation, market risk models etc.).

One of the first authors who used basic statistical techniques in financial distress, e.g. the bankruptcy prediction were Beaver (1966) with the univariate analysis and Altman (1968) who used the MDA, in that he computed an individual firm's discriminant score using a set of financial and economic ratios. Probably due to the huge demand coming from the financial sector at the beginning of 1980s, more ad-

Supported by the Czech Science Foundation (Grant No. GA13-25897S).

doi: 10.17221/374/2015-AGRICECON

vanced estimation methods, such as the Ohlson's logit (1980) and Zmijewski's probit method (1984), were employed. Compared to the MDA, the logit model was easier to understand since the logistic score, taking a value between 0 and 1, was interpretable in a probabilistic way.

In 1990, the NNs technique was used in the field of corporate bankruptcy prediction. However, the MDA and logit analyses have remained as popular tools for the financial distress prediction (see for example Vavřina 2013), unless they have some issues associated to the normally distributed independent variables, sensitivity to multicollinearity, equal variance-covariance matrices for distresses and non-distressed companies, see Balcaen and Ooghe (2006).

The NNs as the data mining technique dominates the literature on the business failure in 1990s, and is still the most frequently used as the performance benchmark in the recent bankruptcy prediction studies. Among other prominent algorithms in the data mining used for the bankruptcy prediction are the DT (there exists a large number of DT variations like CART with Breiman (1993) and C4.5 algorithm by Quinlan (1993)), support vector machine like in Vapnik (1995), k-nearest neighbours (kNN)). Especially, the non-parametric prediction method known as the decision tree (DT) or the recursive partitioning has been used in an attempt to bypass the above mentioned assumptions in MDA and logit by Frydman et al. (1985). The recent studies used the DT in the financial distress prediction like in Gepp et al. (2010), or in Huarng et al. (2005).

Li et al. (2010) demonstrated the applicability of the DT model in the area of the business failure prediction and compared the performance power with four other classification methods including the MDA, logit, kNN, and SVM. Many recent studies used the ensemble methods (algorithms that combine multiple models to obtain a better accuracy) to enhance the prediction accuracy of the DT models, whereas Lin and McClean (2001) predicted the corporate failure by using the MDA, logit, DT, NN. They used 106 failed and 690 non-failed companies for their training set and 48 failed and 289 non-failed companies for their validating set. The authors showed that the DT and NN models performed better than the MDA or logit model.

The Adaptive Boosting algorithm developed by Freund and Schapire (1996) is one of the most important ensemble methods because it has solid theoretical evidence, an accurate prediction performance, a great simplicity, and wide and successful applications.

Most researchers used quantitative techniques to compare the prediction performance with other techniques like the MDA, SVM, NN, DT and logistic regression for a specific data set like Min and Lee (2005), who was testing the accuracy of classifiers on the set of 1888 companies, using the classification by the means of the NN, MDA, logit and SVM with the RBF kernel, which had the accuracy around 83% in opposite to the other models with the accuracy around 80%.

Min et al. (2006) assessed the SVM, NN and logistic regression on the case of the prediction of bankruptcy and assessed its predictive performance. This paper deals with the real data of 614 Korean production companies including 307 bankrupted ones in the time period 1999–2002. The classification accuracy 80.3% was achieved for the SVM model with 32 variables. For the logistic regression, only the accuracy of 68% was achieved – the same as for the NN.

Ding et al. (2008) used data of 250 publicly traded Chinese companies (11 variables) with the SVM and other methods (NN, logit and MDA). The results point out that the SVM with the RBF kernel seems to be the most advantageous method; its accuracy of classification was around 95.2% for training and 83.2% for testing the data set. The use of the neural networks led the authors to the worst results – 76% on the testing data set.

The authors of these studies mostly used the datasets of hundreds or thousands medium-sized or large companies and tens of variables. Aziz and Dar (2006) have summarized a large portion of studies which accommodates the use of different approaches and states about different precision in the bankruptcy prediction. At present, predictions of bankruptcy are not limited to mere estimations based on elementary methods, but the so-called ensembles are created – i.e. the collections of models the output of which are the average data from a higher number of elementary models which show a higher accuracy than the elementary approaches. All of this is sufficient for the management which is able to find out the variables beneficial for the decision-making.

The main aim of this paper is to test the prediction accuracy of different classification methods with a heavier concentration on the decision trees methods for different lag of data connected to the agribusiness domain. Practically, we want to find out if it is possible to predict bankruptcy 1–3 years ahead with a solid accuracy and to obtain comparable results to the previous studies based on the data of non-financial or manufacturing companies.

doi: 10.17221/374/2015-AGRICECON

In the previous studies, the authors compared the accuracy of classification with the time interval left to bankruptcy. The results show that the accuracy decreases. Unfortunately, there are still not enough studies available for the practical decision-making focused on the data of European small and medium-size agribusiness companies and a closer focus on the decision-making rules for setting and concretization of the decision-making process, whether the company is endangered by the default and should fight back the crisis. The previous studies often present results which are of a rather technical character as for the setting of classification inputs, instead of a higher concentration on the practical and easily understandable conclusions. The usage of a method such as the DT is extremely fitting for this purpose as it reduces the number of variables (and in addition to it, it has a high illustrative potential).

In our research, we used the DT to develop the EU agricultural companies prediction model and we extended the boundary of the literature reviews into the area of classification using the DT model because of the limited studies of financial distress prediction using the DT model in comparison to the SVM and more advanced DT models like the ADA boosting (also based on the DT classifiers). The merit of this report is in different classification marks than the default for the classification problem set in this way, which can be beneficial for a further analysis in this field. The results will show whether the data profile of companies is characteristic enough for the financial distress or whether the data profile bears a higher resemblance with the active companies.

Our results enhance the literature with at least one main contribution concerning research studies about the manufacturing business field: assessing of the best classifier types which are useful for the prediction of bankruptcy and inference of Error I and II. Besides the merit for the management of companies involved in the same branch of business in the form of providing information, the merit for banking officers in relation to crediting subjects can be mentioned as well.

MATERIAL AND METHODS

Data and descriptive statistics

The paper is aimed at assessing the bankruptcy prediction precision of the selected classifiers for the agricultural entities based on their annual accounting data obtained from the Amadeus database by the Bureau Van Dijk. In Table 1, the variables observed during the time period from 2009 to 2013 are presented together with the attribute units. The database provides information about the selected companies (including legal form, status of the company and region), and a full description of the accounting data with the selection by the sector or geographic area. The data consist of 250 EU 28 companies, from which 62 business entities went default or fell into insolvency proceedings in 2014. Specifically, we use the data from crop and animal production, hunting and the related service activities, Forestry and logging, fishing and aquaculture companies (based on the NACE Rev. 2 classification and companies with available data). Finally, we have arrived to four different data sets to evaluate how the proportion between the training, validation and test setting changes the accuracy of the classification, see below. A categorical variable Status which indicates whether the company was active (A) or in the financial distress (D) was added to this database. The choice of the likely variables is

Table 1. Financial variables for risk assessment

Variable	
Solvency ratio (%)	Debt/Total assets
EBIT (ths. EUR)	Debt/Equity
ROA (%)	Long-term debt/Total assets
Collection period (days)	Long-term debt/Equity
Credit period (days)	Assets/Equity
Current ratio = Current assets/Current liabilities	Current assets/Equity
Liquidity ratio = Liquid assets/short-term liabilities	Working capital/Total debt
Working capital/Total assets	Current assets/Total assets
Status (categorical value)	

Source: Own processing

doi: 10.17221/374/2015-AGRICECON

bound to the possible data acquisition and the previous empirical studies. We use the usual variables for capturing rentability, liquidity and leverage characteristics of the entity altogether with the selected basic univariate variables. The data analysis processing was realized in the computational system R 3.1.1.

Classification methods used in the study for the distress prediction

Technical details about the well-known logistic regression method can be found in Freedman (2009). The SVM was presented by Vapnik (1995) as a new class of machine learning techniques. In the SVM, the original input space is mapped into a high-dimensional dot product space called a feature space, and in the feature space the optimal hyperplane is determined to maximize the generalization ability of the classifier. The optimal hyperplane is found by exploiting the optimization theory, and respecting insights provided by the statistical learning theory. In the empirical studies, there are often used different kernels like the linear, RBF, ANOVA RBF or Hyperbolic tangent.

Decision trees based on the CART algorithms described in Breiman (1993) are produced by the algorithms that identify various ways of splitting the data set into branch-like segments: the resulting models can be represented as binary trees. This method does not require any statistical assumption concerning the data in the training sample. The data is separated, and then this process is applied separately to each sub-group, and so on recursively until the subgroups either reach a minimum size for its settings or until no improvement can be made.

The second stage of the procedure consists of using the cross-validation to trim back the full tree. The objective of the split is to reduce the impurity of a set by creating subsets that have a greater proportion of members from one of the groups than the original set. The algorithm ends when it achieves the aim of maximizing the homogeneity of the response variable in each of the obtained sub-groups.

Freund and Schapire (1996) formulated the Adaptive Boosting (AdaBoost) algorithm. It can be used by combining other learning algorithms to make an improved learning algorithm from the basic ones. The Adaptive Boosting algorithm first sets the initial distribution on the training set and then iterates it until the stopping criterion is reached by using the adaptive weights. After a classifier is built, the weight of each training sample is changed according to the classification given by the classifier. The next classifier is then built using the reweighted training sample. Finally, once the training process has been completed, the single classifiers are combined into a final highly accurate classifier based on the training set. Because boosting maintains a weight for each instance, the higher the weight, the more the instance influences the learned classifier.

Research steps

After the initial selection of variables, we proceed in the following steps, like Ding et al. (2008), where the authors compared the prediction accuracy of several classifiers. At first, we divide the data set into the particular training, validation and testing (prediction) groups. In this case, we want to test the accuracy for random samples with 75/15/15%, 20/30/50%, 33/33/34% and 20/57/23% as for the training/validation/test partitioning. The first three choices depend on the random guess; the fourth is the one in selection on the basis of testing.

We proceed with training and validation on the data and prediction whereas the target labels are the Distress or Activity in year 2014. The use of different classifiers which all lead to the error evaluation: the SVM with ANOVA RBF, decision trees and the Adaptive Boosting working with five hundred decision trees, see below for details. The SVM with the ANOVA RBF kernel was chosen after the evaluation of different kernels (RBF and linear), hence it offers the best results among them. Technical or mathematical description about these methods is covered in Vapnik (1995).

The accuracy diagnostics deals with the Area Under Curve (AUC) values and the average accuracy evaluation for all chosen classifiers. The obtained Empirical Error I and II types values are established for comparison with other empirical studies. Error I evaluates the number of active companies which were classified as bankrupted. In contrast, Error II type shows how

Table 2. Confusion matrix for testing the subset

		Current category	
		T (active)	F (distressed)
Predicted category	T	True positives (TP)	False Positives (FP)
	F	False Negatives (FN)	True Negatives (TN)

Source: Own processing

doi: 10.17221/374/2015-AGRICECON

many bakrupted companies were wrongly labeled as the active ones. The results of the classification on prediction are evaluated with the aid of the confusion matrix like in Table 2.

According to classes in Table 2, we can calculate the overall accuracy and the more specific misclassification rates:

– Empirical average classification accuracy as the average rate of accuracy bound both classification labels and Type I and II errors are used for more specific accuracies evaluation

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

– Empirical Error I which evaluates the number of true positives which were classified as true negatives

$$Type\ I\ Error = \frac{FN}{TP + FN}$$

– In contrast, Empirical Error II that shows how many of true negatives were labeled as false positives

$$Type\ II\ Error = \frac{FP}{FP + TN}$$

Based on these calculated data, the analy sis of the predictional ability of the classification methods via the construction of the so-called *Receiver Operator Characteristics (ROC)* curves can be carried out. It is a tool for the evaluation and optimization of the binary classification system for which the value Area Under Curve (AUC) is the most common index describing the ROC curve with the usual value between 0.5 and 1:

– AUC from 0.50 to 0.75 – eligible classification
– AUC from 0.75 to 0.92 – good classification
– AUC from 0.92 to 0.97 – very good classification
– AUC from 0.97 to 1.00 – perfect classification.

Afterwards, we analyse the inducted decision tree across the time to get a better recognition about the levels of the selected variables. For this part, we choose the best DT model according to different data partitioning results for the testing part of data set, because it offers an effective way how to visualize the results.

RESULTS

Basic characteristics of the data are summarized in Table 3. The form of median is used which is able to describe the analysed set of data properly. The comparison of companies in default and the active ones is very interesting in the years before bankruptcy. Marked changes of observed variables can be seen clearly between 2012 and 2013. In many cases, the companies show very low levels of the liquidity ratios – the Current Ratio, which is two times lower than for the active companies. It concerns also the

Table 3. Median values for all variables (from 2011 to 2013 for the active, distressed and mixed data set)

Variable/Period	Distressed			Active			Mixed		
	2011	2012	2013	2011	2012	2013	2011	2012	2013
Solvency ratio (%)	14.2	11.2	2.51	51.87	52.43	54.17	42.45	44.2	44.91
EBIT (ths. EUR)	18.51	–1.3	–36	64.92	63.09	58.62	56.95	46.4	38.83
ROA (%)	0.13	–1.5	–8.2	5.61	7.33	6.46	3.68	3.6	3.48
Collection period (days)	116.94	124	103	41.56	42.02	41.47	51.12	48.98	51.32
Credit period (days)	103.47	107	113	31.02	27.62	24.71	39.97	38.62	34.02
Current ratio	0.96	0.93	0.78	1.71	1.82	1.93	1.42	1.43	1.51
Liquidity ratio	0.44	0.45	0.38	1.2	1.25	1.35	0.86	0.93	1.4
Working capital/Total assets	0.15	0.16	0.09	0.17	0.17	0.16	0.155	0.17	0.15
Current assets/Total assets	0.44	0.46	0.44	0.55	0.51	0.53	0.51	0.49	0.52
Total debt/Total assets	0.86	0.89	0.98	0.48	0.48	0.46	0.58	0.55	0.55
Total debt/Equity	3.12	2.85	1.2	0.85	0.91	0.85	1.13	1.6	0.87
Long debt/Total assets	0.25	0.23	0.26	0.06	0.07	0.07	0.08	0.1	0.08
Long debt/Equity	0.63	0.31	0.04	0.09	0.14	0.13	0.13	0.17	0.11
Total assets/Equity	4.12	3.85	2.2	1.85	1.91	1.85	2.13	2.6	1.87
Current assets/Equity	2.3	1.59	0.39	0.98	0.99	0.96	1.5	1.1	0.91
Working capital/Total debt	0.18	0.17	0.09	0.37	0.44	0.48	0.31	0.36	0.36

Source: author's calculation and research processing

doi: 10.17221/374/2015-AGRICECON

performance efficiency measured via the ROA, which even reaches the negative level (–8.2% in 2013) for the default companies. Besides, an apparent financial distress can be illustrated by the negative numbers of the EBIT which has been manifested since 2012.

The companies which will default in the future suffer from the apparent problems with indebtedness – the ratio Total Debt/Assets as the leverage indicator is many times higher for the distressed companies than for the active ones, besides, the more the negative change of the company's standing is, the higher the rate of indebtedness is. The ratio of the contribution of assets from the own sources of company, i.e. the Equity Multiplier, is also apparent. It is much higher than for the companies active more than 1 year before distress. It can be claimed that the companies nearing distress sell out both their short-term and long-term assets to settle the short-term liabilities and their working capital decreases at the same time. A longer time of the outstanding debts maturity and the liabilities settlement over a longer time period are more evident than for the active companies. It corresponds with the final company standing, when the companies either default or enter the state of the insolvency management – it is the logical outcome of an unfavourable company's standing.

After the initial statistical evaluation models are constructed, their classification accuracy is carried out for various methods. In Table 4, the total results of the accuracy of classification together with errors of type I and II can be seen. They are stated via the value AUC and Accuracy for the testing sample, which were based on the data of the companies in each year. The previous testing (using the linear and RBF kernel SVM) leads to results which were not satisfactory; the change to kernels shown below provides results of a higher quality or relevance in case of these data.

Table 4 illustrates that the longer time period till distress, the lower the prediction ability of the models is, i.e. the financial indicators or the real situation for the previous years do not reflect the resulting financial difficulties on the satisfactory level, or the companies have not faced these problems yet. The highest prediction ability can be observed no sooner than in 2013, i.e. one year before the bankruptcy. The evaluation via the AUC value is also problematic. Models for the lagged data with higher lags provide only a low percentage of the appropriately classified bankruptcies; technically the differences between the active companies and those which will be bankrupt in the future are not so obvious.

The results show that the highest values of the average accuracy of classification can be observed mostly in 2013. It is appropriate to add that the models are able to notice correct labels for active companies,

Table 4. Evaluation of classifiers (ordering from I to IV describes data partitioning, from 75/15/15%, 20/30/50%, 33/33/34% to 20/57/23%)

Classifier/Period	Total accuracy (%)			AUC (%)			Type I Error (%)			Type II Error (%)		
	2011	2012	2013	2011	2012	2012	2011	2012	2013	2011	2012	2013
Decision trees (I)	78.94	89.4	94.73	0.83	0.87	0.91	19.2	7.7	0	25	16.67	16.67
ANOVA RBF SVM (I)	78.37	81.5	89.47	0.88	0.93	0.92	11.53	3.84	3.84	45	50	25
Logit (I)	81.08	78.37	92.1	0.84	0.92	0.97	7.69	7.69	7.69	45	54.54	8.33
Adaptive Boosting (I)	89.4	89.47	94.74	0.91	0.94	0.98	7.69	3.84	7.1	17	25	0
Decision trees (II)	65.11	87.2	91.8	0.72	0.91	0.92	21.81	0	7.27	58	35.49	9.68
ANOVA RBF SVM (II)	77.91	80.72	88.37	0.84	0.9	0.92	9.9	1.92	0	45	43.38	32.22
Logit (II)	80.23	87.8	88.37	0.75	0.93	0.86	1.82	11.6	1.82	52	15.38	29.3
Adaptive Boosting (II)	86.04	86.04	94.19	0.93	0.96	0.97	3.64	0	3.64	32	38.71	9.68
Decision trees (III)	72	72.8	92.8	0.75	0.5	0.93	23.7	0	6.59	4.1	1	8.82
ANOVA RBF SVM (III)	76.8	79.2	84	78.86	81.7	88.91	9.89	11	9.89	59	47.08	32.35
Logit (III)	76.8	78.4	80.3	0.76	0.79	0.83	17.58	13.2	19.38	38	44.11	20.58
Adaptive Boosting (III)	82.4	80.8	94.4	0.84	0.86	0.97	4.39	5.49	2.19	53	55.88	14.71
Decision trees (IV)	72.4	67.24	93.1	0.75	0.5	0.94	20.51	0	7.69	42	1	5.26
ANOVA RBF SVM (IV)	68.97	75.86	82.76	0.74	0.80	0.87	15.38	12.8	10.26	63	47.4	31.58
Logit (IV)	79.31	72.41	72.41	0.77	0.76	0.77	12.82	10.3	25.64	37	63.16	31.58
Adaptive Boosting (IV)	81.03	77.59	93.1	0.86	0.86	0.96	2.56	2.56	5.13	53	63.15	10.53

Source: author's calculation and research processing

doi: 10.17221/374/2015-AGRICECON

however, the longer the time period to bankruptcy is, the worse the results are, or the results are inconsistent for the bankrupted companies which have been active till now.

From the side of Error I type and Error II type as the key metrics, we compare the classifier and arive to the conclusion, that the best results are provided by the Adaptive Boosting and Decision Trees, from the point of type II errors. The values of the average accuracy are also important to some degree, however, it is not essential, because we prefer the metrics capturing the missclasification rates for both Error types.

Decision trees

After having gone through the previous stages, the empirical accuracies of the prediction models were found out. The conclusion of this comparison is that the Decision Trees, the Random Forest and the ADA Boost achieve the highest accuracy, especially when compared with models based on the SVM.

Figures 1–3 show classification trees for the method, which were created using the chosen attributes by the means of method for input data from 2011, 2012 and 2013 where the number for the key error of the 2nd level was the lowest one. This method, in addition to its high prediction accuracy, provides an easily understandable illustration of the decision-making rules.

Mark A means active companies and D means distressed companies in 2014. The total of the quotients of percentages in the lowest level means the quotient marked by the model; it is always 100% in total. The quotient of values marked above the quotient in percentages shows the assignment of the company into this subset within the tree. Looking from above downwards, a bifurcation can be seen, which gradually separates the set of data according to the appraising criteria into elements.

The data shown by companies in 2011 lead to the conclusion that the primary criterion for the company standing in 2014 is the ROA and LD/TA (Figure 1). According to the obtained values, the most advantageous method is to observe the number ROA lower than 4.1 and LD/TA lower than 0.1 three years till bankruptcy. However, the model suffers from a quite high number of misclassifications, there is the 43% ratio of active companies even in the case of the LD/TA value higher than 0.1. However, it must be said, that the results are specific for this choice of data and settings. Models, which had the highest accuracy in the prediction part, resp. the lowest number of misclassifications, are represented. In some cases of less accurate models, the classification was performed via the Solvency Ratio, which showed a quite high explanatory capability.

From the perspective of two-year distance to default, the ROA is one of the main separating criteria (in this case the primary one in the hierarchy), see Figure 2. After the evaluation of the ROA under the level 1.1%, it is possible to determine for 41% of companies if the company will have defaulted in the future. The next part evaluates the value of the Current Ratio, i.e. the liquidity indicator under 1.2, which together with the ROA provides an exact view of the risk for the company.

In the last year before bankruptcy, the signals of the default-to-be companies are simpler; the value ROA under 0.29% indicates a more serious financial distress of the company showing the future default. It cannot be concluded that the variable ROA is the only one

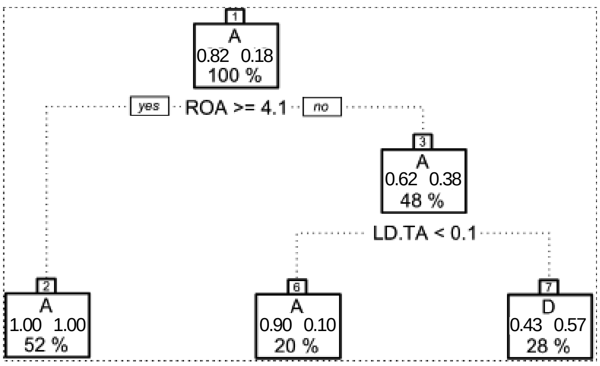

Figure 1. Decision trees (with data from 2011 and bankruptcy in 2014, partitioning 20/30/50)

Source: author's calculation and research processing

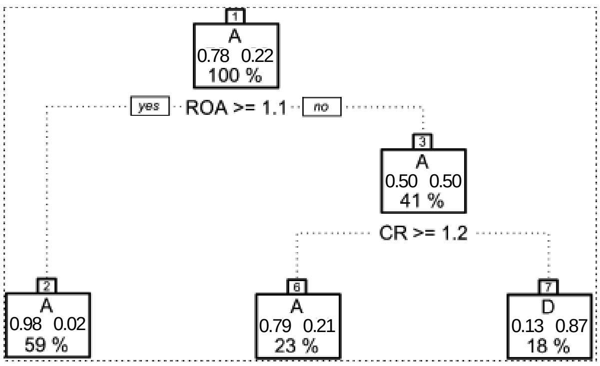

Figure 2. Decision trees (with data from 2012 and distress in 2014, partitioning 33/33/34)

Source: author's calculation and research processing

doi: 10.17221/374/2015-AGRICECON

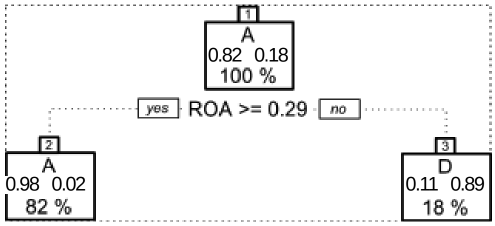

Figure 3. Decision trees (with data from 2013 and distress in 2014, partitioning 20/57/23)

Source: author's calculation and research processing

indicating default in the near future. However, from the point of view of the evaluation DT, the ROA was evaluated to be the variable of the utmost importance with very high discrimination ability.

It is clearly shown that modelling leads to very similar results as in the introductory Table 1, however, there is a difference – it is possible to achieve the exact value of the evaluating criterion and the relations between variables. It is apparent that the unfavourable theoretical value of one variable does not have to lead to bankruptcy. It is important to mention that based on our data, it can be claimed that the running time to default lowers the complexity of the diagram which detects default, the companies are more distinctly profiled which is shown in a more accurate classification.

DISCUSSION AND CONCLUSION

In the previous paragraphs, we compared and analysed the results of the financial distress prediction models with a heavier concentration on the Decision Trees method, with a various model setting applied to the agribusiness companies. Such results suggest important implications.

We literally wanted to find out if it is possible to predict financial distress (default or insolvency proceeding state) 1–3 years ahead with a solid accuracy. The tested financial data acquired from the Amadeus database consisted of accounting variables from 2011 to 2013. Our analysis used the data of 188 active manufacturing small and medium-sized companies and 62 companies which reported the default of payment or insolvency proceedings during 2014.

The obtained average accuracies are on the same levels as in the comparable studies for one year before bankruptcy of Niknya et al. (2013), but in contrast to the previous studies, the SVM based classifiers propose the use of different kernels to the linear or RBF that were used in Min and Lee (2005). The variant comparison with different models led us to the conclusion that the SVM classifier based on the ANOVA RBF kernel performs well for capturing the total accuracy, especially for the formation of 1 year ahead predictions. However, this does not apply for the evaluation of Type I and II errors – the models have significant difficulties in capturing the real bankruptcy or distressed profile, which holds true especially for the active companies. For a longer period before bankruptcy, the models are not efficient enough to predict the bankruptcy – the active companies are assigned with bankruptcy labels.

For this sake and to treat Error II types, it is better to utilize the Decision Trees or to the ensemble Adaptive Boosting classifiers in comparison to the SVM or logit method. Therefore, we used the AdaBoosting to overcome the sensitivity problem of the DT model and to make the DT approach more replicable, as discussed by Alfaro et al. (2008).

A successful prediction depends on the setting of the model and the ratio of data. It was demonstrated that even companies which did not default in reality can be predicted – they show serious financial difficulties, such as the insolvency management or the default of payments and the differences which are not shown for active companies. Primarily, there are distinct shifts in the levels of company indicators in the time period before default. Using the DT method led to the construction of easily represented and understandable diagrams for the representation of limits which mean facing financial difficulties for the companies, i.e. the decision rules.

Indeed, the obtained results could be further enhanced via testing of the accuracy for different classifiers settings and different proportions of the training/validation/testing set. The results can differ according to the ratios of the defaulted companies, however, the availability and incompleteness of the acquired data led to this ratio of the companies applied to the total amount. Apart from a certain imbalance of the dataset, the models enable to predict distress standings even several years ahead of the default, with the highest accuracy when using the methods of Adaptive Boosting and Decision Trees. The prediction of default is marked by the results which are highly influenced by both the data used and the setting of the classification models, therefore, it cannot be equivocally claimed that the results can be generalized for all similar companies. Even other authors refer to empirical researches

doi: 10.17221/374/2015-AGRICECON

with a similar orientation or number of companies in the set of data.

REFERENCES

Alfaro E., Gamez M., Garcia N. (2008): Linear discriminant analysis versus adaboost for failure forecasting. Revista Española de Financiación y Contabilidad, 37: 13–32.

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

Aziz M.A., Dar H.A. (2006): Predicting corporate bankruptcy: where we stand? Corporate Governance: The International Journal of Business in Society, 6: 18–33.

Beaver W.H. (1966). Financial ratios as predictors of failure. Journal of Accounting Research, 4: 71–102.

Balcaen S., Ooghe H. (2006): 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. The British Accounting Review [online], 38: 63–93.

Breiman L. (1993): Classification and Regression Trees. Chapman & Hall, Boca Raton.

Ding Y., Song X., Zen X. (2008): Forecasting financial condition of Chinese listed companies based on support vector machine. Expert Systems with Applications, 34: 3081–3089.

Freedman D.A. (2009): Statistical Models: Theory and Practice. Cambridge University Press.

Freund Y., Schapire R. (1996): Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning. Morgan Kauffman, San Francisco, CA.

Frydman H., Altman E.I., Kao D.L. (1985): Introducing recursive partitioning for financial classification: the case of financial distress. The Journal of Finance, 40: 269–291.

Gepp A., Kumar K., Bhattacharya S. (2010): Business failure prediction using decision trees. Journal of Forecasting, 29: 536–555.

Li H., Sun J., Wu J., (2010): Predicting business failure using classification and regression tree: An empirical comparison with popular classical methods and top classification mining methods. Expert Systems with Applications, 37: 5895–5904.

Lin F., McClean S. (2001): A data mining approach to the prediction of corporate failure. Knowledge-Based Systems, 14: 189–195.

Huarng K., Yu H., Chen C. (2005): The application of decision trees to forecast financial distress companies. In: Proceedings of 2005 International Conference on Intelligent Technologies and Applied Statistics, Taipei, Taiwan.

Min S.H., Lee J., Han I. (2006): Hybrid genetic algorithms and support vector machines for bankruptcy prediction. Expert Systems with Applications, 31: 652–660.

Min J.H., Lee Y.C. (2005): Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications, 28: 603–614.

Niknya A., Darabi R., Vakili F. (2013): Financial distress prediction of Tehran stock exchange companies using support vector machines. European Online Journal of Natural and Social Science, 2 (Special Issue on Accounting and Management).

Ohlson J.A. (1980): Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18: 109–131.

Quinlan J.R. (1993): Programs for Machine Learning. Morgan Kaufmann Publishers.

Vapnik V.M. (1995): The Nature of Statistical Learning Theory. Springer, New York.

Vavřina J., Hampel D., Janová J. (2013): New approaches for the financial distress classification in agribusiness. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 61: 1177–1182.

Zmijewski M. (1984): Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research, 22: 59–86.

Received December 24, 2015
Accepted January 28, 2016
Published online March 21, 2017