Central European Review of Economic Issues
EKONOMICKÁ REVUE

Risk assessment of VAT entities using selected data mining models

Stanislav CÚT*

Department of Quantitative Methods and Information Systems, Faculty of Economics, University of Matej Bel, Tajovského 10, 974 01 Banská Bystrica, Slovak Republic.

Abstract

The goal of the paper was to evaluate the classification ability of selected types of data mining methods, focusing on neural networks, decision trees and random forests, within the risk assessment of VAT entities. The data set used for the testing contained information on the risk of taxpayers who were obliged to file VAT returns in the calendar year 2012. The highest classification ability among the constructed models was achieved by the multilayer perceptron model. The lowest classification ability was demonstrated by the decision tree method, using the default growth exhaustive CHAID algorithm.

The presented study focuses on the risk evaluation of Slovak value-added tax (VAT) payers.

Keywords

Data-mining methods, neural networks, decision trees, random forests, classification analysis, VAT.

JEL Classification: C38, C45, H26

* stanislav.cut@umb.sk

The paper was supported by the grant scheme VEGA No. 1/0765/12 – Research into possibilities and perspectives of employing traditional and alternative approaches in financial management and financial decision-making in the changing economic environment.

ISSN 1212-3951 (Print), 1805-9481 (Online)

ER-CEREI, Volume 18: 5–14 (2015).
doi: 10.7327/cerei.2015.03.01

Risk assessment of VAT entities using selected data mining models

Stanislav CÚT

1. Introduction

Despite the fact that the issue of early warning systems dates to the early 1920s, due to the sharp fluctuations in the global economy and the interconnected rapid development of tools and techniques used in the financial management of businesses, the interest of professionals in this issue has grown significantly, starting in the second half of the 1970s and gaining increasing importance ever since.

In 1966, the first results in the form of Beaver one-dimensional discriminant analysis (Beaver, 1966), belonging to the category of mathematical and statistical methods, appeared in the research on early warning systems. In accordance with the complexity of the current economic problems and the one-dimensional Beaver discriminant analysis, with the aim of undertaking a more comprehensive assessment of the financial performance of an enterprise, the multidimensional discriminant analysis method (MDA) was presented by Altman (1968). Among the authors who used the methodology of the earlier-presented Altman's Z-score in their works belong, for example, Deakin (1972), Blum (1974) and Poston et al. (1994).

As early as 1980, Ohlson was the first to attempt the application of the logit linear probability model to financial distress prediction. The benefit of this approach in comparison with multidimensional discriminant analysis was that the independent variables did not have to fulfil the requirement of multivariate normal distribution and equal covariance matrixes. When the sample data do not meet these two assumptions, the results of an MDA model may be suspicious. For a detailed understanding of financial distress prediction using the logit linear probability model, please refer for example to Zavgren (1983), Gentry et al. (1985), Keasey and Watson (1987), Ooghe et al. (1995) and others.

With the recent rapid development of information technologies and the related need for systematization of the process of selection, exploration and modelling of large amounts of data in the research on financial distress early warning decision-making, progress can be observed from traditional statistical methods to data-mining methods belonging to the category of machine learning based on artificial intelligence.

Due to their performance, ease of use and high learning abilities, even if the input data contain information that is too noisy and incomplete or highly correlated variables, data mining technologies are still widely used, for example in predicting exchange rate movements on securities markets, analysing the reasons for changes in a service provider, determining the probability of machinery failures, scoring an applicant to determine the risk of extending credit or detecting fraudulent transactions in insurance claim databases.

From the large number of existing classification methods based on the principles of artificial intelligence, neural networks (NN), evolutionary algorithms (EA), rough set (RS) based techniques, case-based reasoning (CBR), decision trees (DT) and others have been classified by authors such as Ravi Kumar and Ravi (2007). In the context of this paper, the focus is on the application of selected types of neural networks, classification trees and random forests to evaluate and compare their classification abilities for the available data set.

Regarding the disputes concerning neural networks and their applications in economic disciplines, attention should be paid, for example, to the works of Hornik et al. (1990), Odom and Sharda (1990), Tam and Kiang (1992), Shah and Murtaza (2000), Atiya (2001), Sen et al. (2004), Wallace (2008), Chen and Du (2009) and others. The performance of artificial neural networks in financial distress prediction has often been compared with that of traditional statistical methods (such as MDA and logistic regression), and in many cases, this has led to the confirmation of the expected higher classification abilities of neural networks, as presented by Fletcher and Goss (1993), Pendharkar (2005), Ravi Kumar and Ravi (2007) and Tseng and Hu (2010).

The decision tree and recursive partitioning methods were applied, for example, in the works by Quinlan (1986), McKee and Greenstein (2000), Cho et al. (2010) and others. The accuracy of decision trees has also often been compared with the traditional statistical methods, such as MDA (Gepp et al., 2010), logit analysis (Chen, 2011) or neural networks and support vector machines (Olson et al., 2012).

During the creation of the random forests data mining method, its author – Breiman (1996, 2001) –was inspired by the previously presented works by authors

like Amit and Geman (1997) and the random split selection approach of Diettrich (2000). Among the later-presented works, the most interesting were the ones, for example, by Meinshausen (2006) and Biau and Devroy (2010).

The paper is divided into six chapters. In the second chapter, the focus is on the description and preparation of data samples to be used in the modelling phase. Various data mining techniques are applied to this prepared data set in the third, fourth and fifth chapters, specifically neural networks, classification trees and random forests. The evaluation and comparison of the above-presented model's performance, along with other conclusions and potential options for future research, are summarized in the sixth chapter.

2. Data Selection and Preparation for Modelling

The focus of the presented study is on the risk evaluation of Slovak value-added tax (VAT) payers. The data set used for the present article contains real characteristics related to the risk assessment of selected registered value-added tax payers for the tax period 2012, to prevent unauthorized settlement of funds from the state budget of the Slovak Republic. Using various data mining methods, an attempt is made to create a model, and respectively an appropriate classification rule, to classify selected tax units accurately into a VAT risky payer category.

The selection of the paper's topic and its main aim reflect the current situation in the Slovak tax administration, namely the document *Action Plan in Defiance of Tax Evasions for the Years 2012–2016*, approved by the resolution of the Slovak Government, which foresees the implementation of risk analysis methods and the following selection of tax entities to control to contribute to the fight against tax fraud. Unauthorized VAT refunds from the state budget represent one of the most common forms of tax fraud and it is therefore necessary to adopt prevention efforts to minimize this illegal activity, including efficient choice of entrepreneurs for VAT control.

The input data matrix for the purpose of model creation, minus the missing and noisy data, consists of 27 indicators evaluated for each of the 1 290 randomly selected taxpayers, mostly small and medium-sized enterprises, which requested VAT refunds during the tax period 2012. The abbreviations and brief descriptions of the dependent variable and a set of independent variables included in the models are presented in Table 1.

As the risky taxpayers for the data-mining models' learning procedures, 645 entities were randomly selected that filed VAT returns for which the VAT control subsequently detecting findings over the given threshold for the tax period of 2012.

The dichotomous dependent variable *RISK_DEF* takes the numerical value 1 for risky taxpayers and the value 0 for non-risky taxpayers. The riskiness of the selected entities was assessed using 26 explanatory variables (predictors) with the character of continuous, dichotomous or categorical encoded variables.

To eliminate the impact of significantly different units and ensure the comparability of the selected continuous explanatory variables, their standardized values were used.

To assess the predictive accuracy of the proposed models, a Bernoulli probability distribution split was performed to divide the sample containing all 1 290 observations into 2 subsamples, specifically the training subsample used to build the model during the learning procedure and the second testing subsample used for the evaluation of the final classification ability. The absolute and relative frequencies of the included subsamples are presented in Table 2.

Table 2 Absolute and relative frequencies of the subsamples entering the models

Observations		Number of observations	Share
Subsample	*Training*	1047	81.1 %
	Testing	243	18.8 %
Valid		1290	100.0 %
Excluded		0	
Total		1290	

Each of the selected data mining methods – neural networks, classification trees and random forests – was applied to the data sets defined in this way. To evaluate the performance of the proposed classification methods, their classification abilities were finally compared.

3. Neural Networks

One of the effective solutions to approximate non-linear and complex reliance in the data, mainly due to its simplicity, the absence of restrictive assumptions for the relevant application and applicability even in cases when the input data samples contain noisy data, highly correlated explanatory variables or incomplete time series, appears to be the method of artificial neural networks from the category of models based on neurophysiological knowledge-based studies of the functioning of the human brain (Wallace, 2008).

As the family grew, most of the new models were designed for non-biological applications. Of the large number of existing types of neural networks, the most historically suitable for economic disciplines' applications proved to be the linear network (LN), generalized

regression neural network (GRNN), probabilistic neural network (PNN), multilayer perceptron (MLP), radial basis function (RBF) and Kohen network (SOFM).

In the following chapter, the focus is on the practical application of the MLP neural network as one of the most commonly used types of neural networks to create a model, and respectively a classification rule, to classify selected taxpayers accurately into the appropriate category (default/non-default).

In general, in the framework of the neural networks and artificial intelligence learning systems, it is possible to separate the learning phase from the life phase, when the knowledge gained can be used to solve a particular problem (prediction, classification, etc.) (Berka, 2003).

Because of the relatively wide range of adjustment of the model's input parameters, eight differing network topologies, the type of activation function used for the hidden and the output layer of the neural network, the type of learning and ultimately the optimization algorithms for estimating the synaptic weights were tested. The aim of this procedure was to reach a setting of synaptic weights *wij*, which would be able to minimize the deviation between the predicted and the target output, while optimizing the classification ability of the model.

As shown in Table 3, in the final model, a single-layer neural network topology is used, consisting of 14 nodes, using the backpropagation algorithm of

Table 1 Abbreviation and description of the variables entering the models

Abbreviation of the variable	*Description of the variable*
BUS_ENTITY	Type of the business entity (self-employed person/enterprise)
VAT_PERIOD	Type of the VAT period (monthly/quarterly)
SIZE_TP	Size of the taxpayer (small/medium/large)
VAT_TURN	Turnover on the VAT
CHANGE_BP	Change of business partner (yes/no)
CHANGE_LR	Change in leading role (yes/no)
CHECKS	Number of checks on the spot of the entity
FORM_NOTICE	Sent a formal notice to file a VAT return (yes/no)
CHANGE_ADDRESS	Number of changes in the taxpayer's address
VAT_IRREG	Irregularity in the VAT return filing (yes/no)
ADD_VAT	Number of additional filings of VAT returns
NON_FIN_STAT	Non-filed financial statements by the taxpayer (yes/no)
EXEC_PROC	Starting of the execution proceedings (yes/no)
HIGH_VAT_TURN	Entity with a low amount of assets, but a high turnover (yes/no)
HIGH_VAT_REG	High VAT, right after the registration for VAT (yes/no)
VOL_VAT_REG	Voluntary registration for VAT (yes/no)
VAT_ARREARS	Registered VAT arrears (yes/no)
ONCE_HIGH_REF	One high refund of VAT required (yes/no)
COEF_RANGE	Tax liability or tax refund coefficient within the defined range (yes/no)
LOW_TURN	Low or any turnover in the income statement (yes/no)
RISK_COEF_1	Risk coefficient no. 1
RISK_COEF_2	Risk coefficient no. 2
RISK_SECTOR	High-risk business sector (yes/no)
NEW_REG_VAT	Newly registered VAT entity (yes/no)
VAT_FRAUD_PAST	VAT fraud in the past (yes/no)
BUS_FORM	Business form of the entity
RISK_DEF	Risk of default (yes/no)

errors.[1] In the first, forward step of the algorithm, the values of input matrix p_i are multiplied by the synaptic weights w_i and through the summary node raised by the scalar threshold b of the node. In general, the input a of the activation function $f(a)$ can be defined as:

$$a = w_1 p_1 + w_2 p_2 + ... + w_R p_R + b = \sum_{i=1}^{R} w_i p_i + b. \quad (1)$$

As the activation function combining the weighted sum of inputs in the hidden layer, the hyperbolic tangent function was chosen. It takes real-valued arguments and transforms them into the range (–1, 1). This function has the form:

$$f(a)_{\tanh} = \frac{e^a - e^{-a}}{e^a + e^{-a}}. \quad (2)$$

The same type was also chosen in the output layer activation function $f(a)_{OUT}$ transforming the output of the hidden layer activation function $f(a)_{HIDDEN}$ on the scalar n to which the following applies:

$$n = f(a). \quad (3)$$

The number of nodes in each layer of the neural network was established using an automatic algorithm, and with regard to the type of activation function selected, the sum of squares error function is used. In the second, back step of the algorithm, the synaptic weights w_{ij} are then recalculated to minimize the sum of squares error.

Table 3 MLP information

Hidden Layer	Number of Hidden Layers	1
	Number of Units in Hidden Layer	12
	Activation Function	Hyperbolic tangent
Output Layer	Dependent Variable	*RISK_DEF*
	Number of Units	2
	Activation Function	Hyperbolic tangent
	Error Function	Sum of Squares

The type of learning procedure of the neural network was due to the medium-sized data set considered, defined by the mini-batch method, and as the optimization algorithm to estimate the synaptic weights, the gradient descent method was finally used.[2] The estimation algorithm was stopped because the maximum number of epochs without reducing a sum of squares error was reached, in this case limited to the default value of 1.

The classification table shows the practical results of using the network. From Table 4, we can conclude that of the 130 risky taxpayers included in the testing subsample, 122 entities were correctly classified by the network; in relative terms, this is nearly 94%. Of the 111 non-defaulters, 105 were classified correctly, which signals that the network performs considerably better in predicting non-defaulters than defaulters. Globally, 94.2% of the testing subsample cases containing exactly 241 observations (2 observations were excluded) were classified correctly, corresponding to 5.8% classified incorrectly.

Table 4 Classification ability of the MLP

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	507	25	95.3 %
	1	30	485	94.2 %
	Overall %			94.7 %
Testing	0	105	6	94.6 %
	1	8	122	93.8 %
	Overall %			94.2 %

Due to the fact that the numbers of correctly classified entities are roughly equal under the training and the testing subsample, it can be concluded that the network was not over-trained during the learning phase and thus it is able to generalize new unseen data. The fact that during the process of learning a relatively sharp decrease in the sum of squares errors across the samples (from 44 028 in the training subsample to 11 901 in the testing subsample) was observed can also be considered as positive.

4. Decision Trees

In the following chapter, the decision trees method is applied to the data set containing information about the riskiness of VAT payers as another data mining method used to predict or classify a selected set of units.

The advantages of decision trees lie mainly in the illustration and clarity of the inductive learning process, leading to rapid evaluation and interpretability of the obtained results, even without perfect knowledge of the statistical methods.

[1] For the backpropagation of errors algorithm, see Haykin (1998).

[2] For the gradient descent method, see IBM SPSS Statistics 20 (2011b).

From the large choice of available algorithms to generate decision trees, it was decided to use the growth exhaustive CHAID method, particularly because of its generally high classification ability and frequency of practical use.

The exhaustive CHAID growing method originated as a modification of the CHAID algorithm (*chi-squared automatic interaction detection*), in which the basic criterion for selecting the branching is the effort to increase the purity of child nodes and respectively to reduce the impurity measure called entropy. As the name of the algorithm suggests, the determination of the number of nodes and merging and respectively eventual splitting of the categories is conducted through the similarity based on the chi-square statistic. In each step, CHAID chooses the independent variable that has the strongest interaction with the dependent variable. Unlike the CHAID method, the exhaustive CHAID growth method assesses the possibility of partitioning and merging for each of the observed predictors. Since the method works only with nominal or ordinal variables, continuous variables are categorized under the algorithm (Terek et al., 2010).

Since the aim is to keep the tree fairly simple, the tree's growth is limited by raising two conditions. The first condition applies to the minimum number of cases for parent nodes, which is limited to 100. The second condition limits the minimum number of cases for child nodes to 50. Nodes that do not satisfy these criteria will not be split. The maximum depth is limited to 3 levels using an automatic algorithm.

In this model, the significance level for merging categories and splitting nodes for both is 0.05, the maximum number of iterations is limited to 100 and the maximum number of categories to classify continuous explanatory variables is limited to 10.

To grow the final decision tree, the same data sample and the statistical software were used as in the case of the MLP network. . For the graphic representation of the tree structure growth over the training subsample, see Appendix 1.

The information about the variables included in the model at various levels is presented by the tree structure representation using the tree diagram. The frequency tables for each node contain the number of cases (count and percentage) for each category of the dependent variable *RISK_DEF*. The predicted category with the highest count in each node is highlighted.

As can be seen in Appendix 1, the final decision tree is characterized by a 3-level hierarchical structure consisting of 12 nodes. The root node containing the dependent variable *RISK_DEF* represents the entire training subsample, exactly 1 047 observations. Using the exhaustive CHAID growing method, the *VAT_PERIOD* variable is the best predictor of default. The classification tree has 4 parent nodes consisting of 3 categories of the independent variables *VAT_PERIOD* and *VAT_FRAUD_PAST*. The other predictors entering the model are *RISK_COEF_1* and *CHANGE_BP*.

According to Table 5, which presents the classification ability of the growth decision tree, it can be concluded that the overall accuracy of the classification model reached 90.9% on the testing subsample. A higher percentage of correctly classified entities, exactly 93.8%, was achieved by a tree in the category of non-risky VAT payers. In the category of risky VAT payers, 115 of the 130 defaulters were correctly classified.

Table 5 Classification ability of the exhaustive CHAID decision tree

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	485	47	91.2%
	1	45	470	91.3%
	Overall %			91.2%
Testing	0	106	7	93.8%
	1	15	115	88.5%
	Overall %			90.9%

5. Random Forests

Random forests, as a classification and prediction tool, consist of predictions obtained by aggregating over the ensemble of many decision trees in which each tree in the ensemble is grown in accordance with a random parameter (Biau, 2012).

According to the speed of the current algorithm, its resistance to overfitting and sufficient flexibility of usage even in the case that the input data sample contains more variables than observations, and since standard parametric methods do not work, random forests are considered a very powerful tool for both the classification and the regression. Among the limitations that may appear in their practical implementation belong loss of interpretability and additional information on the impact of different variables on the classification or prediction accuracy of the model as a whole. To solve this problem, it is possible to use the help of the graphical presentation of impurity measure decrease or the permutational significance of the predictors entering the model (Berk, 2008).

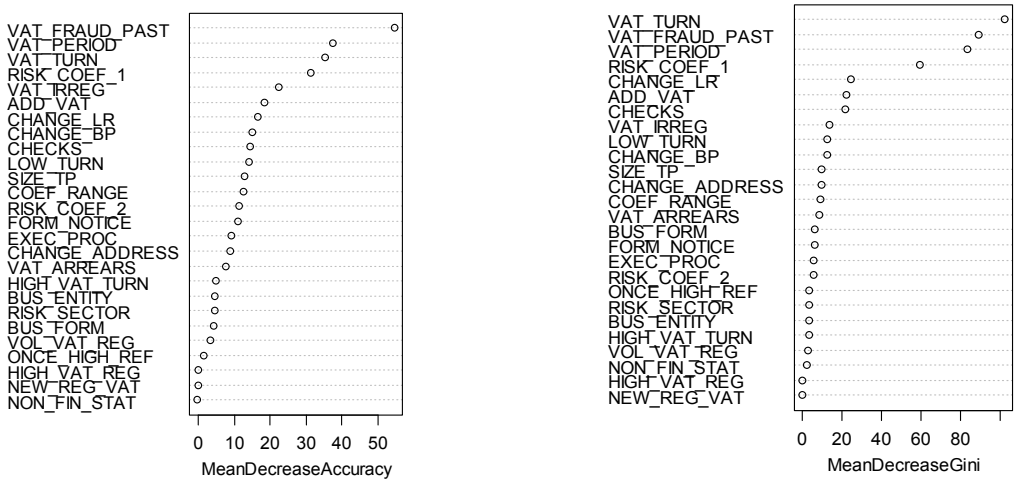

Figure 1 Significance of predictors entering the model

The random forest was obtained by using the statistical software R 3.0.2 and its function *randomForest( )* from the package randomForest (Liaw and Wiener, 2002).

In the analysis, the default number of trees is set to 500 and the number of predictors selected in each division is determined by an adjustable value *mtry*. With regard to the fact that the estimation of out-of-bag (OOB) error, in each case of an *mtry* setting, does not differ significantly, as the default number of *mtry* on the level 3 is used. This means that for each splitting of trees, the algorithm uses 3 randomly selected predictors. The unseen observation is then classified into the category following the majority of the trees grown in the random forest.

Figure 1 presents the significance of the independent variables entering the model. According to the mean decrease in accuracy and the mean decrease of impurity represented by the Gini index, the most important independent variables seem to be *VAT_FRAUD_PAST*, *VAT_PERIOD*, *VAT_TURN* and *RISK_COEF_1*. Despite the fact that the significance level of some predictors appears to be relatively low, all of the 26 original explanatory variables are retained in the analysis. Based on Table 6, which represents the classification ability of the resulting random forest, it can be stated that within the testing subset the random forest is able to classify correctly about 93.8 % of risky VAT payers and about 91.2% of non-risky entities. Of the 243 VAT payers, the random forest is able to classify correctly 225, which corresponds to an overall error rate of 7.4%.

Table 6 Classification ability of the random forest

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	502	30	94.4%
	1	26	489	95.0%
	Overall %			94.7%
Testing	0	103	10	91.2%
	1	8	122	93.8%
	Overall %			92.6%

6. Conclusion

The aim of the paper was to apply selected types of classification and decision-making data mining methods, including the evaluation of their classification abilities, to the selected data set of 1 290 potentially risky VAT payers.

From Table 7, it is evident that the best overall classification ability of 94.20% was achieved by the feed-forward MLP neural network using the backpropagation algorithm of errors. The neural network reached the highest level of accuracy in the category of non-risky entities (94.60%), and in accordance with a random forest model, also in the category of risky VAT payers (93.80%). The lowest classification ability from the previously mentioned models at the level of 90.90% was achieved by the exhaustive CHAID growth decision tree.

Despite the above-presented evaluation of the classification abilities, in which the most appropriate data mining tool appears to be the MLP neural network, it is also necessary to keep in mind that it has limitations

that may appear with its practical implementation, which can eventually result in highly biased estimates. Among these limitations, the following often appear problematic: randomization of the data sample selection due to the infrequent occurrence of default cases, from which quite a high level of sensitivity to the ratio of default and non-default entities included in the sample compilation is derived, the influence of various setting possibilities of the appropriate parameters and also the difficult or sometimes even impossible interpretability of the resulting synaptic weights.

When assessing the usability of the particular data mining method, it is necessary to consider the main purpose of the research and respectively the type of decision-making task. In the case that it is necessary to describe the existing relationships between the dependent and the ensemble of independent variables entering the model, the best choice, due to its simple interpretability, seems to be the method of decision trees. If the interpretability of the model is put into the background, better and faster results can usually be achieved by neural networks or random forests.

Future research may discuss further non-standard tools used in economic forecasting and confront their strengths and weaknesses in relation to standard and well-established statistical classification methods.

Table 7 Comparison of the classification abilities of the data mining models created

Testing subsample	Classification tool					
	Multilayer Perceptron		*Exhaustive CHAID Classification Tree*		*Random Forest*	
Observed	Predicted		Predicted		Predicted	
	0	1	0	1	0	1
0	94.6%	5.4%	93.8%	6.2%	91.2%	8.8%
1	6.2%	93.8%	11.5%	88.5%	6.2%	93.8%
Overall %	94.2%		90.90%		92.6%	

References

ALTMAN, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23(4): 589–609. http://dx.doi.org/10.1111/j.1540-6261.1968.tb00843.x

AMIT, Y., GEMAN, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation* 9(7): 1545–1588. http://dx.doi.org/10.1162/neco.1997.9.7.1545

ATIYA, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks* 12(4): 929–935. http://dx.doi.org/10.1109/72.935101

BEAVER, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research* 4: 71–111. http://dx.doi.org/10.2307/2490171

BERK, R. A. (2008). *Statistical Learning from a Regression Perspective (Springer series in Statistics)*. New York: Springer-Verlag. http://dx.doi.org/10.1007/978-0-387-77501-2_1

BERKA, P. (2003). *Dobývání znalostí z databází*. Praha: Academia.

BIAU, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research* 13: 1063–1095.

BIAU, G., DEVROYE, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis* 101(10): 2499–2518. http://dx.doi.org/10.1016/j.jmva.2010.06.019

BLUM, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research* 12(1): 1–25. http://dx.doi.org/10.2307/2490525

BREIMAN, L. (2001). Random forests. *Machine Learning* 45(1): 5–32. http://dx.doi.org/10.1023/A:1017934522171

BREIMAN, L. (1996). Bagging predictors. *Machine Learning* 24(2): 123–140. http://dx.doi.org/10.1023/A:1018054314350

DEAKIN, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research* 10(1): 167–179. http://dx.doi.org/10.2307/2490225

DIETTERICH, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning* 40(2): 139–157. http://dx.doi.org/10.1023/A:1007607513941

FLETCHER, D., GOSS, E. (1993). Forecasting with neural networks: an application using bankruptcy data. *Information and Management* 24(3): 159–167. http://dx.doi.org/10.1016/0378-7206(93)90064-Z

GENTRY, J. A., NEWBOLD, P., WHITFORD, D. T. (1985). Classifying bankrupt firms with funds flow components. *Journal of Accounting Research* 23(1): 146–160. http://dx.doi.org/10.2307/2490911

GEPP, A., KUMAR, K., BHATTACHARYA, S. (2010). Business failure prediction using decision trees. *Journal of Forecasting* 29(6): 536–555. http://dx.doi.org/10.1002/for.1153

HAYKIN, S. (1998). *Neural Networks: A Comprehensive Foundation*. Singapore: Pearson Education.

HORNIK, K., STINCHCOMBE, M., WHITE, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feed forward network. *NeuralNetworks* 3(5): 551–560. http://dx.doi.org/10.1016/0893-6080(90)90005-6

CHEN, M.-Y. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications* 38(9): 11261–11272. http://dx.doi.org/10.1016/j.eswa.2011.02.173

CHEN, W.-S., DU, Y.-K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications: An International Journal* 36(2): 4075–4086. http://dx.doi.org/10.1016/j.eswa.2008.03.020

CHO, S., HONG, H., HA, B.-C. (2010). A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction. *Expert Systems with Applications* 37(4): 3482–3488. http://dx.doi.org/10.1016/j.eswa.2009.10.040

KEASEY, K., WATSON, R. (1987). Non-financial symptoms and the prediction of small company failure: a test of Argenti's hypotheses. *Journal of Business Finance & Accounting* 14(3): 335–354. http://dx.doi.org/10.1111/j.1468-5957.1987.tb00099.x

LIAW, A., WIENER, M. (2002). Classification and regression by random forest. *R News* 2(3): 18–22.

MCKEE, T. E., GREENSTEIN, M. (2000). Predicting bankruptcy using recursive partitioning and a realistically proportioned data set. *Journal of Forecasting* 19(3): 219–230. http://dx.doi.org/10.1002/(SICI)1099-131X(200004)19:3<219::AID-FOR752>3.3.CO;2-A

MEINSHAUSEN, N. (2006). Quantile regression forests. *Journal of Machine Learning Research* 7: 983–999.

ODOM, M. R., SHARDA, R. (1990). A neural networks model for bankruptcy prediction. In: *Proceedings of the IEEE International Conference on Neural Network (San Diego)*. New York: IEEE, 163–168.

OLSON, D. L., DELEN, D., MENG, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems and Electronic Commerce* 52(2): 464–473. http://dx.doi.org/10.1016/j.dss.2011.10.007

OOGHE, H., JOOS, P., DE BOURDEAUDHUIJ, C. (1995). Financial distress models in Belgium: The results of a decade of empirical research. *International Journal of Accounting* 30(3): 245–274.

PENDHARKAR, P.C. (2005). A threshold varying artificial neural network approach for classification and its application to bankruptcy prediction problem. *Computers and Operations Research* 32(10): 2561–2582. http://dx.doi.org/10.1016/j.cor.2004.06.023

POSTON, K. M., HARMON, W. K., GRAMLICH, J. D. (1994). Financial ratios as predictors of turnaround versus failure among financially distressed firms. *The Journal of Applied Business Research* 10: 41–56.

QUINLAN, L. (1986). Induction of decision trees. *Machine Learning* 1(1): 81–106. http://dx.doi.org/10.1007/BF00116251 http://dx.doi.org/10.1023/A:1022643204877

RAVI KUMAR, P., RAVI, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – a review. *European Journal of Operational Research* 180(1): 1–18. http://dx.doi.org/10.1016/j.ejor.2006.08.043

SEN, T., GHANDFOROUSH, T. P., STIVASON, C. T. (2004). Improving prediction of neural networks: a study of two financial prediction tasks. *Journal of Applied Mathematics and Decision Sciences* 8(4): 219–233. http://dx.doi.org/10.1155/S1173912604000148

SHAH, J. R., MURTAZA, M. B. (2000). A neuralnetwork based clustering procedure for bankruptcy prediction. *American Business Review* 18(2): 80–86.

TAM, K. Y., KIANG, M. Y. (1992). Managerial applications of neural networks: the case of bank failure prediction. *Management Science* 38(7): 926–947. http://dx.doi.org/10.1287/mnsc.38.7.926

TEREK, M., HORNÍKOVÁ, A., LABUDOVÁ, V. (2010). *Hĺbková analýza údajov*. Bratislava: Iura Edition.

TSENG, F.-M., HU, Y.-C. (2010). Comparing four-bankruptcy prediction models: logit, quadratic interval logit, neural and fuzzy neural networks. *Expert Systems with Applications* 37(3): 1846–1853. http://dx.doi.org/10.1016/j.eswa.2009.07.081

WALLACE, M. P. (2008). Neural networks and their application to finance. *Business Intelligence Journal* 1(1): 67–77.

ZAVGREN, C. V. (1983). Assesing the vulnerability to failure of American industrial firm: A logistic analysis. *Journal of Bussines Finance & Accounting* 12(1): 19–45. http://dx.doi.org/10.1111/j.1468-5957.1985.tb00077.x

Additional sources

IBM SPSS Statistics 20 (2011a). *Decision Trees*. [Online], accessed at 05. 04. 2014. Available at: <http://www.csun.edu/sites/default/files/decision-trees 20-64bit.pdf>.

IBM SPSS Statistics 20 (2011b). *Neural Networks*. [Online], accessed at 05. 04. 2014. Available at: <http://www.csun.edu/sites/default/files/neural-network20-64bit.pdf>.

IBM SPSS Statistics 20 (2011c). *Algorithms*. [Online], accessed at 05. 04. 2014. Available at: <http://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Statistics_Algo rithms.pdf>.

Appendix 1 Structure of exhaustive CHAID decision tree growth over the training subsample