STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 155–171, DOI 10.21307/stattrans-2019-020

EXTREME GRADIENT BOOSTING METHOD IN THE PREDICTION OF COMPANY BANKRUPTCY

Barbara Pawełek[1]

ABSTRACT

Machine learning methods are increasingly being used to predict company bankruptcy. Comparative studies carried out on selected methods to determine their suitability for predicting company bankruptcy have demonstrated high levels of prediction accuracy for the extreme gradient boosting method in this area. This method is resistant to outliers and relieves the researcher from the burden of having to provide missing data. The aim of this study is to assess how the elimination of outliers from data sets affects the accuracy of the extreme gradient boosting method in predicting company bankruptcy. The added value of this study is demonstrated by the application of the extreme gradient boosting method in bankruptcy prediction based on data free from the outliers reported for companies which continue to operate as a going concern. The research was conducted using 64 financial ratios for the companies operating in the industrial processing sector in Poland. The research results indicate that it is possible to increase the detection rate for bankrupt companies by eliminating the outliers reported for companies which continue to operate as a going concern from data sets.

Key words: XGBoost, company bankruptcy, machine learning, outlier.

1. Introduction

An important issue in economic and financial decision-making is to predict business failure (bankruptcy prediction, credit scoring) (Nanni and Lumini, 2009). A number of data classification methods are used in company bankruptcy prediction and in credit scoring (Baesens et al., 2003; Lessmann et al., 2015). In the paper by Baesens et al. (2003) the authors evaluate and compare different types of classifiers, for example logistic regression, discriminant analysis, *k*-nearest neighbour, neural networks, decision trees, support vector machines, least-squares support vector machines. Their results suggest that the neural network, least-squares support vector machines, logistic regression and linear discriminant analysis yield a very good performance. The authors of the paper by Lessmann et al. (2015) update the study of Baesens et al. (2003) and compare 41 different classification algorithms such as individual classifiers (Bayesian network, CART, extreme learning machine, kernalized ELM, *k*-nearest neighbour,

[1] Department of Statistics, Cracow University of Economics, Kraków, Poland.
E-mail: barbara.pawelek@uek.krakow.pl. ORCID ID: https://orcid.org/0000-0002-9589-6043.

J4.8, linear discriminant analysis, linear support vector machine, logistic regression, multilayer perceptron artificial neural network, naive Bayes, quadratic discriminant analysis, radial basis function neural network, regularized logistic regression, SVM with radial basis kernel function, voted perceptron), homogenous ensemble classifiers (alternating decision tree, bagged decision trees, bagged MLP, boosted decision trees, logistic model tree, random forest, rotation forest, stochastic gradient boosting), heterogeneous ensemble classifiers (simple average ensemble, weighted average ensemble, stacking, complementary measure, ensemble pruning via reinforcement learning, GASEN, hill-climbing ensemble selection, HCES with bootstrap sampling, matching pursuit optimization ensemble, top-*T* ensemble, clustering using compound error, *k*-means clustering, kappa pruning, margin distance minimization, uncertainty weighted accuracy, probabilistic model for classifier competence, *k*-nearest oracle). Their results suggest that heterogeneous ensembles classifiers perform well.

The main criterion for assessing the suitability of a bankruptcy prediction model (and a credit scoring model) is its prediction ability. In general, performance measures split into three types: the measures that assess the discriminatory ability of the model; the measures that assess the accuracy of the model's probability predictions; the measures that assess the correctness of the model's categorical predictions (Lessmann et al., 2015).

Researchers are seeking to identify sources of the errors committed when predicting company bankruptcy. One of the reasons for misclassification of objects is the heterogeneity of a research data set. Bankruptcy prediction models are developed on the basis of the financial ratios included in financial statements. An analysis of the financial details of the companies which went bankrupt and those which continue to operate as a going concern leads to the conclusion that some of the companies in Poland with unfavourable financial ratios do not go bankrupt (Pawełek et al., 2017). In light of the above considerations, the homogeneity (in terms of financial condition assessment) of the set of companies which continue to operate as a going concern is called into doubt.

Machine learning methods are increasingly used in company bankruptcy prediction (e.g. Brown and Mues, 2012; García et al., 2019; Pawełek, 2017). Comparative studies carried out on selected methods to determine their suitability for predicting company bankruptcy have demonstrated high levels of prediction accuracy for the extreme gradient boosting method (Xia et al., 2017; Zięba et al., 2016). The study by Zięba et al. (2016) adopted a bankruptcy prediction model for the companies operating in the industrial processing sector in Poland over time horizons of one, two, three, four and five years, using a number of machine learning methods (e.g. linear discriminant analysis, multilayer perceptron with a hidden layer, decision rules inducer, decision tree model, logistic regression, boosting algorithm AdaBoost, cost-sensitive boosting algorithm AdaCost, support vector machines, random forest, boosted trees trained with extreme gradient boosting). The databases subject to analysis were not free from outliers and the missing data were not imputed.

This research was undertaken to investigate the combined findings concerning the heterogeneous nature of the set of companies which continue to operate as a going concern in terms of their financial condition and the high

accuracy of the extreme gradient boosting method in predicting company bankruptcy.

The aim of this study is to present the results of our empirical research on the impact that the elimination of outliers from data may have on the accuracy of the extreme gradient boosting method in predicting company bankruptcy. The added value of the study is demonstrated by the proposed application of the extreme gradient boosting method in bankruptcy prediction based on data free from the outliers reported for companies which continue to operate as a going concern.

The study is divided as follows: Section 2 provides a description of the relevant databases and the extreme gradient boosting method; Section 3 outlines the research procedure used; the results of the empirical research are presented and discussed in Section 4; and the main findings of the study are summarised in Section 5.

Table 1. Financial ratios

Ratio	Description	Ratio	Description
W_1	net profit / total assets	W_{33}	operating expenses / short-term liabilities
W_2	total liabilities / total assets	W_{34}	operating expenses / total liabilities
W_3	working capital / total assets	W_{35}	profit on sales / total assets
W_4	current assets / short-term liabilities	W_{36}	total sales / total assets
W_5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	W_{37}	(current assets - inventories) / long-term liabilities
W_6	retained earnings / total assets	W_{38}	constant capital / total assets
W_7	EBIT / total assets	W_{39}	profit on sales / sales
W_8	book value of equity / total liabilities	W_{40}	(current assets - inventory - receivables) / short-term liabilities
W_9	sales / total assets	W_{41}	total liabilities / ((profit on operating activities + depreciation) * (12/365))
W_{10}	equity / total assets	W_{42}	profit on operating activities / sales
W_{11}	(gross profit + extraordinary items + financial expenses) / total assets	W_{43}	rotation receivables + inventory turnover in days
W_{12}	gross profit / short-term liabilities	W_{44}	(receivables * 365) / sales
W_{13}	(gross profit + depreciation) / sales	W_{45}	net profit / inventory
W_{14}	(gross profit + interest) / total assets	W_{46}	(current assets - inventory) / short-term liabilities
W_{15}	(total liabilities * 365) / (gross profit + depreciation)	W_{47}	(inventory * 365) / cost of products sold
W_{16}	(gross profit + depreciation) / total liabilities	W_{48}	EBITDA (profit on operating activities - depreciation) / total assets

Table 1. Financial ratios (cont.)

Ratio	Description	Ratio	Description
W_{17}	total assets / total liabilities	W_{49}	EBITDA (profit on operating activities - depreciation) / sales
W_{18}	gross profit / total assets	W_{50}	current assets / total liabilities
W_{19}	gross profit / sales	W_{51}	short-term liabilities / total assets
W_{20}	(inventory * 365) / sales	W_{52}	(short-term liabilities * 365) / cost of products sold)
W_{21}	sales (n) / sales (n-1)	W_{53}	equity / fixed assets
W_{22}	profit on operating activities / total assets	W_{54}	constant capital / fixed assets
W_{23}	net profit / sales	W_{55}	working capital
W_{24}	gross profit (in 3 years) / total assets	W_{56}	(sales - cost of products sold) / sales
W_{25}	(equity - share capital) / total assets	W_{57}	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
W_{26}	(net profit + depreciation) / total liabilities	W_{58}	total costs / total sales
W_{27}	profit on operating activities / financial expenses	W_{59}	long-term liabilities / equity
W_{28}	working capital / fixed assets	W_{60}	sales / inventory
W_{29}	logarithm of total assets	W_{61}	sales / receivables
W_{30}	(total liabilities - cash) / sales	W_{62}	(short-term liabilities *365) / sales
W_{31}	(gross profit + interest) / sales	W_{63}	sales / short-term liabilities
W_{32}	(current liabilities * 365) / cost of products sold	W_{64}	sales / fixed assets

Source: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data.

2. Data and method

The data used in this study are derived from the Emerging Markets Information Service (https://www.emis.com/pl). The empirical research was carried out using the five databases available at, https://archive.ics.uci.edu/ml/datasets/Polish+companies+ bankruptcy+data.

The research is primarily concerned with the companies operating in the industrial processing sector in Poland. A total of 64 financial ratios were used (Table 1). The research covers the period from 2000 to 2013.

The time horizons of one, two, three, four and five years were used to predict company bankruptcy. A prediction horizon of one or two years is typically adopted in the literature on bankruptcy prediction. However, in some cases the three-year time horizon is used. A time horizon of four or five years is also possible but is

rarely applied; this is due to the dynamic nature of the immediate or more distant business environment. An example of such research is the work by Zięba et al. (2016), which adopted a time horizon of one to five years. Due to the fact that the above work has inspired us to undertake this study, we have decided to analyse all five databases (i.e. five prediction horizons).

Machine learning methods are used in a number of research areas (e.g. Friedman et al., 2000). Two major factors determining the suitability of a machine learning method for predicting various developments are: the application of statistical methods for detecting and modelling the existing links between complex phenomena and the use of calculation algorithms designed for large data sets. One example of the machine learning method is gradient tree boosting (GTB) (Friedman, 2001). The gradient tree boosting method is also known as gradient boosting machine (GBM) or gradient boosted regression tree (GBRT). For the purposes of this research, we have adopted the extreme gradient boosting (XGBoost) method (Chen and Guestrin, 2016). XGBoost is a GTB algorithm, which is particularly useful in analysing large data sets. The innovative nature of the XGBoost method - in comparison with other tree boosting algorithms - lies in its use of a novel sparsity-aware algorithm for parallel tree learning.

The necessary calculations for the research work were made using the R software and the 'xgboost' package (Chen and Guestrin, 2016).

3. Research method description

The first step of our empirical research was to split the input data set into a training set and a test set. To ensure generalisability of the research findings, the splitting operation was repeated 30 times. The next step was to remove the outliers reported for companies which continue to operate as a going concern from the training set. The outliers were removed using quantiles (Pawełek et al., 2017; Wu et al., 2010). Then, the extreme gradient boosting method was applied, followed by calculations concerning two model selection criteria to compare the results obtained. The *error* criterion is used to verify the share of wrongly classified objects in the total number of objects, while the *AUC* criterion stands for *Area under the ROC Curve*. Once developed, the models were assessed in terms of their prediction accuracy (*Accuracy* – the share of correctly classified companies in the total number of objects, *Sensitivity* – the share of correctly classified bankrupts in the total number of bankrupts, *Specificity* – the share of correctly classified non-bankrupts in the total number of non-bankrupts). The final step was to verify the hypothesis concerning the location parameters for the populations from which the prediction accuracy measure concerned had been derived, as calculated for models developed based on outlier-free training sets.

Phases of research:

1) Random division of the set X^h, where $h = 1, 2, 3, 4, 5$, into the training set U^h and the test set T^h ($X^h = U^h \cup T^h$, where $\overline{\overline{U^h}} = \frac{2}{3}\overline{\overline{X^h}}$ and $\overline{\overline{T^h}} = \frac{1}{3}\overline{\overline{X^h}}$), while preserving the current structure to take account of bankrupt (B) and non-bankrupt (NB) companies. The splitting operation was repeated 30 times.

2) Providing four variants for each set U^h $(h = 1, 2, 3, 4, 5)$ using the quantiles Q_q and Q_{1-q}, where:

   a) $q = 0.00$ (i.e. outliers are not removed from the set),

   b) $q = 0.01$,

   c) $q = 0.05$,

   d) $q = 0.10$.

   If:

   $U^h = U^{h,B} \cup U^{h,NB}$ $(h = 1, 2, 3, 4, 5)$,

   where:

   $U^{h,B} = \{u_i^{h,B}: u_i^{h,B} \in U^h$ i $u_i^{h,B}$ is bankrupt$\}$,

   $U^{h,NB} = \{u_i^{h,NB}: u_i^{h,NB} \in U^h$ i $u_i^{h,NB}$ is not bankrupt$\}$.

   The following conversion formula is used:

$$\forall j = 1, 2, \dots, 64: u_{ij}^{h,NB} = \begin{cases} u_{ij}^{h,NB} \text{ if } Q_q^j \le u_{ij}^{h,NB} \le Q_{1-q}^j \\ Q_q^j \quad \text{if} \quad u_{ij}^{h,NB} < Q_q^j, \\ Q_{1-q}^j \text{ if} \quad Q_{1-q}^j < u_{ij}^{h,NB} \end{cases}$$

   the result of which is:

   $U_{0.00}^h$, $U_{0.01}^h$, $U_{0.05}^h$, $U_{0.10}^h$ $(h = 1, 2, 3, 4, 5)$.

3) Developing two models per training set U_q^h $(q = 0.00, 0.01, 0.05, 0.10; h = 1, 2, 3, 4, 5)$:

   – establishing the number of iterations (from 1 to 100) against the two criteria: *error* and *AUC*, which are used to measure the classification accuracy of the model based on the training set by means of cross-validation $(v = 3)$; the result is $M_q^{h,error}$ and $M_q^{h,auc}$ $(q = 0.00, 0.01, 0.05, 0.10;\ h = 1, 2, 3, 4, 5)$; Figures 1 and 2 show trends in *error* and *AUC* criteria values for the sample training set $U_{0.00}^1$, while Figures 3-5 show trends in *Accuracy*, *Sensitivity* and *Specificity* values for the sample test set $T_{0.00}^1$ (case of the *error* criterion);

– assessing the prediction accuracy of the developed models on the basis of the test set T^h: $M_q^{h,error}(T^h)$ and $M_q^{h,auc}(T^h)$ $(q = 0.00, 0.01, 0.05, 0.10; h = 1, 2, 3, 4, 5)$, using the following measures: *Accuracy*, *Sensitivity* and *Specificity*.

4) Verifying the hypothesis that the prediction accuracy measure concerned (*Accuracy*, *Sensitivity* or *Specificity*) – as calculated for models developed on the basis of outlier-free training sets $M_q^{h,error}(T^h)$ and $M_q^{h,auc}(T^h)$ $(q = 0.00, 0.01, 0.05, 0.10;\ h = 1, 2, 3, 4, 5)$ – originated from the populations with the same location parameters.

We used the following tests:

- the Kruskal-Wallis test,
- Dunn's post hoc test (the version including the Bonferroni correction for multiple testing).

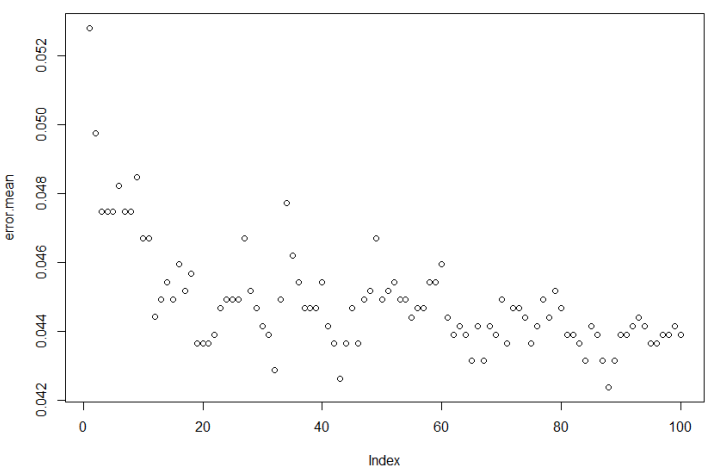

Figure 1. *Error* criteria values for training set $\mathrm{U}_{0.00}^{1}$

Source: Own work.

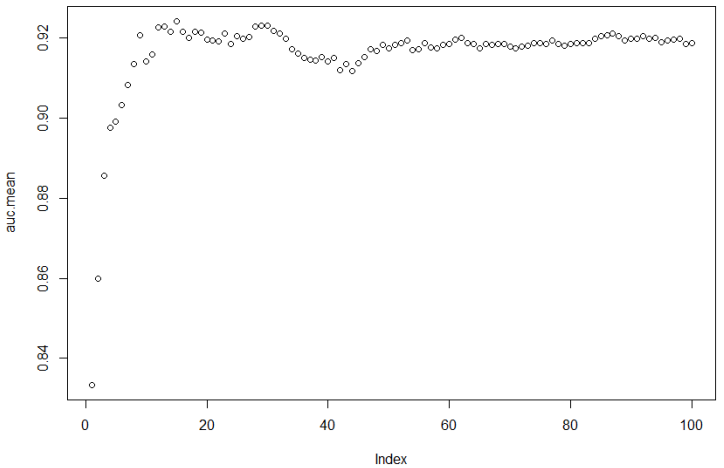

Figure 2. *AUC* criteria values for training set $U_{0.00}^{1}$
Source: Own work.

Figure 1 indicates that the rise in the number of iterations is accompanied by a corresponding decrease in *error* values, which fluctuate around 0.044, whereas the *AUC* criterion value (Figure 2) - initially on the rise - is finally stabilised around 0.92.

As regards the *error* criterion, we selected a model with the lowest share of wrongly classified objects in the total number of objects. As regards the *AUC* criterion, we selected a model with the largest area under the ROC curve.

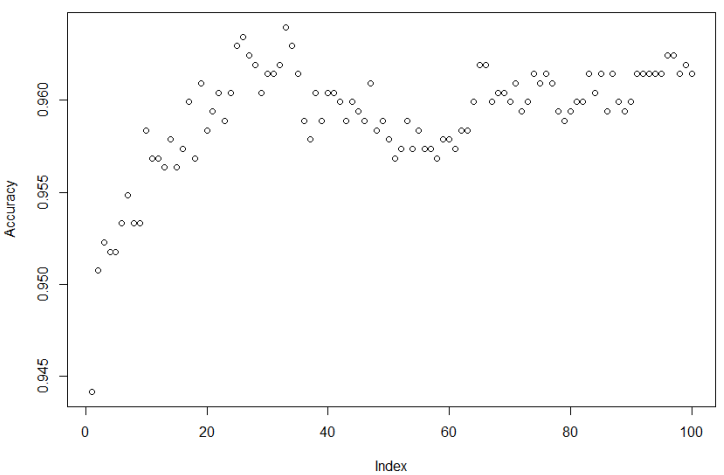

Figure 3. *Accuracy* measure for test set $T_{0.00}^{1}$ (case of the *error* criterion)
Source: Own work.

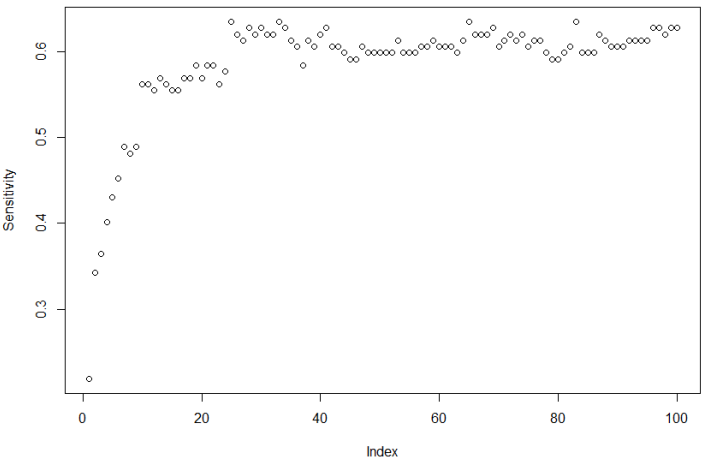

Figure 4. *Sensitivity* measure for test set $T_{0.00}^{1}$ (case of the *error* criterion)
Source: Own work.

The process of model selection may also be based on such prediction accuracy measures as *Accuracy*, *Sensitivity* and *Specificity*. Figures 3-5 show changes in the value of these measures for the selected test set $T_{0.00}^{1}$ by number of iterations (case of the *error* criterion). An analysis of the charts indicates that the *Accuracy* and *Specificity* measures for the test set under consideration stabilised around 0.960 and 0.987 after approx. 70 iterations, whereas the *Sensitivity* measure values fluctuated around 0.60 after approx. 40 iterations.

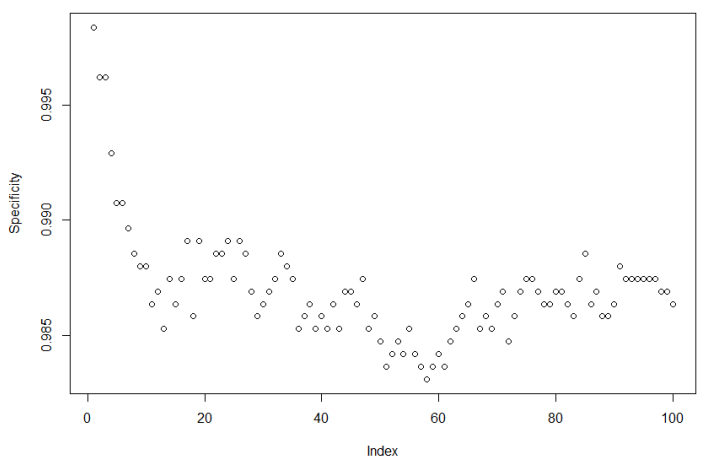

Figure 5. *Specificity* measure for test set $T_{0.00}^{1}$ (case of the *error* criterion)
Source: Own work.

The following section of this document describes the results of empirical research for the two criteria: *error* and *AUC*. For further research purposes, other criteria may also be used for comparison purposes.

4. Empirical results

Table 2 contains results of the Dunn-Bonferroni post hoc test $(\alpha = 0.05)$ carried out on the pairs of research approaches based on data free from outliers which have been removed using the quantiles $\{Q_{q_1}, Q_{1-q_1}\}$ and $\{Q_{q_2}, Q_{1-q_2}\}$, where $q_1, q_2 = 0.00, 0.01, 0.05, 0.10$. Due to the fact that the results obtained for all prediction horizons under consideration $(h = 1, 2, 3, 4, 5)$ are the same, they were put in a single table.

Table 2. Results of the Dunn-Bonferroni post hoc test $(\alpha = 0.05)$ carried out on the pairs of research approached based on data free from outliers which have been removed using the quantiles $\{Q_{q_1}, Q_{1-q_1}\}$ and $\{Q_{q_2}, Q_{1-q_2}\}$ for $h = 1, 2, 3, 4, 5$

	q_1:q_2					
Measure	**0.00:0.01**	**0.00:0.05**	**0.00:0.10**	**0.01:0.05**	**0.01:0.10**	**0.05:0.10**
Criterion: *error*						
Accuracy	S	S	S	S	S	S
Sensitivity	NS	S	S	S	S	S
Specificity	S	S	S	S	S	S
Criterion: *AUC*						
Accuracy	S	S	S	S	S	S
Sensitivity	NS	S	S	S	S	S
Specificity	S	S	S	S	S	S

Significant (S) or non-significant (NS) at $\alpha = 0.05$.

Source: Own work.

At a significant level of 0.05, the test results obtained in most cases under consideration indicate that there is a statistically significant difference between the location parameters for the populations from which relevant prediction accuracy measures are derived, obtained as a result of different variants having been adopted for the training sets $U^h_{0.00}$, $U^h_{0.01}$, $U^h_{0.05}$, $U^h_{0.10}$ $(h = 1, 2, 3, 4, 5)$. The results obtained can be interpreted to mean that the proposed removal of outliers from the data sets has a statistically significant impact on the accuracy of the XGBoost method in predicting company bankruptcy.

An exception to this rule is the *Sensitivity* measure for the pairs of research approaches based on data free from outliers which have been removed using the quantiles $\{Q_{0.00}, Q_{1.00}\}$ and $\{Q_{0.01}, Q_{0.99}\}$. In this case (significant level of 0.05), the removal of outliers from data in the training set has not affected the accuracy of the XGBoost prediction method in a statistically significant way.

To verify that changes in the prediction accuracy of the XGBoost method, occurring as a result of outliers being removed from data sets, constitute a positive trend in bankruptcy prediction, Table 3 provides aggregate information on the arithmetic mean, standard deviation and median values for such measures as *Accuracy*, *Sensitivity* and *Specificity*. For example, Figures 6-8 show results assuming the one-year prediction horizon (prior to bankruptcy).

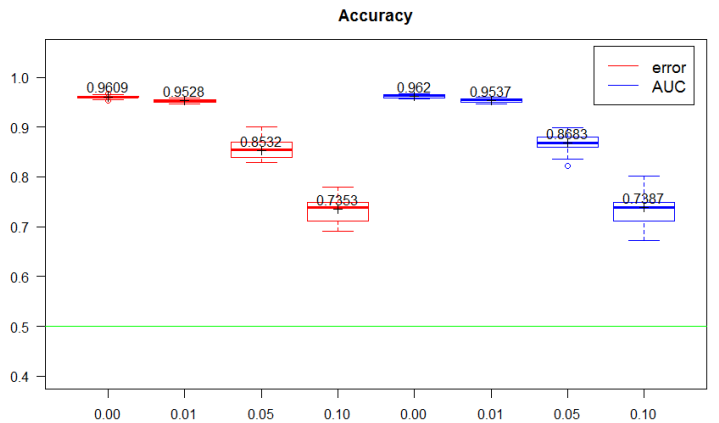

Figure 6. *Accuracy* values obtained assuming the one-year prediction horizon (prior to bankruptcy)

Source: Own work.

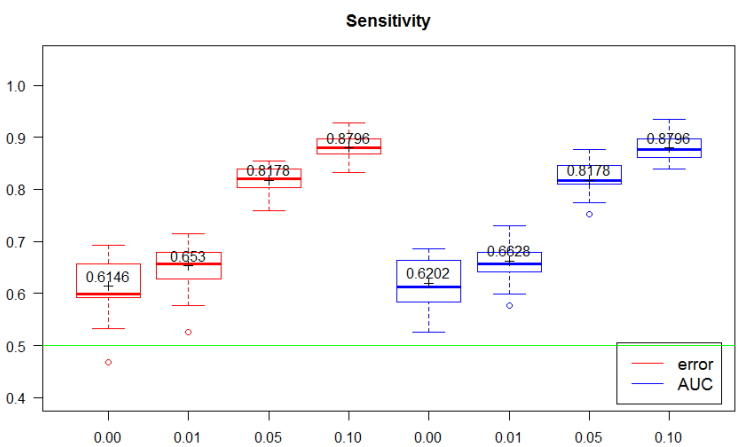

Figure 7. *Sensitivity* values obtained assuming the one-year prediction horizon (prior to bankruptcy)

Source: Own work.

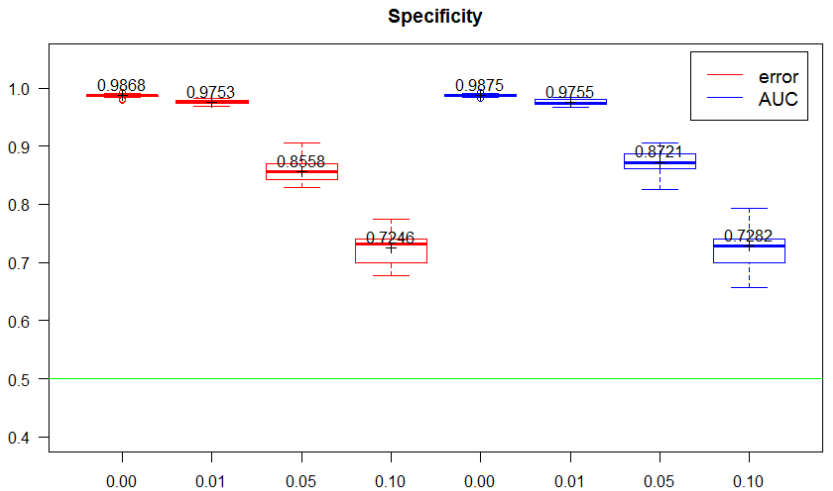

Figure 8. *Specificity* values obtained assuming the one-year prediction horizon (prior to bankruptcy)

Source: Own work.

Figures 6-8 indicate that for the one-year prediction horizon an increase in *q* parameter value (which indicates the level of quantiles used to eliminate outliers from data) is accompanied by corresponding decreases in *Accuracy* and *Specificity* values and increases in *Sensitivity* values. The observed increase in the average value of the *Sensitivity* measure, i.e. from 0.6146 $(q = 0.00)$ to 0.8796 $(q = 0.10)$ (case of the *error* criterion) and from 0.6202 $(q = 0.00)$ to 0.8796 $(q = 0.10)$ (case of the *AUC* criterion), is a positive development, whereas the reported decrease in *Specificity* values, i.e. from 0.9868 $(q = 0.00)$ to 0.7246 $(q = 0.10)$ (case of the *error* criterion) and from 0.9875 $(q = 0.00)$ to 0.7282 $(q = 0.10)$ (case of the *AUC* criterion), is a negative development. From the perspective of credit granting institutions, which commission bankruptcy prediction models, the accuracy of bankruptcy risk prediction is more important than the prediction accuracy for absence of bankruptcy risks. Given the above and the results obtained, the use of quantiles $\{Q_{0.05}, Q_{0.95}\}$ during removal of outliers from the data sets is the most preferred option. The *Sensitivity* measure rose to 0.8178 (case of the *error* criterion) and 0.8178 (case of the *AUC* criterion), while the *Specificity* measure fell only slightly, i.e. to 0.8558 (case of the *error* criterion) and 0.8721 (case of the *AUC* criterion).

The results presented in Tables 3-7 show that the patterns observed for the prediction model with the one-year time horizon can also be identified for the results obtained using the other four databases. The trends reported for the arithmetic mean values of the prediction accuracy measures are also identified for the median values.

Table 3. Arithmetic mean, standard deviation and median values for *Accuracy*, *Sensitivity* and *Specificity* measures (h = 1)

Measure	q	Criterion: *error*			Criterion: *AUC*		
		Average	Standard deviation	Median	Average	Standard deviation	Median
Accuracy	0.00	0.9609	0.0032	0.9609	0.9620	0.0032	0.9629
	0.01	0.9528	0.0036	0.9530	0.9537	0.0034	0.9548
	0.05	0.8532	0.0173	0.8541	0.8683	0.0165	0.8680
	0.10	0.7353	0.0208	0.7381	0.7387	0.0315	0.7391
Sensitivity	0.00	0.6146	0.0487	0.5985	0.6202	0.0431	0.6131
	0.01	0.6530	0.0426	0.6569	0.6628	0.0365	0.6569
	0.05	0.8178	0.0279	0.8212	0.8178	0.0286	0.8175
	0.10	0.8796	0.0265	0.8796	0.8796	0.0245	0.8759
Specificity	0.00	0.9868	0.0026	0.9869	0.9875	0.0023	0.9875
	0.01	0.9753	0.0036	0.9749	0.9755	0.0045	0.9733
	0.05	0.8558	0.0188	0.8560	0.8721	0.0176	0.8723
	0.10	0.7246	0.0226	0.7310	0.7282	0.0332	0.7278

Source: Own work.

Table 4. Arithmetic mean, standard deviation and median values for *Accuracy*, *Sensitivity* and *Specificity* measures (h = 2)

Measure	q	Criterion: *error*			Criterion: *AUC*		
		Average	Standard deviation	Median	Average	Standard deviation	Median
Accuracy	0.00	0.9645	0.0020	0.9645	0.9641	0.0023	0.9635
	0.01	0.9465	0.0064	0.9459	0.9490	0.0050	0.9505
	0.05	0.7748	0.0241	0.7713	0.7882	0.0202	0.7874
	0.10	0.6175	0.0251	0.6131	0.6269	0.0331	0.6176
Sensitivity	0.00	0.5060	0.0389	0.5000	0.5062	0.0372	0.5058
	0.01	0.5554	0.0507	0.5698	0.5665	0.0433	0.5901
	0.05	0.7556	0.0630	0.7791	0.7653	0.0439	0.7791
	0.10	0.8762	0.0279	0.8779	0.8624	0.0247	0.8605
Specificity	0.00	0.9900	0.0017	0.9897	0.9896	0.0017	0.9897
	0.01	0.9682	0.0055	0.9683	0.9703	0.0045	0.9702
	0.05	0.7758	0.0266	0.7705	0.7895	0.0218	0.7908
	0.10	0.6031	0.0266	0.5983	0.6138	0.0353	0.6025

Source: Own work.

Table 5. Arithmetic mean, standard deviation and median values for *Accuracy*, *Sensitivity* and *Specificity* measures (h = 3)

		Criterion: *error*			Criterion: *AUC*		
Measure	***q***	**Average**	**Standard deviation**	**Median**	**Average**	**Standard deviation**	**Median**
Accuracy	0.00	0.9642	0.0017	0.9640	0.9645	0.0028	0.9646
	0.01	0.9512	0.0042	0.9514	0.9529	0.0027	0.9520
	0.05	0.7675	0.0240	0.7775	0.7876	0.0210	0.7832
	0.10	0.5648	0.0199	0.5701	0.5879	0.0361	0.5920
Sensitivity	0.00	0.3582	0.0978	0.3758	0.4154	0.0439	0.4182
	0.01	0.4317	0.0379	0.4303	0.4564	0.0555	0.4303
	0.05	0.6919	0.0165	0.6909	0.6846	0.0193	0.6848
	0.10	0.8220	0.0204	0.8182	0.8265	0.0187	0.8242
Specificity	0.00	0.9941	0.0038	0.9922	0.9917	0.0021	0.9912
	0.01	0.9768	0.0051	0.9769	0.9775	0.0024	0.9781
	0.05	0.7713	0.0255	0.7830	0.7927	0.0221	0.7891
	0.10	0.5520	0.0205	0.5579	0.5761	0.0381	0.5808

Source: Own work.

Table 6. Arithmetic mean, standard deviation and median values for *Accuracy*, *Sensitivity* and *Specificity* measures (h = 4)

		Criterion: *error*			Criterion: *AUC*		
Measure	***q***	**Average**	**Standard deviation**	**Median**	**Average**	**Standard deviation**	**Median**
Accuracy	0.00	0.9730	0.0019	0.9732	0.9727	0.0021	0.9729
	0.01	0.9643	0.0031	0.9653	0.9647	0.0033	0.9655
	0.05	0.7810	0.0276	0.7846	0.8069	0.0311	0.8178
	0.10	0.5527	0.0306	0.5520	0.5789	0.0478	0.5619
Sensitivity	0.00	0.4341	0.0277	0.4361	0.4368	0.0289	0.4361
	0.01	0.4476	0.0291	0.4511	0.4574	0.0254	0.4586
	0.05	0.6629	0.0536	0.6692	0.6634	0.0463	0.6692
	0.10	0.8253	0.0338	0.8346	0.8183	0.0350	0.8158
Specificity	0.00	0.9950	0.0015	0.9952	0.9946	0.0016	0.9945
	0.01	0.9854	0.0032	0.9865	0.9855	0.0036	0.9863
	0.05	0.7858	0.0293	0.7893	0.8128	0.0325	0.8241
	0.10	0.5416	0.0319	0.5428	0.5691	0.0502	0.5525

Source: Own work.

Table 7. Arithmetic mean, standard deviation and median values for *Accuracy*, *Sensitivity* and *Specificity* measures (h = 5)

Measure	q	Criterion: *error*			Criterion: *AUC*		
		Average	Standard deviation	Median	Average	Standard deviation	Median
Accuracy	0.00	0.9793	0.0014	0.9791	0.9789	0.0012	0.9795
	0.01	0.9730	0.0021	0.9718	0.9736	0.0021	0.9735
	0.05	0.8898	0.0177	0.8915	0.9018	0.0171	0.9065
	0.10	0.7350	0.0234	0.7400	0.7476	0.0191	0.7447
Sensitivity	0.00	0.5411	0.0465	0.5556	0.5426	0.0392	0.5556
	0.01	0.5715	0.0136	0.5667	0.5707	0.0169	0.5778
	0.05	0.6756	0.0520	0.6889	0.6711	0.0421	0.6778
	0.10	0.7967	0.0349	0.8222	0.7911	0.0402	0.8000
Specificity	0.00	0.9968	<0.0001	0.9969	0.9964	<0.0001	0.9964
	0.01	0.9890	0.0021	0.9880	0.9897	0.0019	0.9893
	0.05	0.8984	0.0188	0.9028	0.9111	0.0168	0.9156
	0.10	0.7326	0.0249	0.7393	0.7459	0.0211	0.7411

Source: Own work.

In the case of standard deviation values, no pattern that would be common to all cases under consideration has been observed. It is often the case that an increase in q parameter values is accompanied by corresponding increases in standard deviation values for *Accuracy* and *Specificity* measures. Further research is needed to assess how the elimination of outliers from data sets could affect measures other than the arithmetic mean and median of the data set.

5. Conclusions

The above considerations can be summed up by stating that in most cases where the significant level was set at 0.05, an analysis of the prediction accuracy measures resulted in the null hypothesis being rejected in favour of the alternative hypothesis stating that the resulting data sets were not derived from the populations with the same location parameters. An exception to this rule is the pair of research approaches based on the training sets free from outliers which have been removed using the quantiles $\{Q_{0.00}, Q_{1.00}\}$ and $\{Q_{0.01}, Q_{0.99}\}$ for the *Sensitivity* measure.

It can therefore be concluded that the removal of the outliers reported for companies which continue to operate as a going concern from data sets affects the accuracy of the extreme gradient boosting method in predicting company bankruptcy.

The results show that the use of quantiles for the removal of the outliers reported for companies which continue to operate as a going concern from training sets increases the accuracy of the extreme gradient boosting method in detecting bankrupt companies (*Sensitivity*), while reducing the prediction

accuracy of that method when measured as total (*Accuracy*) and for a group of non-bankrupt companies (*Specificity*). In addition, the following pattern was observed: the more the accuracy is affected, the higher the *q* parameter $\left(Q_q \text{ and } Q_{1-q}\right)$.

The results of the empirical research are consistent with the statement that the longer the prediction horizon (*h*), the less accurate the bankruptcy detection model (*Sensitivity*).

Among the quantiles examined, the pair $Q_{0.05}$ and $Q_{0.95}$ should be highlighted due to (among others) the fact that when it was used to remove the outliers reported for companies which continue to operate as a going concern from the training set, the average value of the *Sensitivity* measure for *h* = 3, 4 rose above 0.50.

Acknowledgements

Publication was financed from the funds granted to the Faculty of Management at Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

REFERENCES

BAESENS, B., VAN GESTEL, T., VIAENE, S., STEPANOVA, M., SUYKENS, J., VANTHIE, J., (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54, pp. 627–635. DOI: http://dx.doi.org/10.1057/palgrave.jors.2601545.

BROWN, I., MUES, Ch., (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39 (3), pp. 3446–3453. DOI: https://doi.org/10.1016/j.eswa.2011.09.033.

CHEN, T., GUESTRIN, C., (2016). XGBoost: A Scalable Tree Boosting System. DOI: http://dx.doi.org/10.1145/2939672.2939785.

GARCIA, V., MARQUES, A. I., SANCHEZ, J. S., (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. Information Fusion, 47, pp. 88–101. DOI: https://doi.org/10.1016/j.inffus.2018.07.004.

FRIEDMAN, J. H., (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29 (5), pp. 1189–1232.

FRIEDMAN, J. H., HASTIE, T., TIBSHIRANI, R., (2000). Additive logistic regression: a statistical view of boosting. The Annals of Statistics, 28 (2), pp. 337–407.

LESSMANN, S., BAESENS, B., SEOW, H. V., THOMAS, L. C., (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. European Journal of Operational Research, 247 (1), pp. 124–136. DOI: https://doi.org/10.1016/j.ejor.2015.05.030.

NANNI, L., LUMINI, A., (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications, 36, pp. 3028–3033.
DOI: http://dx.doi.org/10.1016/j.eswa.2008.01.018.

PAWEŁEK, B., GAŁUSZKA, K., KOSTRZEWSKA, J., KOSTRZEWSKI, M., (2017). Classification Methods in the Research on the Financial Standing of Construction Enterprises After Bankruptcy in Poland. In: Palumbo, F. et al. (Eds.), Data Science, Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Switzerland, pp. 29–42.
DOI: http://dx.doi.org/10.1007/978-3-319-55723-6_3.

PAWEŁEK, B., (2017). Prediction of Company Bankruptcy in the Context of Changes in the Economic Situation. In: Papież, M., Śmiech, S. (Eds.), The 10th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Conference Proceedings. Cracow: Foundation of the Cracow University of Economics, pp. 290–299.

WU, Y., GAUNT, C., GRAY, S., (2010). A comparison of alternative bankruptcy prediction models. Journal of Contemporary Accounting & Economics, 6, pp. 34–45. DOI: http://dx.doi.org/10.1016/j.jcae.2010.04.002.

XIA, Y., LIU, Ch., LI, Y., LIU, N., (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems With Applications, 78, pp. 225–241.

ZIĘBA, M., TOMCZAK, S. K., TOMCZAK, J. M., (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems With Applications, 58, pp. 93–101.
DOI: http://dx.doi.org/10.1016/j.eswa.2016.04.001.