Bank i Kredyt 49(1), 2018, 45-62

BANK I KREDYT
BANK & CREDIT

The importance of financial and non-financial ratios in SMEs bankruptcy prediction

Aneta Ptak-Chmielewska*, Anna Matuszyk#

Submitted: 17 October 2017. Accepted: 15 November 2017.

Abstract

Credit risk is considered to be a key risk in banking activity. Statistical and data mining bankruptcy prediction models can be used in assessing the credit risk of enterprises. In the case of small and medium enterprises, qualitative factors are as important as financial ones. In this paper those financial ratios and qualitative factors that are the most frequently used in assessing bankruptcy prediction of small and medium enterprises were discussed. They were analysed and assessed with the use of data mining techniques, and they were also considered from the point of view of their inclusion in the bankruptcy prediction model.

Keywords: bankruptcy prediction, non-financial ratios, random forests

JEL: G17, C1, C5, R15

* Warsaw School of Economics; e-mail: aptak@sgh.waw.pl.
# Warsaw School of Economics; e-mail: amatus1@sgh.waw.pl.

1 Introduction

Small and medium enterprises (SMEs) play an important role in the economy. The SMEs sector represents around 99% of enterprises in the European Union. However, the position of this sector varies among European countries. The bankruptcy risk of SMEs is difficult to predict mainly due to the lack of credible databases (Yoon, Kwon 2010). In this study we carry out the analysis of data acquired from a consulting company operating on the Polish market in order to identify both financial and non-financial factors that can be used to assess the probability of a firm bankruptcy.

The aim of this paper is to contribute to the research associated with the determinants of SMEs' bankruptcy prediction making use of data mining techniques namely: decision trees (DT), random forests (RF), neural network (NN) and logistic regression (LR). We want to check whether RF outperforms the other methods. All models built were compared in terms of prediction accuracy as measured by AUC and the misclassification rate.

The input data contain financial ratios calculated on the basis of the companies' financial statements and non-financial ratios kindly provided by one of the consulting companies. The dataset consists of 435 bankrupt and 533 non-bankrupt companies operating on the Polish market. Bankruptcy was considered as legal act only, the economic bankruptcy was not considered. The analysis was conducted using the RF approach. By way of comparison logistic regression and other data mining models were used. Furthermore, the inclusion of non-financial factors should provide wider view on the determinants influencing companies' bankruptcy risk.

The remainder of the paper is structured as follows. The next section briefly reviews the previous empirical research on the SMEs bankruptcy models. The subsequent section describes the random forest methodology and the data set used in our research. This is followed by the description of results. The final section concludes.

1 Literature review

Since the announcement of the Altman's Z-Score model (Altman 1968), a large number of statistical bankruptcy prediction studies were written using the traditional methods, like discriminant analysis (Back et al. 1996), logistic regression (Aziz, Dar 2004; Back et al. 1996) and probit analysis (Zmijewski 1984). Recent studies in this area focus on more advanced and sophisticated methods, like case-based reasoning (Bryant 1997; Yip 2006; Sartori, Mazzucchelli, Di Gregorio 2016), genetic algorithms (Back et al. 1996) and artificial neural networks (Wilson, Sharda 1994).

Sartori, Mazzucchelli and Di Gregorio (2016) applied the case-based reasoning (CBR) paradigm to forecast the bankruptcy and compared the results received with the Z-Score model. The CBR method turned out to be good in predicting bankruptcy. The authors found that this approach could be useful to cluster enterprises according to opportune similarity metrics.

Genetic algorithms (GAs) was another method used in SMEs default prediction analysis. Gordini (2014) compared the potential of genetic algorithms with two other methods: logistic regression (LR) and support vector machine (SVM). The results obtained suggest that GAs are a very effective and promising method in assessing the probability of SMEs bankruptcy compared with LR and SVM, especially in reducing type II misclassification rate. The author also investigated whether the size of

firms and the geographical area of their operation can influence the accuracy of the models and, again, the results obtained from separate models built to custom for separate geographical areas show that GAs prediction accuracy in each area is superior to that of the other models.

Sohn, Kim and Yoon (2016) proposed an approach based on fuzzy logistic regression that can be used in the default prediction models. Moreover, the authors showed that the proposed approach outperforms the logistic regression model in terms of discriminatory power. Similarly, Chaudhuri and De (2011) used the fuzzy support vector machine which outperformed traditional bankruptcy prediction methods.

Traditional analysis of company financial condition is based on financial factors. However, it is worth considering whether other indicators can be significant. This problem was addressed by few researchers. Jiménez and Saurina (2004) discussed the role of a limited set of variables, namely: collateral, type of lender and bank-borrower relationship. According to their results, collateralised loans have higher probability of default and loans granted by savings banks are riskier. Additionally, authors found that a close relationship between the bank and the customer increases the willingness to take more risk.

One of the first who observed improving forecasting performance by including industry groupings in their models were Chava and Jarrow (2004). Using a dataset covering over 30 years (1962–1999) they built bankruptcy hazard rate models for US companies. According to their results, there are industry effects in hazard rate estimation. Industry groupings significantly affected the intercept and slope coefficients in the forecasting equations.

Psillaki, Tsolas and Margaritis (2010) showed that non-financial performance indicators are useful *ex ante* determinants of business failure. Using the companies' datasets from three different French manufacturing industries they proved that managerial inefficiencies are an important *ex ante* indicator of a firm's financial risk. The results suggest that more efficient firms as well as firms with more liquid assets are less likely to fail.

A similar approach was taken by Fabling and Grimes (2005), who used regional as well as national data. They analysed the role of property prices, which influenced the collateral values. According to the authors' findings the interactions between economic activity, leverage and property price (collateral) shocks indicate that region-specific shocks can compound into significant localised economic cycles.

El Kalak and Hudson (2016) investigated how company size can affect bankruptcy probability for SMEs. The authors used a data set of 11,117 US non-financial firms of which 465 filed for insolvency between 1980 and 2013 and built discrete-time duration-dependent hazard models for 4 groups of companies separately, namely: SMEs, micro, small, and medium firms. They found micro and small firms should be considered separately when modelling the credit risk.

The majority of models built for predicting bankruptcy of Polish firms were based on the linear discriminant analysis, including Mączyńska (1994), Pogodzińska and Sojak (1995), Gajdka and Stos (1996), Wierzba (2000), Holda (2001), Sojak and Stawicki (2001) and Prusak (2005).

The literature suggests that there are many advantages of using other than traditionally accepted methods. The same applies to the financial ratios in building bankruptcy risk models for companies. Two important groups of factors can additionally be considered, namely non-financial and macroeconomic variables. A part of our project will be devoted exclusively to checking the benefits of usage of such factors. We want to check their influence on the model prediction accuracy. We assume that the usage of these additional factors improves the results of the models. Moreover, the usage of macro variables

(that are time-dependent) will also give the dynamic view of the firm's situation, providing us with some additional findings. The traditional approach in modelling the bankruptcy of companies was mainly based on simple discriminate analysis and financial factors. Our contribution to the literature consists in bankruptcy models with non-financial factors, such as size, age, region and the legal form of the company. In our modelling approach we are trying a more sophisticated statistical method, namely random forests, and compare it with logistic regression, the decision tree and the neural network.

2 Methodology

Random forests enhanced by a boosting technique was applied in the modelling phase. As a comparison logistic regression (stepwise selection), decision tree (recursive division algorithm) and neural network (multi-layer perceptron) were chosen. In the first step the hierarchical clustering was applied in order to transform the nominal variable regions (voivodeships) into homogenous groups.

Decision tree is a tool mainly used in hierarchical segmentation (division) of the data set. The main element is the so-called root that includes the entire data set. Subsequent splits of the data (observations) are carried out in the so-called nodes, or segments, according to the rules created on the basis of the values of explanatory variables. A segment that is subdivided into subsegments is being referred to as the parent node (or intermediate node) and the subsegments as children nodes. The tree branch creates a node with further subsegments. A leaf (group) is the final segment that is no longer divided. Each observation from the output node is assigned to only one final leaf. The decision tree contains intermediate and final nodes, while the decision tree model contains only the final leaves that are used to predict or classify data (see Graph 1).

In order for decision trees to be used, a large collection of observations is required as well as sufficiently numerous cases for the dependent variable. Any (very) unusual observations may distort the results, though this is not a major risk.

A big risk in building the tree is overfitting, which can cause instability of the model. The decision tree, unlikely the binary logistic regression, does not contain any equations or coefficients, it is based only on the rules of dividing the dataset into separate groups. As estimation of probabilities posteriori probabilities for each leaf are used. The rules generated by the model from the learning set can be used for prediction (resulting in binary decisions).

The basic ways to measure the quality of the division for binary dependent variables or discrete dependent variables with several categories include:

1 The degree of separation achieved by the division (measured by the Pearson Chi-squared test),

2 The degree of pollution reduction achieved by the division (measured by the reduction of entropy or by the Gini coefficient).

The stopping criteria may be the following: the minimal number of observations in any final leaf, the critical size of any node, the number of splits in any path. After building a tree, it should be pruned into an optimal size. The advantages of a decision tree are twofold: the results are easily interpretable and the model is flexible. Besides, decision trees are not sensitive to missing data and do not require the normality of distributions or the equality of covariance matrices (as discriminant analysis does). The explanatory variables may differ in character, being either qualitative or quantitative. Decision trees automatically select important variables and may explain non-linear dependencies.

The disadvantage of the decision trees is the fact that they can prove unstable and sensitive to the size of the training sample, validation or test sample results. A big size of the training sample is critical. Probabilities are approximated on the final leaf level. Overtraining is quite common in decision trees and the results for the training sample are usually much better than for the testing sample. All those disadvantages must be considered while building a model.

Nowadays, a more popular method is the random forest, initially proposed by Breiman. It is a method that takes together many classification trees. Firstly, we draw K bootstrap samples, then we create a classification tree for each of them in such a way that in each node we draw m (fewer than the number of all features) features which will participate in the selection of the best division. Trees are built without pruning. Finally, the observations are classified by the voting method. The only parameter of the method is the m coefficient, which should be much smaller than the dimension of data $p(m = \sqrt{p})$. The ease and speed with which random forests can be created makes them a feasible option even for very large data.

Random forests are currently one of the most efficient classification methods, apart from the SVM and boosting. The boosting method makes it possible to cope with an opposite situation: it allows to aggregate many stable but less efficient classifiers (weak learners). The classification abilities of weak learners are small – the probability of correct classification slightly exceeds 1/2. The main idea is that in the iteration process the observations should be assigned weights which suggest to weak learners on which examples they should concentrate in their next approach to the classification task. The final decision regarding the classification of observations is made in majority voting. The main feature of boosting is the ability to decrease the training error: a group of weak learners acts together as a single good learner. What is more important, the error decreases exponentially, which is very important in practical usage. An additional advantage is that the boosting algorithms are not subject to overfitting.

The logistic regression model is the third analysed approach. The general form of the logistic regression model is as follows:

$$Y \sim B(1, p)$$
$$p = E\left(Y|X\right) = \frac{\exp\left(\beta X\right)}{1 + \exp\left(\beta X\right)}$$

where:

$B(1, p)$ – the binomial distribution with the probability of success p,

Y – the dependent variable,

$X = (X_1, \ldots, X_k)$ – the independent explanatory variables,

β – structural parameters.

The greatest advantage of this model is that it provides the probability estimation of bankruptcy at the level of each observation. Unfortunately, this model makes some assumptions and suffers from certain restrictions. The model needs to meet the following assumptions: sampling randomness, numerous observations, the lack of correlations of the explanatory variables, observation independence.

The neural network, i.e. the fourth analysed method, is formed by the neurons (information processing elements) along with the connections among them (weights modified during the learning process) (see Graph 2). This network is a simplified model of the human brain. A neuron contains many inputs x_i, where $i = 1, 2, \ldots, n$ and one output. Neural inputs are selected by the explanatory variables.

When neural networks are used to forecast the risk of bankruptcy, these are typically financial ratios. Each input variable is assigned a specific weight w_i. Once the weights are determined, the total neuron activation (e) is calculated as the sum of the product of the explanatory variables and their weights assigned. Then y is calculated, which is the difference between the value of e and the threshold value Θ. The output signal depends on the neuron activation and the activation function $\phi(y)$. The form of this function determines the neuron type.

In practice, artificial neural networks are usually made of a large number of interconnected neurons. We can distinguish the following neural networks:

– double layer neural networks – consisting of the input and output layers,
– multi-layered neural networks – consisting of input and output layers and hidden layers between them.

In predicting the bankruptcy of enterprises multi-layer perceptron (MLP) neural networks are frequently used. Neural networks are flexible and they quickly adapt to changes. They are resistant to any chaotic information and do not require assumptions like normality etc. The explanatory variables can be both qualitative or quantitative in type. Neural networks enable the modeling of any type of non-linear dependencies in the data.

Unfortunately, neural network models also have significant limitations. The long-term learning process for networks with extensive structures prevents the model from achieving an optimal level of error reduction. The weights selection process is difficult and complex. Neural networks do not select explanatory variables for the model. The analyst conducts a selection of explanatory variables by himself. Similarly as in the case of decision trees, there is a risk of overtraining. Selecting network architecture is a subjective choice. The worst disadvantage of the neural networks approach is the fact that they operate on the "black box" basis – without the ability to provide the rules that resulted in the obtained outcome.

3 Data description and results

The dataset consists of 435 bankrupt and 533 non-bankrupt small and medium (SMEs) companies operating on the Polish market. The sample was selected randomly from the available database with almost equal proportion of bankrupt and non-bankrupt companies. A balanced sample makes the robust estimation of misclassification errors possible. The financial statements covered the period of 2008–2010. The bankruptcy events took place between 2009 and 2012, a 12-month observation period was considered. The data were kindly provided by a consultancy firm operating on the Polish market. Due to security reasons, the provider asked to stay anonymous. It was individual data that were made available and on the basis of these data we found that we were able to calculate financial ratios for one year only. Eventually, we decided to use 16 ratios (statistically significant) for further analysis and they were presented in Table 1. Additionally, we considered five non-financial factors related to the analysed companies, namely: the sector of the company's activity, the region of the activity, company's legal form, year of the company's establishment, the number of employees. Information on the non-financial factors is presented in Table 2. The selection of variables plays an important role (Du Jardin 2009), but financial ratios are the most often selected variables in bankruptcy prediction models. For modelling purposes, the whole data sample was divided into the training and test sample with the proportion of 70% and 30% respectively.

The nominal variable *region* contains 16 categories/regions, the so-called voivodeships. It was clustered with the use of the hierarchical clustering average linkage method and bankruptcy rate for each region. The optimal number of groups (in this case four groups) was selected on the basis of the pseudo t^2 statistic.

The final groups for regions are presented in Graph 3 and Table 3.

To merge the categories for this variable, cluster analysis was used. According to this approach, the grouping takes place on the basis of the smallest average distance between the clusters. The variable used for grouping was the percentage of companies in bankruptcy. Observations were combined in pairs while aggregations were combined according to the smallest average distance between them. Cluster analysis can be used to group the categories of the nominal variable and to reduce the variability and dimensionality of the analysis. The grouping of the categories makes it possible to include a nominal variable with a quasi-complete separation into the logistic regression.

The logistic regression was estimated with the stepwise selection method and at the 0.05 significance level, both for entry and stay. The decision tree was estimated with the recursive partitioning method applying the F test for interval variables, the Chi-square test for nominal variables and entropy for ordinal variables. The p-value for splitting was set at 0.2. The stopping criterion was set at the maximum depth of 6. Random forests were estimated with the gradient boosting algorithm and M Huber rules with maximum branch equal 2 and maximum depth equal 2. Subtree selection was based on the misclassification rate. 50 interactions were trained on the 60% of the training set containing 70% of observations.

Model evaluation was based on the Gini (accuracy ratio – AR) coefficient on the test sample, which is a measure based on ROC, i.e. the curve used to measure the discriminative power of the model. It is applied in the case when the dependent variable is binary (it has two unique values). The chart presents the relation of the specificity to the sensitivity of the model. Both of those measures provide information on how effective the classification is in the context of both levels of the dependent variable. The ROC curve is a sensitivity function (on the vertical axis) and 1-specificity (on the horizontal axis). Each point of the curve corresponds to a given point of split (section). Points in the right upper corner correspond to a low q level. Points in the left bottom corner relate to a high q level. ROC does not depend on the assumed point of split. The rates are drawn for all points of split. While selecting a given point of split we can establish the specificity and sensitivity of the model for that point. Selecting a given point of split we can establish the number of successes and failures predicted by the model, and then calculate the sensitivity and the specificity of the model. The correspondent sensitivity and specificity levels are easy to read from the graph of the ROC curve. A good model has the ROC curve close to the upper-left boundary of the graph. Then we can find points on the curve representing high values both in terms of sensitivity and specificity (e.g. so that $c > 0.8$ and $s > 0.8$). The random model has the ROC curve lying on the diagonal. Then the sensitivity + specificity = 1 for all the threshold values of q. In such a case while establishing the value $c > 0.8$ we cannot ensure that the specificity is bigger than 0.2. The ROC curve is helpful when selecting the optimal point of division. For example, we choose the threshold that gives equal probability of misclassification in each class. We also have to take into account the different cost of both types of misclassification and decide whether to provide high sensitivity or high specificity. The area under the ROC curve is a measure of the quality of the model. This way we can compare the quality of different models. The AUC (area under the ROC curve) for an ideal model equals 1 and for a random model 0.5 (see Graph 4).

The similar measure to ROC is the CAP curve, where the cumulative frequencies for good customers are substituted by frequencies for all customers. The area under the CAP curve is called accuracy ratio. The CAP curve represents the *y*% of bankrupted enterprises that can be found in the *x*% of the worst assessed enterprises within the model. The curve should be concave. The accuracy ratio (Gini coefficient) based on the CAP curve is defined as:

$$AR = \frac{\int_0^1 f(x)dx - \frac{1}{2}}{\frac{1}{2} - \frac{1}{2}BR}$$

where BR – bankruptcy rate; $\int_0^1 f(x)dx - \frac{1}{2}$ – area under the CAP curve. The value of AR is normalized in the range of [0; 1].

Table 4 indicates that the best model, gradient boosting random forests, had the AR amounting to 0.614, which is a satisfactory figure. The misclassification rate was 0.27 for the test sample and 0.19 for the training sample. This model seems to be slightly over trained due to the small sample used in modelling.

The first type error (the wrong classification of defaults) amounted to 0.25 and the second type error (wrong classification of good customers) amounted to 0.15 (see Table 5).

The importance of the variables used in gradient boosting is presented in Table 6. It is possible to evaluate the importance of a given variable in prediction by adding up the weighted impurity decreases for all the nodes where the variable is used averaged over all trees in the forest, but actually, it can be used on a single tree as well.

Variable *region* was not important, it appeared in 3 rules only, more important were the following variables: *legal form*, *employment* and *sector of activity*. Bankruptcies are not diversified interregionally. Ratios with low discriminatory power (AR < 0.1) X8, X9, X10, X11 were not very important in the random forest either. The most significant ratios were as follows:

X16: share of net financial surplus in total liabilities,

X13: capital ratio,

X12: inventory turnover,

X3: liquidity cash,

X15: coverage of fixed assets by equity.

In the logistic regression model only three of the included variables were significant: X13: capital ratio, X7: operating profitability of assets and a non-financial factor called legal form.

For the decision tree important variables were as follows (importance in brackets): X16 (1.00), X11 (0.45), X13 (0.35), X12 (0.27), X9 (0.25), X7 (0.25), X2 (0.21), X10 (0.21) and legal form (0.21), age of the company (0.19), employment (0.19).

Contrary to gradient boosting the results of the decision tree model can be presented in a graphical form. These results were presented on Graph 5. The most important split was the split on enterprises with X16 < 0.068 and X16 ≥ 0.068. Enterprises from the first group with the share of the net financial surplus in total liabilities below 0.068 run a high risk of bankruptcy.

In neural network model all variables were used. Results are not as easily interpretable as in the decision tree model.

4 Concluding remarks

This paper presented the results of the analysis of the impact of non-financial ratios on bankruptcy classification for SMEs. Four methods were used: gradient boosting, logistic regression, decision trees and neural networks, with and without non-financial ratios. In terms of the accuracy ratio, gradient boosting outperforms the remaining three methods. However the difference in accuracy power between the training and testing sample was the smallest for the logistic regression. For gradient boosting the difference amounted to ca. 15 percentage points, other models were overtrained because the difference between the training sample and the testing sample was quite sizeable and amounted to around 25 percentage points.

The usage of non-financial ratios improves the results of all models (see Table 4) which confirmed our expectations and research conducted by Keasey and Watson (1987). The legal form of the company seems to be the most important variable among all the considered non-financial factors. Employment and sector also plays a role, which confirms the results obtained by Chava and Jarrow (2004). Gordini (2004) confirmed that building models tailored to specific geographical areas increases the accuracy. However in our models two variables: region and age of the company seem to play a much less important role.

References

Aziz M.A., Dar H.A. (2004), *Predicting corporate bankruptcy: Whither do we stand?*, Economic Research Papers, 04-01, Loughborough University Institutional Repository, https://dspace.lboro.ac.uk/2134/325.

Back B., Laitinen T., Sere K., van Wezel M. (1996), *Choosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms*, Technical Report, Turku Centre for Computer Science.

Bryant S.M. (1997), A case-based reasoning approach to bankruptcy prediction modelling, *Intelligent Systems in Accounting, Finance and Management*, 6(3), 195–214.

Chaudhuri A., De K. (2011), Fuzzy support vector machine for bankruptcy prediction, *Applied Soft Computing*, 11(2), 2472–2486.

Chava S., Jarrow R.A. (2004), Bankruptcy prediction with industry effects, *Review of Finance*, 8, 537–569.

Du Jardin P. (2009), Bankruptcy prediction models: How to choose the most relevant variables?, *Bankers, Markets & Investors*, 98, 39–46.

El Kalak I., Hudson R. (2016), The effect of size on the failure probabilities of SMEs: an empirical study on the US market using discrete hazard model, *International Review of Financial Analysis,* 43, 135–145.

Fabling R., Grimes A. (2005), Insolvency and economic development: regional variation and adjustment, *Journal of Economics and Business*, 57(4), 339–359.

Gajdka J., Stos D. (1996), Wykorzystanie analizy dyskryminacyjnej w ocenie kondycji finansowej przedsiębiorstw, in: R. Borowiecki (ed.), *Restrukturyzacja w procesie przekształceń i rozwoju przedsiębiorstw*, Wydawnictwo AE w Krakowie.

Jiménez G., Saurina J. (2004), Collateral, type of lender and relationship banking as determinants of credit risk, *Journal of Banking & Finance*, 28(9), 2191–2212.

Keasey K., Watson R. (1987), Non-financial symptoms and the prediction of small company failure: a test of Argenti's hypothesis, *Journal of Business Finance and Accounting*, 14(3), 335–354.

Korol T., Prusak B. (2009), *Upadłość przedsiębiorstwa a wykorzystanie sztucznej inteligencji*, CeDeWu.

Mączyńska E. (1994), Ocena kondycji przedsiębiorstwa (uproszczone metody), *Życie Gospodarcze*, 38, 42–45.

Pogodzińska M., Sojak S. (1995), Wykorzystanie analizy dyskryminacyjnej w przewidywaniu bankructwa przedsiębiorstw w AUNC, *Ekonomia* XXV, 299.

Prusak B. (2005), *Nowoczesne metody prognozowania zagrożenia finansowego przedsiębiorstw*, Difin.

Psillaki M., Tsolas I.E., Margaritis D. (2010), Evaluation of credit risk based on firm performance, *European Journal of Operational Research*, 201(3), 873–881.

Sartori F., Mazzucchelli A., Di Gregorio A. (2016), Bankruptcy forecasting using case-based reasoning: the CRePERIE approach, *Expert Systems with Applications*, 64, 400–411.

Sohn S.Y., Kim D. H., Yoon J.H. (2016), Technology credit scoring model with fuzzy logistic regression, *Applied Soft Computing*, 43, 150–158.

Wierzba D. (2000), *Wczesne wykrywanie przedsiębiorstw zagrożonych upadłością na podstawie wskaźników finansowych – teoria i badania empiryczne*, Zeszyty Naukowe, 9, Wydawnictwo Wyższej Szkoły Ekonomiczno-Informacyjnej w Warszawie.

Wilson R.L., Sharda R. (1994), Bankruptcy prediction using neural networks, *Decision Support Systems*, 11, 545–557.

Yip A.Y.N. (2006), Business failure prediction: a case-based reasoning approach, *Review of Pacific Basin Financial Markets and Policies*, 09, 491–508.

Yoon J.S., Kwon Y.S. (2010), A practical approach to bankruptcy prediction for small businesses: substituting the unavailable financial data for credit card sales information, *Expert Systems with Applications*, 37, 3624–3629.

Zmijewski M.E. (1984), Methodological issues related to the estimation of financial distress prediction models, *Journal of Accounting Research*, 22, 59–82.

Appendix

Table 1
Financial ratios used in the analysis

Ratio	Name	Formula
X1	current liquidity	$\frac{current\ assets}{short\text{-}term\ liabilities}$
X2	quick ratio	$\frac{current\ assets - inventory - prepayments}{short\text{-}term\ liabilities}$
X3	liquidity cash	$\frac{cash}{short\text{-}term\ liabilities}$
X4	capital share in assets	$\frac{current\ assets - short\text{-}term\ liabilities}{total\ assets}$
X5	gross margin	$\frac{gross\ profit/loss\ on\ sale}{operating\ expenses}$
X6	operating profitability of sales	$\frac{profit/loss\ on\ operating\ activities}{total\ revenues}$
X7	operating profitability of assets	$\frac{profit/loss\ on\ operating\ activities}{total\ assets}$
X8	net profitability of equity	$\frac{net\ profit/loss}{equity}$
X9	assets turnover	$\frac{total\ revenues}{total\ assets}$
X10	current assets turnover	$\frac{total\ revenues}{current\ assets}$
X11	receivables turnover	$\frac{total\ revenues}{receivables}$
X12	inventory turnover	$\frac{total\ revenues}{inventory}$
X13	capital ratio	$\frac{equity}{total\ liabilities}$
X14	coverage of short-term liabilities by equity	$\frac{equity}{short\text{-}term\ liabilities}$
X15	coverage of fixed assets by equity	$\frac{equity}{fixed\ asstes}$
X16	share of net financial surplus in total liabilities	$\frac{net\ profit/loss + amortisation + interest}{total\ liabilities}$

Table 2
Non-financial factors used in the analysis

Name	Attributes/categories
Sector of the company's activity	Production, trade, services
Region of the company's activity (voivodeship)	16 regions so called voivodeships
Company's legal form	Self-employed, joint stock company, limited liability company, limited partnership company, other (e.g. cooperative, association, etc.)
Year of the company's establishment (age of the company)	Interval variable (age in years)
Number of employees	Discrete variable (number of employed workers on the date of financial statement)

Table 3
Clusters of regions and their characteristics

Cluster	Number of enterprises	Bankruptcy rate (in %)
Małopolskie, Lubelskie	107	32.7
Zachodniopomorskie, Pomorskie, Podkarpackie, Mazowieckie, Warmińsko--Mazurskie, Lubuskie	432	46.5
Świętokrzyskie, Wielkopolskie, Podlaskie, Śląskie, Opolskie, Kujawsko-Pomorskie	305	41.6
Łódzkie, Dolnośląskie	125	58.4

Table 4
Model comparison – accuracy ratio (Gini)

Financial and non-financial ratios			**Financial ratios only**		
model	**test sample**	**training sample**	**model**	**test sample**	**training sample**
Gradient boosting	0.614	0.759	Gradient boosting	0.574	0.727
Logistic regression	0.556	0.595	Logistic regression	0.522	0.546
Decision tree	0.501	0.731	Decision tree	0.517	0.682
Neural network	0.486	0.754	Neural network	0.490	0.721

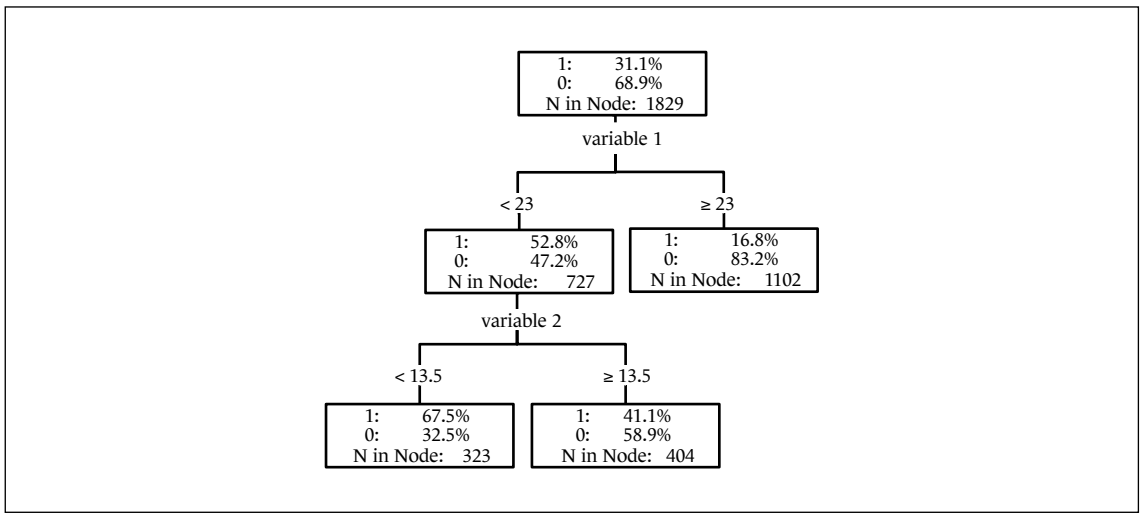
Table 5
Classification table – train sample. Gradient boosting random forests

	Model = 1	**Model = 0**	**Total**
Actual = 1	228	76	304
Actual = 0	55	319	374
Total	283	395	678

Table 6
Importance of the variables in gradient boosting

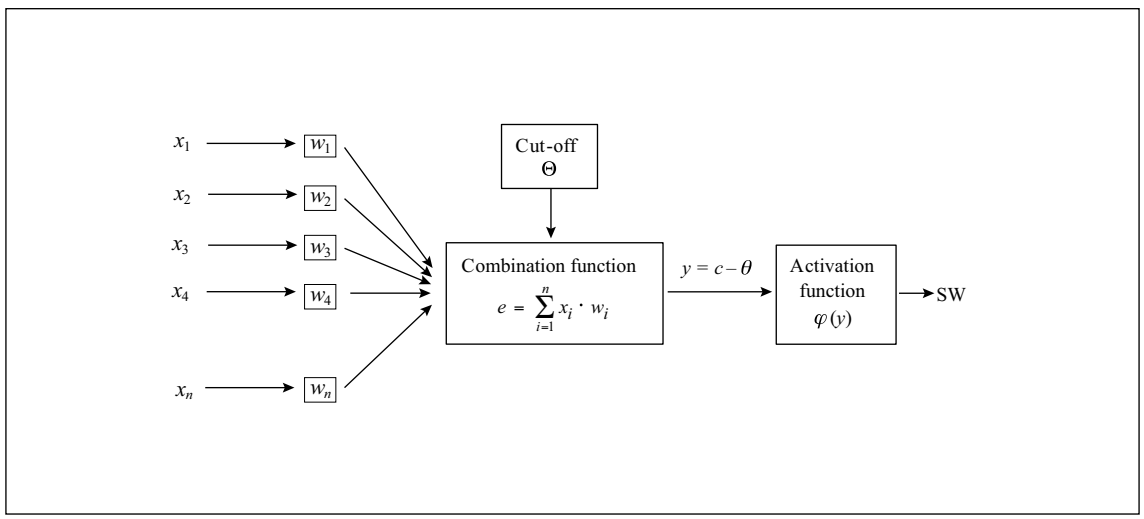
Variable	Number of rules	Importance
X16	11	1.00000
X13	8	0.61897
X12	11	0.42262
X3	11	0.40412
X15	6	0.34499
Legal form	6	0.32095
X14	5	0.31401
X11	6	0.30986
Employment	8	0.29354
Sector	6	0.28866
X9	5	0.24959
X1	5	0.23360
X2	4	0.21964
X6	4	0.21590
X5	3	0.20981
X10	3	0.20178
X8	4	0.19464
Region	3	0.17857
Age of the company	2	0.13371
X7	1	0.11488

Graph 1
Decision tree model



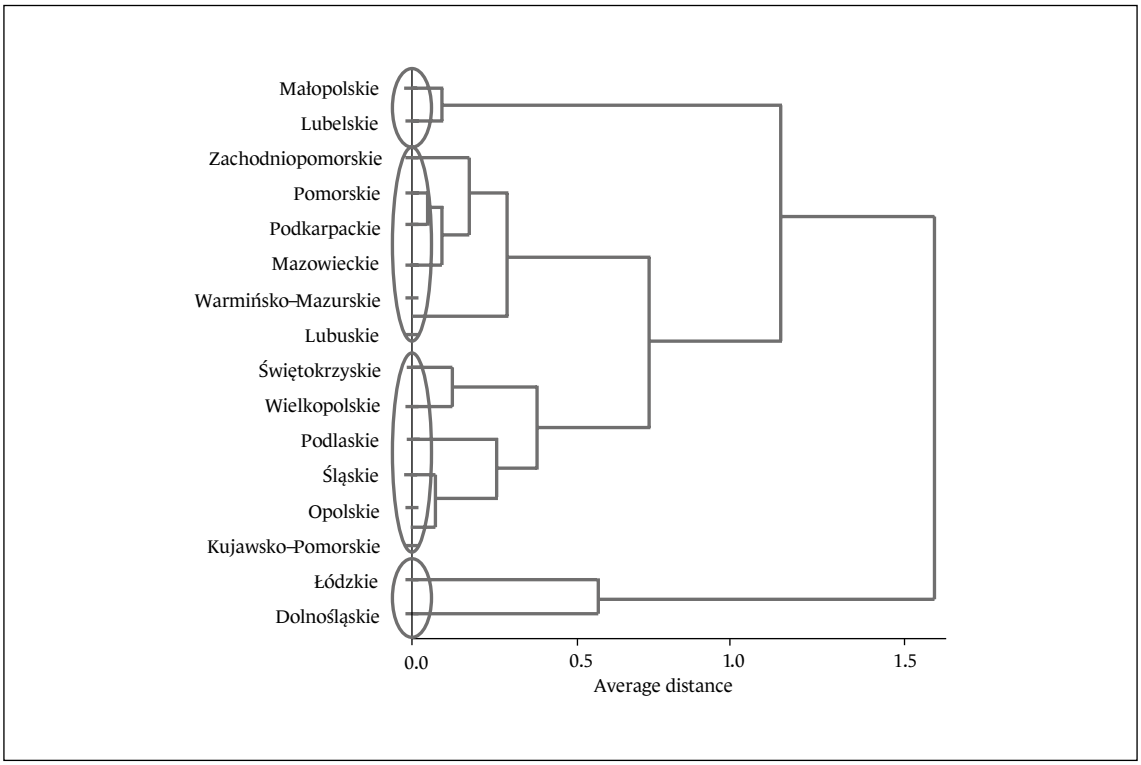
Note: variable 1: for example *age* of the customer; variable 2: for example *duration* of the credit (in months).

Graph 2
Neural network model – neuron scheme

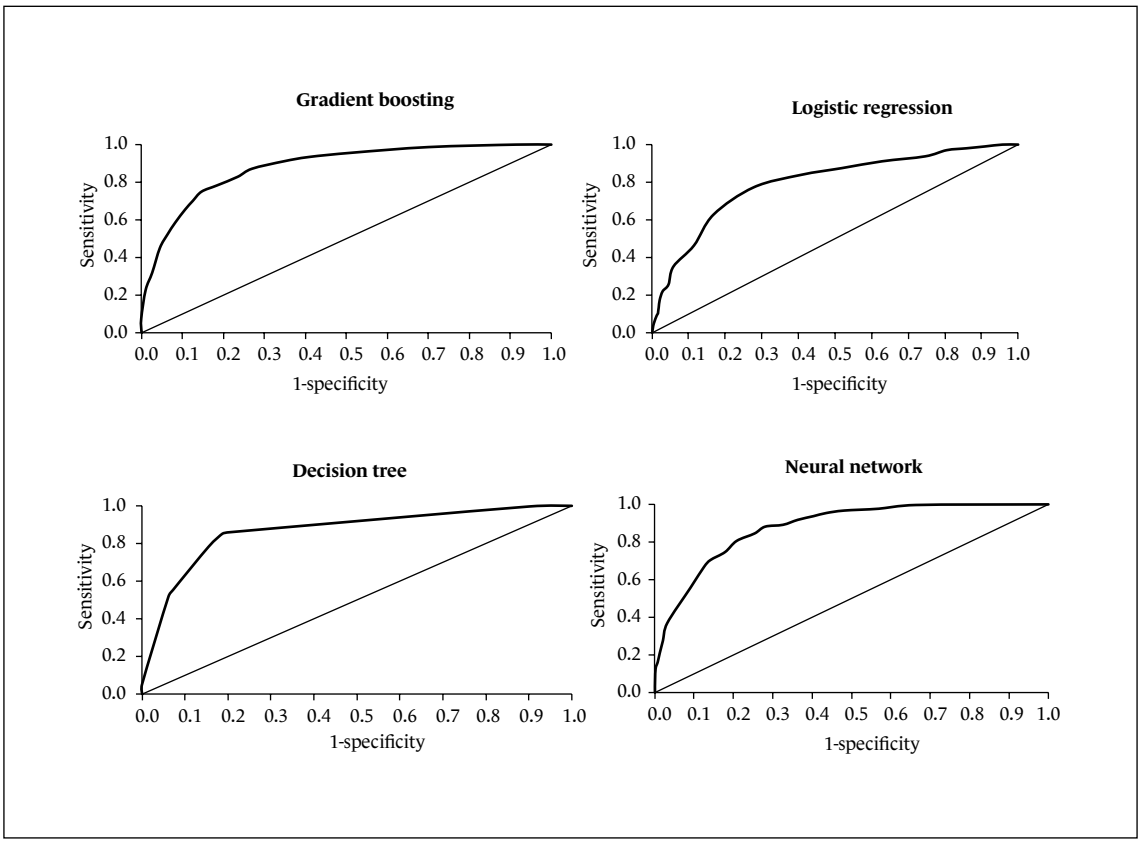


Source: analysis based on Korol, Prusak (2009, p. 57).

Graph 3
Clustering of the *region* variable – dendrogram



Graph 4
The ROC curve for estimated models (training sample)



Graph 5
Decision tree – graphical representation of results