Expert Systems With Applications 58 (2016) 93–101

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction

Maciej Zięba[a,*], Sebastian K. Tomczak[b], Jakub M. Tomczak[a]

[a] *Department of Computer Science, Faculty of Computer Science and Management, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370, Wrocław, Poland*

[b] *Department of Operations Research, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370, Wrocław, Poland*

ARTICLE INFO

Article history:
Received 26 January 2016
Revised 1 April 2016
Accepted 2 April 2016
Available online 6 April 2016

Keywords:
Bankruptcy prediction
Extreme gradient boosting
Synthetic features generation
Imbalanced data

ABSTRACT

Bankruptcy prediction has been a subject of interests for almost a century and it still ranks high among hottest topics in economics. The aim of predicting financial distress is to develop a predictive model that combines various econometric measures and allows to foresee a financial condition of a firm. In this domain various methods were proposed that were based on statistical hypothesis testing, statistical modeling (e.g., generalized linear models), and recently artificial intelligence (e.g., neural networks, Support Vector Machines, decision tress). In this paper, we propose a novel approach for bankruptcy prediction that utilizes Extreme Gradient Boosting for learning an ensemble of decision trees. Additionally, in order to reflect higher-order statistics in data and impose a prior knowledge about data representation, we introduce a new concept that we refer as to synthetic features. A synthetic feature is a combination of the econometric measures using arithmetic operations (addition, subtraction, multiplication, division). Each synthetic feature can be seen as a single regression model that is developed in an evolutionary manner. We evaluate our solution using the collected data about Polish companies in five tasks corresponding to the bankruptcy prediction in the 1st, 2nd, 3rd, 4th, and 5th year. We compare our approach with the reference methods.

1. Introduction

Prediction of an enterprise bankruptcy is of great importance in economic decision making. A business condition of either small or large firm concerns local community, industry participants and investors, but also influences policy makers and global economy. Therefore, the high social and economic costs as a consequence of corporate bankruptcies have attracted attention of researchers for better understanding of bankruptcy causes and eventually prediction of business distress (Zhang, Wang, & Ji, 2013).

The purpose of the bankruptcy prediction is to assess the financial condition of a company and its future perspectives within the context of long-term operation on the market (Constand & Yazdipour, 2011). It is a vast area of finance and econometrics that combines expert knowledge about the phenomenon and historical data of prosperous and unsuccessful companies. Typically, enterprises are quantified by a numerous indicators that describe their business condition that are further used to induce a mathematical model using past observations (Altman & Hotchkiss, 2010).

There are different issues that are associated with the bankruptcy prediction. Two main problems are the following: First, the econometric indicators describing the firm's condition are proposed by domain experts. However, it is rather unclear how to combine them into a successful model. Second, the historical observations used to train a model are usually influenced by imbalanced data phenomenon, because there are typically much more successful companies than the bankrupted ones. As a consequent, the trained model tends to predict companies as successful (majority class) even when some of them are distressed firms. Both of these issues mostly influence the final predictive capability of the model.

Previous works. First attempts of the formal bankruptcy prediction trace back to the beginnings of the 20th century when first econometric indicators were proposed to describe predictive abilities of business failure (Fitzpatrick, 1932; Merwin, 1942; Winakor & Smith, 1935). The sixties of the twentieth century brought a turning point in the survey of the early recognition of the business failure symptoms. First of all, the work of Beaver (1966) initiated

* Corresponding author. Tel.: +48713204150.
E-mail addresses: maciej.zieba@pwr.edu.pl (M. Zięba), sebastian.tomczak@pwr.edu.pl (S.K. Tomczak), jakub.tomczak@pwr.edu.pl (J.M. Tomczak).

http://dx.doi.org/10.1016/j.eswa.2016.04.001
0957-4174/

application of statistical models to the bankruptcy prediction. Following this line of thinking, Altman (1968) proposed to use multidimensional analysis to predict corporate bankruptcy that was further developed by others (Altman & Loris, 1976; Blum, 1974; Deakin, 1972; Edmister, 1972; Ketz, 1978; Koh & Killough, 1990; Laitinen, 1991; Libby, 1975; Meyer & Pifer, 1970; Pettway & Sinkey, 1980; Rujoub, Cook, & Hay, 1995; Sinkey, 1975; Wilcox, 1973). In parallel, a great interest was paid to the generalized linear models that can be used in both decision making and providing certainty of the prediction (Aziz, Emanuel, & Lawson, 1988; Grice & Dugan, 2003; Hopwood, McKeown, & Mutchler, 1994; Koh, 1991; Li & Miu, 2010; Ohlson, 1980; Platt & Platt, 1990; Platt, Platt, & Pedersen, 1994; Zavgren, 1983; Zmijewski, 1984). Additionally, the generalized linear models are of special interest because estimated weights of the linear combination of economic indicators in the model can be further used to determine importance of the economic indicators.

Since nineties of the 20th century artificial intelligence and machine learning have become a major research direction in the bankruptcy prediction. In the era of increasing volumes of data it turned out that the linear models like the logistic regression or logit (probit) models are unable to reflect non-trivial relationships among economic metrics. Moreover, the estimated weights of the linear models are rather unreliable to indicate the importance of the metrics.

In order to obtain comprehensible models with an easy to understand knowledge representation, decision rules expressed in terms of first-order logic were induced using different techniques, naming only a few, like rough sets (Dimitras, Slowinski, Susmaga, & Zopounidis, 1999) or evolutionary programming (Zhang et al., 2013). However, the classification accuracy of the decision rules are very often insufficient, therefore, more accurate methods were applied to the bankruptcy prediction. One of the most successful model was support vector machines (SVM) (Shin, Lee, & Kim, 2005). The disadvantages of SVM are that the kernel function must be carefully hand-tuned and it is impossible to obtain comprehensible model.

A different approach aims at automatic feature extraction from data, i.e., automatic non-linear combination of econometric indicators, which alleviates the problem of a specific kernel function determination in the case of SVM. This approach applies neural networks to the bankruptcy prediction (Bell, Ribar, & Verchio, 1990; Cadden, 1991; Coats & Fant, 1991; Geng, Bose, & Chen, 2015; Koster, Sondak, & Bourbia, 1991; Salchenberger, Cinar, & Lash, 1992; Serrano-Cinca, 1996; Tam, 1991; Tam & Kiang, 1992; Wilson & Sharda, 1994; Zhang, Hu, Patuwo, & Indro, 1999). The main problem of the neural networks lies in the fact that they can fail in case of multimodal data. Typically the econometric metrics need to be normalized/standardized in order to have all features of the same magnitude. This is also necessary for training neural networks so that the errors could be backpropagated properly. However, the normalization/standardization of data do not reduce the problem of the data multimodality that may drastically reduce predictive capabilities of the neural networks. That is why it has been advocated to take advantage of different learning paradigm, namely, the ensemble of classifiers (Kittler, Hatef, Duin, & Matas, 1998). The idea of the ensemble learning is to train and combine typically weak classifiers to obtain better predictive performance. First approaches but still very successful were bagging (Breiman, 1996) and boosting (Freund & Schapire, 1996; Friedman, 2001; 2002; Zięba, Tomczak, Lubicz, & Świątek, 2014). The idea of boosting was further developed to the case of unequal classification costs (Fan, Stolfo, Zhang, & Chan, 1999) and imbalanced data (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012). Recently, the boosting method was modified to optimize a Taylor expansion of the loss functions, an approach known as *Extreme Gradient Boosting* (Chen & He, 2015a) that obtains state-of-the-art results in many problems on Kaggle competitions.[1] Recently, it has been shown that the ensemble classifier can be successfully applied to the bankruptcy prediction (Nanni & Lumini, 2009) and it significantly beats other methods (Alfaro, García, Gámez, & Elizondo, 2008).

Contribution. In this paper we propose a novel method for bankruptcy prediction that makes use of Extreme Gradient Boosting (Chen & He, 2015b) for developing regularized boosted trees (Chen & He, 2015a; Johnson & Zhang, 2011). Best to our knowledge, such an approach was not applied to solve the problem of predicting financial condition of the companies. However, this method is successfully applied to many classification problems (Chen & He, 2015a) and widely used in winning Kaggle competitions. The model is also insensitive to imbalanced data phenomenon because it enables to select AUC measure for evaluation and forces proper ordering of the imbalanced data. To improve the prediction of the model we use ensemble of boosted trees, where each base learner is constructed using additional *synthetic features*. The synthetic features are developed at each boosting step in an evolutionary fashion by combining features using an arithmetic operation. Each synthetic feature can be seen as a single regression model. The purpose of the synthetic features is to combine the econometric indicators proposed by the domain experts into a complex features. The synthetic features can be seen as hidden features extracted by the neural networks but the fashion they are extracted is different. At the end, we test our solution using collected data about Polish companies.

Organization of the paper. The paper is organized as follows. In Section 2 the ensemble boosted trees is introduced as the model for bankruptcy prediction. In Section 3 we present the experimental results gained on real dataset representing the financial condition of the polish companies. The paper is summarized by the conclusions in Section 4.

2. Methodology

2.1. Extreme Gradient Boosting Framework

Let us denote by $\mathbf{x} \in \mathcal{X}$ a vector of features describing an enterprise, where $\mathcal{X} \subseteq \mathbb{R}^D$ and by $y \in \{0, 1\}$ a label representing whether the enterprise is bankrupt, $y = 1$, or not, $y = 0$. Further, we utilize decision trees as discriminative models, more precisely, Classification and Regression Trees (CART). A CART tree can be represented by the weights associated with the leaves in the tree structure

$$f_k(\mathbf{x}_n) = w_{q(\mathbf{x})}, \tag{1}$$

where $q(\mathbf{x}_n)$ is the function that takes an example $\mathbf{x}$ and returns the path id in the structure of the tree, $q : \mathbb{R}^D \to \{1, \dots, T\}$, T is the number of paths (leaves). A path is ended with a leaf that contains weight w_i.

We aim at learning an ensemble of K decision trees(Chen & He, 2015a)

$$h_K(\mathbf{x}) = \sum_{k=1}^{K} f_k(\mathbf{x}), \tag{2}$$

where $f_k \in \mathcal{F}$, for $k = 1, \dots, K$, and $\mathcal{F}$ is a space of all possible decision trees (CART). In order to obtain a decision for new $\mathbf{x}$ one could calculate a conditional probability of a class for h_K as follows:

$$p(y = 1|\mathbf{x}) = \sigma\left(h_K(\mathbf{x})\right), \tag{3}$$

where $\sigma(a) = \frac{1}{1+\exp(-a)}$ is the sigmoid function.

[1] www.kaggle.com/

For given training data $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$, the model is trained by minimizing the following criterion:

$$L_{\Omega}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \Omega(\boldsymbol{\theta}) = \sum_{n=1}^{N} l(y_n, h_K(\mathbf{x}_n)) + \sum_{k=1}^{K} \Omega(f_k), \tag{4}$$

where $\boldsymbol{\theta}$ represents the parameters of the model, i.e., $\boldsymbol{\theta} = \{f_1, \ldots, f_K\}$, $\Omega(\boldsymbol{\theta}) = \sum_{k=1}^{K} \Omega(f_k)$ is a regularization term and $L(\boldsymbol{\theta}) = \sum_{n=1}^{N} l(y_n, h_K(\mathbf{x}_n))$ is a loss function. In this work we consider the binary classification task, for which we use the logistic loss

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} [y_n \log(1 + \exp\{-h_K(\mathbf{x}_n)\}) + (1 - y_n) \log(1 + \exp\{h_K(\mathbf{x}_n)\})]. \tag{5}$$

The ensemble model for this loss function is known as LogitBoost model (Chen & He, 2015a).

The problem of learning such model can be solved iteratively by adding a new weak learner $f_k(\cdot)$ in the k-th training iteration assuming that models $f_1(\cdot), \ldots, f_{k-1}(\cdot)$ are already trained. We can present the loss function for single example $l(y_n, h_k(\mathbf{x}_n))$ in the following manner:

$$l(y_n, h_k(\mathbf{x}_n)) = l(y_n, h_{k-1}(\mathbf{x}_n) + f_k(\mathbf{x}_n)) \tag{6}$$

We assumed additive regularization term, therefore we can represent it in the following form:

$$\sum_{i=1}^{k} \Omega(f_i) = \Omega(f_k) + \Omega(h_{k-1}) = \Omega(f_k) + constant \tag{7}$$

As a consequence, we can represent the general learning criterion (4) as

$$L_{\Omega}(\boldsymbol{\theta}) = \sum_{n=1}^{N} l(y_n, h_{k-1}(\mathbf{x}_n) + f_k(\mathbf{x}_n)) + \Omega(f_k) + constant \tag{8}$$

Further, approximating the objective function using the Taylor expansion with respect to $h_{k-1}(\mathbf{x}_n)$ yields

$$L_{\Omega}(\boldsymbol{\theta}) \simeq \sum_{n=1}^{N} [l(y_n, h_{k-1}(\mathbf{x}_n)) + g_n \cdot f_k(\mathbf{x}_n) + \frac{1}{2} \cdot h_n \cdot f_k^2(\mathbf{x}_n)] + \Omega(f_k) + constant, \tag{9}$$

where g_n is the first derivative with respect to $h_{k-1}(\mathbf{x}_n)$

$$g_n = \frac{\partial l(y_n, h_{k-1}(\mathbf{x}_n))}{\partial h_{k-1}(\mathbf{x}_n)}, \tag{10}$$

and h_n is the second derivative with respect to $h_{k-1}(\mathbf{x}_n)$

$$h_n = \frac{\partial^2 l(y_n, h_{k-1}(\mathbf{x}_n))}{\partial h_{k-1}^2(\mathbf{x}_n)}. \tag{11}$$

Considering the logistic loss (4) we have

$$\begin{aligned} g_n &= -y_n \frac{\exp\{-h_{k-1}(\mathbf{x}_n)\}}{1 + \exp\{-h_{k-1}(\mathbf{x}_n)\}} + (1 - y_n) \frac{\exp\{h_{k-1}(\mathbf{x}_n)\}}{1 + \exp\{h_{k-1}(\mathbf{x}_n)\}} \\ &= -y_n \frac{1}{1 + \exp\{h_{k-1}(\mathbf{x}_n)\}} + (1 - y_n) \frac{1}{1 + \exp\{-h_{k-1}(\mathbf{x}_n)\}} \\ &= -y_n(1 - \sigma(h_{k-1}(\mathbf{x}_n))) + (1 - y_n)\sigma(h_{k-1}(\mathbf{x}_n)) \\ &= \sigma(h_{k-1}(\mathbf{x}_n)) - y_n, \end{aligned} \tag{12}$$

In calculating the first derivative we took advantage of the sigmoid function property, namely, $\sigma(-a) = 1 - \sigma(a)$. It can be observed, that $\sigma(h_{k-1}(\mathbf{x}_n))$ has interpretation of the probability of observing the class indexed by 1 for the example $\mathbf{x}_n$.

We can make use of $\sigma'(a) = \sigma(a)(1 - \sigma(a))$ property to calculate the second derivative, h_n

$$h_n = \sigma(h_{k-1}(\mathbf{x}_n))(1 - \sigma(h_{k-1}(\mathbf{x}_n))) \tag{13}$$

There are different possible regularization terms. However, in our considerations we focus on the regularizer in the following form:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{t=1}^{T} w_t^2, \tag{14}$$

where λ and γ are the parameters of the regularization term. For the tree representation with weights the objective function given in (9) can be presented in the following manner:

$$\begin{aligned} L_{\Omega}(\boldsymbol{\theta}) &\simeq \sum_{n=1}^{N} [g_n\ w_{q(\mathbf{x}_n)} + \frac{1}{2}\ h_n \cdot w_{q(\mathbf{x}_n)}^2] + \gamma T \\ &\quad + \frac{1}{2}\lambda \sum_{t=1}^{T} w_t^2 + constant \\ &= \sum_{t=1}^{T} [(\sum_{j \in I_t} g_j) w_t + \frac{1}{2}(\sum_{j \in I_t} h_j + \lambda) w_t^2] + \gamma T + constant \\ &= \sum_{t=1}^{T} [G_t w_t + \frac{1}{2}(H_t + \lambda) w_t^2] + \gamma T + constant, \end{aligned} \tag{15}$$

where $I_t = \{n | q(\mathbf{x}_n) = t\}$ is the set of indexes of instances associated with the t-th leaf in the tree, $G_t = \sum_{j \in I_t} g_j$ and $H_t = \sum_{j \in I_t} h_j$. Assuming the known structure of the tree, the optimal value of the weight in the t-th leaf is as follows:

$$w_t^* = -\frac{G_t}{H_t + \lambda} \tag{16}$$

The optimal value of the approximated objective function is given by

$$L_{\Omega}(\boldsymbol{\theta}) \simeq -\frac{1}{2} \sum_{t=1}^{T} \frac{G_t^2}{H_t + \lambda} + \gamma T + constant \tag{17}$$

The key problem in the above consideration is that the structure of the tree is not given in advanced and searching all possible structures is computationally infeasible. To overcome this issue the tree is being constructed starting from the root and further the best attribute to be located in the root is selected and the best split point for the attribute is chosen. The splitting process is performed until the quality of the model is improved. As the splitting criterion we take the info gain:

$$\mathcal{G} = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_R + H_L + \lambda} - \gamma, \tag{18}$$

where $\frac{G_L^2}{H_L+\lambda}$ is the score value calculated for the left child, $\frac{G_R^2}{H_R+\lambda}$ for is the score value for the right ancestor and $\frac{(G_L+G_R)^2}{H_R+H_L+\lambda}$ is the score value if splitting is not performed. Parameter γ penalizes addition of more leaves to the tree structure.

The model can be also regularized by setting minimal number of examples combined with each of the leaves, by setting maximal depth of the tree, by setting the percentage of features randomized for each iteration of constructing the tree or by adding the new tree with corrected influence of the trees in the committee

$$h_k(\mathbf{x}_n) = h_{k-1}(\mathbf{x}_n) + \epsilon f_k(\mathbf{x}_n), \tag{19}$$

where $\epsilon \in [0, 1]$ is called step-size or shrinkage.

2.2. Ensemble of boosted trees for bankruptcy prediction

Motivation. The motivation of applying boosted trees trained with the Extreme Gradient Boosting method to the bankruptcy prediction is that estimators of economic indicators describing the companies are characterized by high variance caused by relatively small number of samples. Practically it means that most of the values of some indicators are accumulated in some narrow segment but there are some companies that are described by relatively high/small values of those features. As a consequence, the application of gradient-based models like neural networks or logistic regression leads to the training issues and eventually poor prediction. The problem is also difficult to overcome when data is normalized or standardized. Contrary to this approaches ensemble tree-based learners take into account the order of feature values, not the values itself. Therefore, they are resistant to huge values of the economic indicators and do not need any pre-processing stage.

Synthetic features. Ensemble tree-based models can also effectively learn from data described by many features. We take advantage of this property by proposing the ensemble of boosted trees model dedicated to solve the problem of bankruptcy prediction. The central idea in our approach is to generate synthetic features that may have better influence on prediction than typical economic factors. The synthetic features are generated by random selection of two existing features and random selection of arithmetical operation to be performed on them. To estimate the probability of selecting the seed features we make use of popularity of the feature in the already created forest. The popularity of the feature in the forest is described by the total number of occurrences in trees that constitutes the forest. Let us denote the total number of occurrences of the d-th feature in the forest structure by m_d. We define the categorical distribution $\boldsymbol{\theta}_F = [\theta_F^{(1)}, \dots, \theta_F^{(d)}, \dots \theta_F^{(D)}]$ for selecting the features to be replicated in the following manner:

$$\theta_F^{(d)} = \frac{m_d}{\sum_{d=1}^{D} m_d}. \tag{20}$$

As a consequence, the most popular features are going to be selected for reproduction. The proposed procedure can be seen as a kind of an evolutionary approach that selects the strongest parents for the child feature.

The arithmetic operation is selected from uniform distribution defined on the set of possible values, $\{+, -, *, /\}$.

We motivate introduction of the synthetic features twofold. First, the synthetic features can be seen as regression models that represent complex relationships among features. Notice that such regression models cannot be trained using CART. Moreover, the synthetic features can be seen as a kind of hidden units in the neural networks but the manner they are extracted is completely different. Second, in case of small sample size, the synthetic features are easier to learn that training a complicated tree structure.

Learning algorithm. The procedure of constructing ensemble of base learners is described by Algorithm 1 . In each of the training iterations one of the base learners h_k that represents boosted trees is trained with Extreme Gradient Boosting approach using dataset $\mathcal{D}$. Basing on feature importance m_d, $d = 1, \dots, D$ gathered from trained model h_k we select only those features, for which m_d is above given threshold value η. The trained model is further used to determine the popularity of the features and estimate the distribution $\boldsymbol{\theta}_F$.

Further, the synthetic features are generated using the following framework. Two features f_1 and f_2 are sampled from distribution $\boldsymbol{\theta}_F$. Next, the operation $\circ$ is uniformly sampled from the set $\{+, -, *, /\}$. The value of new feature $f_{new} = f_1 \circ f_2$ is calculated for

Input : $\mathcal{D}$: training set, D_{new}: number of synthetic features, L: number of base learners, η: features acceptance threshold
Output: $H = \{h_1, \dots, h_K\}$: set of base learners
for $k = 1, \dots, K$ **do**
  Train h_k using $\mathcal{D}$;
  Remove features from $\mathcal{D}$ for which $m_d < \eta$;
  Estimate $\boldsymbol{\theta}_F$ from model h_k;
  for $d = 1, \dots, D_{new}$ **do**
    Sample features f_1 and f_2 from distribution $\boldsymbol{\theta}_F$;
    Sample operation $\circ$ from $\{+, -, *, /\}$;
    Generate new feature $f_{new} = f_1 \circ f_2$;
    Extend $\mathcal{D}$ with new values of f_{new};
  end
end
return $H = \{h_1, \dots, h_K\}$;

Algorithm 1: Ensemble of boosted trees with synthetic features

all examples in dataset $\mathcal{D}$. The process of creating synthetic features is repeated until the desired number synthetic features, D_{new}, is reached. The extended dataset is further used to construct the h_{k+1} base model.

3. Experiments

3.1. Dataset

Data preparation. To evaluate the quality of the approach we collected the data about financial condition of Polish companies. The process of selection data consists of choosing the sector, the database, the research period, the number of companies and the number of financial indicators that will be analyzed. First of all, in Poland, since 2004, many companies in the manufacturing sector went bankrupt, therefore we decided to analyze this sector. Then, we chose the database Emerging Markets Information Service (EMIS).[2] which is a database containing information on emerging markets around the world including the Polish one. The service provides access to 540 publications containing financial information, political, macroeconomic and companies news in local languages and English. Moreover, the source of base includes articles, news agency messages, the financial statements of companies, industry reports, stock quotes and statistics and analyzes macroeconomic data.

Next, the period of time was established which is 2007–2013 for bankrupt and 2000–2012 for still operating companies and it is due to the availability of data in the database EMIS. The research sample consists of bankrupt and still operating companies (imbalanced sample). In the period of 2007–2013 nearly 700 bankrupt enterprises (almost 2400 financial statements) were analyzed. In the period 2000–2012 more than 10,000 still operating ones, in this sample the company, which declared bankruptcy is excluded (more than 65 thousand financial statements) were taken into account. Finally, we determined the 64 financial indicators to be analyzed. This figure is due to the availability of data and the intensity of the occurrence in integrated models and financial analysis presented in related works (Tomczak, 2014a; 2014b; 2014c). The detailed methodology of collecting the data is described in Table 1.

The features considered in the research studies are described in details in Table 2. Basing on the collected data we distinguished five classification cases, that depends on the forecasting period

- *1stYear* – the data contains financial rates from 1st year of the forecasting period and corresponding class label that indicates

[2] http://www.securities.com

Table 1
The methodology of collecting the training data.

Name	Criterion	Selection
sector	the highest number of bankruptcies in the sector compared to other sectors	the manufacturing sector
database of financial statements	The availability of databases	EMIS
period	financial statements availability	5 years before bankruptcy in the period of 2007–2013, 2000–2012 for still operating companies
bankrupt companies	availability of at least one financial report in the analyzed period of five years before the bankruptcy of a company	nearly 700 from 1000 bankrupt enterprises were selected in the period of 2007–2013 (almost 2400 financial statements were analyzed)
still operating companies	the availability of a minimum of three consecutive financial statements in the period 2000–2012	more than 10,000 from 17,000 businesses still functioning were chosen (more than 65 thousand financial statements were taken into consideration)
financial indicators	used in the integrated models and financial analysis	64 financial ratios were analyzed (see Table 2).

Table 2
The set of features considered in classification process.

ID	Description	ID	Description
X1	net profit / total assets	X33	operating expenses / short-term liabilities
X2	total liabilities / total assets	X34	operating expenses / total liabilities
X3	working capital / total assets	X35	profit on sales / total assets
X4	current assets / short-term liabilities	X36	total sales / total assets
X5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] ∗ 365,	X37	(current assets - inventories) / long-term liabilities
X6	retained earnings / total assets	X38	constant capital / total assets
X7	EBIT / total assets	X39	profit on sales / sales
X8	book value of equity / total liabilities	X40	(current assets - inventory - receivables) / short-term liabilities
X9	sales / total assets	X41	total liabilities / ((profit on operating activities + depreciation) ∗ (12/365))
X10	equity / total assets	X42	profit on operating activities / sales
X11	(gross profit + extraordinary items + financial expenses) / total assets	X43	rotation receivables + inventory turnover in days
X12	gross profit / short-term liabilities	X44	(receivables ∗ 365) / sales
X13	(gross profit + depreciation) / sales	X45	net profit / inventory
X14	(gross profit + interest) / total assets	X46	(current assets - inventory) / short-term liabilities
X15	(total liabilities ∗ 365) / (gross profit + depreciation)	X47	(inventory ∗ 365) / cost of products sold
X16	(gross profit + depreciation) / total liabilities	X48	EBITDA (profit on operating activities - depreciation) / total assets
X17	total assets / total liabilities	X49	EBITDA (profit on operating activities - depreciation) / sales
X18	gross profit / total assets	X50	current assets / total liabilities
X19	gross profit / sales	X51	short-term liabilities / total assets
X20	(inventory ∗ 365) / sales	X52	(short-term liabilities ∗ 365) / cost of products sold)
X21	sales (n) / sales (n-1)	X53	equity / fixed assets
X22	profit on operating activities / total assets	X54	constant capital / fixed assets
X23	net profit / sales	X55	working capital
X24	gross profit (in 3 years) / total assets	X56	(sales - cost of products sold) / sales
X25	(equity - share capital) / total assets	X57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X26	(net profit + depreciation) / total liabilities	X58	total costs /total sales
X27	profit on operating activities / financial expenses	X59	long-term liabilities / equity
X28	working capital / fixed assets	X60	sales / inventory
X29	logarithm of total assets	X61	sales / receivables
X30	(total liabilities - cash) / sales	X62	(short-term liabilities ∗365) / sales
X31	(gross profit + interest) / sales	X63	sales / short-term liabilities
X32	(current liabilities ∗ 365) / cost of products sold	X64	sales / fixed assets

bankruptcy status after 5 years. The data contains 7027 instances (financial statements), 271 represents bankrupted companies, 6756 firms that did not bankrupt in the forecasting period.

- *2ndYear* – the data contains financial rates from 2nd year of the forecasting period and corresponding class label that indicates bankruptcy status after 4 years. The data contains 10173 instances (financial statements), 400 represents bankrupted companies, 9773 firms that did not bankrupt in the forecasting period.
- *3rdYear* – the data contains financial rates from 3rd year of the forecasting period and corresponding class label that indicates bankruptcy status after 3 years. The data contains 10503 instances (financial statements), 495 represents bankrupted companies, 10008 firms that did not bankrupt in the forecasting period.
- *4thYear* – the data contains financial rates from 4th year of the forecasting period and corresponding class label that indicates bankruptcy status after 2 years. The data contains 9792 instances (financial statements), 515 represents bankrupted companies, 9277 firms that did not bankrupt in the forecasting period.
- *5thYear* – the data contains financial rates from 5th year of the forecasting period and corresponding class label that indicates bankruptcy status after 1 year. The data contains 5910 instances (financial statements), 410 represents bankrupted companies, 5500 firms that did not bankrupt in the forecasting period.

3.2. Experiment setup

The goal of the experiment was to identify the best classification model for each of the bankruptcy prediction cases represented by the training data described in previous subsection. We took under consideration the following classification methods:

- **LDA**, linear discriminant analysis (Altman, 1968);
- **MLP**, multilayer perceptron with a hidden layer (Back, Laitinen, & Sere, 1996);
- **JRip**, decision rules inducer (Cohen, 1995);
- cost-sensitive variation of **JRip** (**CJRip**);
- **J48**, decision tree model (Quinlan, 1993);
- cost-sensitive variation of **J48** (**CJ48**);
- Logistic Regression (**LR**);
- cost-sensitive variation of Logistic Regression (**CLR**);
- AdaBoost (**AB**) (Freund & Schapire, 1996);
- AdaCost (**AC**) (Fan et al., 1999);
- Support Vector Machines (**SVM**) (Cortes & Vapnik, 1995);
- Cost-sensitive Support Vector Machines (**CSVM**);
- Random Forest (**RF**) (Ho, 1995);
- Boosted trees trained with Extreme Gradient Boosting (**XGB**);
- only the last tree of the ensemble of boosted trees, i.e., f_K, trained with the Algorithm 1 (**XGBE**);
- Ensemble of boosted trees trained with the Algorithm 1 (**EXGB**);

Most of the reference approaches were considered in the filed of bankruptcy prediction. Moreover, we had formulated the problem of predicting financial condition of the companies in terms of binary classification. Therefore, we had an opportunity to examine the quality of various machine learning approaches dedicated to solve two-class problems, even those, that are non-standard in the field of bankruptcy prediction.

Due to the imbalanced nature of training data, we utilized the Area Under ROC Curve (**AUC**) curve criterion to evaluate the quality of the models. For each of considered models we examined the quality of various settings of training parameters using 10 folds cross validation methodology. In Table 3 we present only the best results for each type of the considered classifiers.[3]

For the cost-sensitive models we set misclassification costs equal $\frac{N_-}{N_+}$ for minority examples and 1 for majority cases, where N_+ denotes number of minority examples and N_- stays behind the number of majority (Tomczak & Zięba, 2015).

For Ensemble of boosted trees we obtained the best results for the number of base learners equal 10 and number of synthetic features generated in each iteration equal 60. The feature is accepted for next iteration if was observed at least in 5% of trees in the forest. For testing boosted trees we used *xgboost*[4] library for *Python*. For testing other methods we used *Weka Data Mining Tool*[5] for *Java*.

3.3. Results

The experimental results are presented in Table 3. For each of the considered periods and examined models we present the mean (**MN**) and standard deviation (**STD**) for AUC measure that was calculated basing on 10 cross validation folds.

It can be noticed, that boosted trees significantly outperforms other models. For the reference classifiers the best results are gained by AdaBoost and AdaCost. For all datasets the last base learner of the ensemble f_K alone trained with the synthetic features (**XGBE**) gained slightly better results than the boosted tree model (**XGB**). The results were further improved if the base learners were formed in the ensemble structure using the synthetic features (**EXGB**).

To investigate the significance of difference between results gained by **XGB** vs. **XGBE, XGB** vs. **EXGB** and **XGBE** vs. **XGB** we applied signed rank Wilcoxson test. The *p*-values for considered pairs are as follows:

- for **XGB** vs. **XGBE** *p*-value is equal 0.003;
- for **XGB** vs. **EXGB** *p*-value is less than 0.001;
- for **XGBE** vs. **EXGB** *p*-value is equal 0.003;

Assuming the significance level equal 0.05, regarding the Wilcoxson test, we can reject all stated null median difference hypotheses. Concluding, **EXGB** performs better than **XGBE** and **XGB, XGBE** gained significantly better results then **XGB**.

3.4. Features importance evaluation

We evaluated the importance of the features by calculating the total number of the feature being observed in the nodes of forest structure by total number of nodes in trees that constitute the forest. In the other words, we take under consideration the categorical distribution $\theta_F^{(d)}$ defined in Eq. (20). In Table 4 we present 20 most important features for each of the considered classification cases. Analyzing the results presented in Table 4, it can be said that only three indicators $X25$ (adjusted share of equity in financing of assets), $X40$ (current ratio, the most frequently used ratio in the integrated models (Tomczak, 2014b)), $X52$ (liabilities turnover ratio) appeared in each research year. Therefore, they can be considered as useful in predicting bankruptcy of enterprises. It is worth noting that beside these three indicators, during the period considered the following indicators may also be useful: $X13$, $X22$, $X31$, $X42$ (profitability ratios), $X15$ (leverage ratios), $X9$, $X36$, $X48$, $X52$ (operating performance ratios), $X5$, $X27$, $X58$ (others). Because they occurred in 4 out of 5 years.

Further, we examine the popularity of the synthetic features generated to construct each of the base learners for *1stYear* dataset (see Table 5). For instance, we have a very popular feature that is observed in more than 3% of the nodes and can be calculated using formula ($X47/X27$). Days inventory ratio is divided by financial expenses coverage ratio which means that operating performance and profitability of a company do matter. For the third base learner over 4% popularity was gained by the feature that can be calculated with formula $((X22 * X25) + X27)$. Return on investment times adjusted share of equity in financing of assets plus financial expenses coverage ratio which means that profitability and leverage of a company also do matter. Therefore, the presented approach can be used to discover synthetic features that are so far undefined by experts as relevant financial factors. However, it should be checked which ones are characterized by a higher correctness of classification of companies.

For the 10th base learner the generated synthetic samples are presented in Table 6. Most of the features are too complicated to be interpreted in straightforward way. However, some of the features like $X46$ or $X29$ survived replication procedure with high popularity measure. The new features are characterized by decreasing average popularity. Therefore the model is prone to overfitting. That issue should be controlled by proper value of features acceptance threshold (η, see Algorithm 1).

[3] We selected the best model of each type according to the average value of AUC calculated from ten folds.

[4] https://xgboost.readthedocs.org/en/latest/

[5] http://www.cs.waikato.ac.nz/ml/weka/

Table 3
Experimental results for considered classification models.

	1stYear		2ndYear		3rdYear		4thYear		5thYear	
	MN	STD	MN	STD	MN	STD	MN	STD	MN	STD
LDA	.639	.083	.660	.037	.688	.030	.714	.063	.796	.041
MLP	.543	.042	.514	.042	.548	.041	.596	.049	.699	.059
JRip	.523	.030	.540	.025	.535	.022	.538	.026	.654	.049
CJRip	.745	.112	.774	.073	.804	.054	.799	.070	.778	.035
J48	.717	.059	.653	.068	.701	.062	.691	.076	.761	.049
CJ48	.658	.047	.652	.047	.618	.061	.611	.025	.719	.046
LR	.620	.065	.513	.042	.500	.000	.500	.000	.632	.119
CLR	.704	.065	.671	.032	.714	.034	.724	.041	.821	.037
AB	.916	.020	.850	.029	.861	.023	.885	.031	.925	.026
AC	.916	.023	.849	.022	.859	.022	.886	.015	.928	.023
SVM	.502	.006	.502	.006	.500	.000	.500	.000	.505	.006
CSVM	.578	.040	.517	.064	.614	.040	.615	.034	.716	.039
RF	.851	.044	.842	.028	.831	.031	.848	.027	.898	.035
XGB	.945	.033	.917	.027	.922	.025	.935	.024	.951	.024
XGBE	.953	.024	.941	.019	.929	.049	.940	.027	.954	.018
EXGB	**.959**	.018	**.944**	.021	**.940**	.032	**.941**	.025	**.955**	.019

Table 4
Ranking of features for each of the datasets.

	1stYear		2ndYear		3rdYear		4thYear		5thYear	
Rank	ID	$\theta_F^{(d)}$	ID	$\theta_F^{(d)}$	ID	$\theta_F^{(d)}$	ID	$\theta_F^{(d)}$	ID	$\theta_F^{(d)}$
1	X16	.0519	X40	.0473	X15	.0506	X22	.0461	X25	.0627
2	X52	.0380	X15	.0449	X22	.0382	X52	.0446	X22	.0480
3	X32	.0378	X27	.0404	X52	.0369	X15	.0413	X27	.0379
4	X28	.0355	X5	.0342	X27	.0337	X25	.0385	X15	.0356
5	X5	.0347	X25	.0341	X40	.0325	X27	.0345	X52	.0326
6	X40	.0333	X36	.0336	X5	.0309	X40	.0322	X53	.0284
7	X9	.0319	X22	.0277	X25	.0260	X58	.0257	X14	.0248
8	X11	.0308	X42	.0277	X31	.0257	X42	.0252	X40	.0247
9	X59	.0308	X31	.0268	X12	.0250	X13	.0250	X42	.0238
10	X23	.0266	X13	.0266	X42	.0234	X36	.0237	X36	.0236
11	X25	.0245	X12	.0220	X13	.0233	X31	.0234	X54	.0236
12	X55	.0245	X35	.0218	X53	.0230	X5	.0232	X12	.0216
13	X17	.0233	X9	.0216	X57	.0229	X53	.0228	X58	.0215
14	X14	.0221	X58	.0212	X37	.0216	X6	.0215	X41	.0210
15	X29	.0214	X11	.0209	X48	.0205	X35	.0209	X44	.0193
16	X13	.0210	X48	.0209	X6	.0202	X48	.0205	X48	.0193
17	X58	.0210	X52	.0208	X35	.0198	X9	.0201	X9	.0192
18	X30	.0192	X57	.0205	X41	.0188	X24	.0197	X31	.0192
19	X57	.0192	X55	.0184	X32	.0184	X38	.0197	X32	.0190
20	X56	.0174	X6	.0179	X36	.0182	X29	.0186	X16	.0189

Table 5
Ranking of features for first 3 base learners.

	1st Learner		2nd Learner		3rd Learner	
Ranking	ID	$\theta_F^{(d)}$	ID	$\theta_F^{(d)}$	ID	$\theta_F^{(d)}$
1	X16	.0519	X46	.0402	((X22∗X25)+X27)	.0437
2	X52	.0380	(X47/X27)	.0332	X29	.0242
3	X32	.0378	X29	.0311	X46	.0240
4	X28	.0355	X27	.0300	((X18-X34)/X56)	.0240
5	X5	.0347	(X18-X34)	.0281	X9	.0238
6	X40	.0333	X34	.0247	(X11∗X44)	.0224
7	X9	.0319	X9	.0228	X27	.0199
8	X11	.0308	(X13-X46)	.0221	((X32/X15)+X27)	.0192
9	X59	.0308	(X11∗X44)	.0217	(X24/X27)	.0181
10	X23	.0266	(X21+X62)	.0206	((X46/X61)∗(X61+X21))	.0174
11	X25	.0245	(X2-X45)	.0191	((X18-X34)/(X50∗X52))	.0167
12	X55	.0245	X11	.0183	(X18-X34)	.0162
13	X17	.0233	X37	.0174	X11	.0155
14	X14	.0221	X21	.0170	((X18-X34)+X57)	.0149
15	X29	.0214	(X61+X21)	.0168	((X61+X21)∗X32)	.0149
16	X13	.0210	X58	.0166	(X29-X58)	.0126
17	X58	.0194	(X17-X5)	.0160	X25	.0123
18	X30	.0192	X22	.0153	((X13-X46)-X6)	.0123
19	X57	.0192	X25	.0138	X58	.0112
20	X56	.0194	(X64-X37)	.0134	X34	.0107

Table 6
Ranking of features for considered by 10-th base learner.

Ranking	ID	$\theta_F^{(d)}$
1	(((((((X18-X34)/X56)/X46)/(X24/X27))∗(X11∗X44))/(((X18-X34)/(X36∗X58)) +((X38/X30)-X22)))/X46)	.0121
2	(((((X38/X30)-X22)+((X46/X61)∗(X61+X21)))∗(X30/(((X22∗X25)+X27)-((((X47/X27)-X46)∗(X33∗X39))∗(((X38/X30)-X22)+((X18-X34)/(X36∗X58)))))))+X46)	.0109
3	(((X18-X34)∗((X2-X45)∗X46))+X46)	.0106
4	(((((X18-X34)/X56)/X46)-X49)-((((X22∗X25)+X27)/(X29∗(X18-X34)))∗X22))	.0102
5	((((X11-((X2-X45)/X25))-((X18-X34)/(X36∗X58)))+(((X11∗((X11∗X44)-((X1/X61)-(X2-X45))))/(X46/((((X18-X34)/X56)/X46)/(X24/X27))))∗X50)) ∗(X11/((X46+X34)∗(X29-X58))))	.0094
6	((((X22∗X25)+X27)/(X29∗(X18-X34)))∗X22)	.0084
7	X46	.0077
8	((X29+X29)+(((X22∗X25)+X27)/(X29∗(X18-X34))))	.0074
9	(((X11∗((X11∗X44)-((X1/X61)-(X2-X45))))/(X46/((((X18-X34)/X56)/X46)/(X24/X27))))∗X50)	.0074
10	((((X38/X30)-X22)+((X46/X61)∗(X61+X21)))∗(X30/(((X22∗X25)+X27)-((((X47/X27)-X46)∗(X33∗X39))∗(((X38/X30)-X22)+((X18-X34)/(X36∗X58)))))))	.0074
11	(((((X56+X48)-X11)+X46)+((((X56+X48)-X11)+X46)-((X38/X30)-X22)))-X30)	.0074
12	(((((X22∗X25)+X27)/(X29∗(X18-X34)))∗X22)+(X48+(X25/X31)))	.0072
13	(((X48+(X25/X31))/((X35/(X26/X57))-X29))+(X48+(X25/X31)))	.0069
14	(X25-(((X22∗X25)+X27)/(X29∗(X18-X34))))	.0067
15	((((((X22∗X25)+X27)/(X29∗(X18-X34)))∗X22)+(X48+(X25/X31)))-X46)	.0067
16	(X46/((X46/((((X18-X34)/X56)/X46)/(X24/X27)))-((X13∗(((X61+X21)+X41)/((X2-X45)∗X46)))/((X25/X31)-(X47/X27)))))	.0067
17	X29	.0067
18	(((((X38/X30)-X22)+((X18-X34)/(X36∗X58)))+((((X47/X27)-X46)∗(X33∗X39))∗(((X38/X30)-X22)+((X18-X34)/(X36∗X58)))))-((((X38/X30)-X22)+((X46/X61)∗(X61+X21)))∗(X30/(((X22∗X25)+X27)-((((X47/X27)-X46)∗(X33∗X39))∗(((X38/X30)-X22)+((X18-X34)/(X36∗X58))))))))	.0064
19	((X46/((((X18-X34)/X56)/X46)/(X24/X27)))∗(X29+X29))	.0064
20	(X58∗(((X18-X34)/(X36∗X58))+((X38/X30)-X22)))	.0064

4. Conclusions

The paper presents the novel approach for the problem of predicting the bankruptcy basing on the financial factors. We took under consideration the financial condition of Polish companies from 2007 to 2013 (bankrupt) and from 2000 to 2012 for (still operating). To solve the stated classification problem we applied the Extreme Gradient Boosting model. The results gained by the selected classifier were significantly better than the results gained by all reference methods that were applied to the problem of predicting financial condition of the companies before. Further, we proposed some extension of the Extreme Gradient Boosting that randomly generates new synthetic features. The application of such approach led to significant improvement of the quality of prediction. We also thoroughly discussed the relevance of the newly created features. The presented model is not limited to Polish companies but represents a general framework that can be applied to an arbitrary given data from the considered domain.

Acknowledgments

The research conducted by the authors has been partially co-financed by the Ministry of Science and Higher Education, Republic of Poland, namely, Maciej Zięba: grant No. B50083/W8/K3.

References

Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of adaboost and neural networks. *Decision Support Systems, 45*, 110–122.
Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*, 589–609.
Altman, E. I., & Hotchkiss, E. (2010). *Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze and invest in distressed debt*: 289. Hoboken, New Jersey: John Wiley & Sons.
Altman, E. I., & Loris, B. (1976). A financial early warning system for over-the–counter broker-dealers. *The Journal of Finance, 31*, 1201–1217.
Aziz, A., Emanuel, D. C., & Lawson, G. H. (1988). Bankruptcy prediction-an investigation of cash flow based models. *Journal of Management Studies, 25*, 419–437.
Back, B., Laitinen, T., & Sere, K. (1996). Neural networks and genetic algorithms for bankruptcy predictions. *Expert Systems with Applications, 11*, 407–413.
Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, 71–111.
Bell, T. B., Ribar, G. S., & Verchio, J. (1990). Neural nets versus logistic regression: a comparison of each model's ability to predict commercial bank failures. In *Proceedings of the 1990 deloitte and touche/university of kansas symposium of auditing problems, lawrence, KS* (pp. 29–58).
Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research, 12*, 1–25.
Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123–140.
Cadden, D. T. (1991). Neural networks and the mathematics of chaos-an investigation of these methodologies as accurate predictors of corporate bankruptcy. In *First international conference on artificial intelligence applications on wall street* (pp. 52–57). IEEE.
Chen, T., & He, T. (2015a). Higgs boson discovery with boosted trees. In *JMLR: Workshop and conference proceedings*.
Chen, T., & He, T. (2015b). xgboost: extreme gradient boosting. R package version 0.3-0. *Technical Report*.
Coats, P. K., & Fant, L. F. (1991). A neural network approach to forecasting financial distress. *The Journal of Business Forecasting, 10*, 9.
Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the twelfth international conference on machine learning* (pp. 115–123).
Constand, R. L., & Yazdipour, R. (2011). Firm failure prediction models: a critique and a review of recent developments. *Advances in Entrepreneurial Finance* (pp. 185–204). Springer.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*, 273–297.
Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 167–179.
Dimitras, A., Slowinski, R., Susmaga, R., & Zopounidis, C. (1999). Business failure prediction using rough sets. *European Journal of Operational Research, 114*, 263–280.
Edmister, R. O. (1972). An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative analysis, 7*, 1477–1493.
Fan, W., Stolfo, S. J., Zhang, J., & Chan, P. K. (1999). Adacost: Misclassification cost-sensitive boosting. In *Icml* (pp. 97–105).
Fitzpatrick, P. (1932). A comparison of ratios of successful industrial enterprises with those of failed firms. *Certified Public Accountant, 2*, 598–605.
Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *Icml* (pp. 148–156).
Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.
Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*, 367–378.
Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 42*, 463–484.
Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed chinese companies using data mining. *European Journal of Operational Research, 241*, 236–247.
Grice, J. S., & Dugan, M. T. (2003). Re-estimations of the Zmijewski and Ohlson bankruptcy prediction models. *Advances in Accounting, 20*, 77–93.

Ho, T. K. (1995). Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on* (pp. 278–282). IEEE.

Hopwood, W., McKeown, J. C., & Mutchler, J. F. (1994). A reexamination of auditor versus model accuracy within the context of the going-concern opinion decision. *Contemporary Accounting Research, 10*, 409–431.

Johnson, R., & Zhang, T. (2011). Learning nonlinear functions using regularized greedy forest. *Machine Learning* arXiv preprint arXiv:1109.0887.

Ketz, J. E. (1978). The effect of general price-level adjustments on the predictive ability of financial ratios. *Journal of Accounting Research, 16*, 273–284.

Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20*, 226–239.

Koh, H. C. (1991). Model predictions and auditor assessments of going concern status. *Accounting and Business Research, 21*, 331–338.

Koh, H. C., & Killough, L. N. (1990). The use of multiple discriminant analysis in the assessment of the going-concern status of an audit client. *Journal of Business Finance & Accounting, 17*, 179–192.

Koster, A., Sondak, N., & Bourbia, W. (1991). A business application of artificial neural network systems. *Journal of Computer Information Systems, 31*, 3–9.

Laitinen, E. K. (1991). Financial ratios and different failure processes. *Journal of Business Finance & Accounting, 18*, 649–673.

Li, M. Y. L., & Miu, P. (2010). A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based information: A binary quantile regression approach. *Journal of Empirical Finance, 17*, 818–833.

Libby, R. (1975). Accounting ratios and the prediction of failure: Some behavioral evidence. *Journal of Accounting Research, 13*, 150–161.

Merwin, C. L. (1942). Financing small corporations in five manufacturing industries. *NBER Books* p. 1926-1936. New York: National Bureau of Economic Research, Inc.

Meyer, P. A., & Pifer, H. W. (1970). Prediction of bank failures. *Journal of Finance, 24*, 853–868.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 36*, 3028–3033.

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

Pettway, R. H., & Sinkey, J. F. (1980). Establishing on-site bank examination priorities: An early-warning system using accounting and market information. *Journal of Finance, 35*, 137–150.

Platt, H. D., & Platt, M. B. (1990). Development of a class of stable predictive variables: the case of bankruptcy prediction. *Journal of Business Finance & Accounting, 17*, 31–51.

Platt, H. D., Platt, M. B., & Pedersen, J. G. (1994). Bankruptcy discrimination with real variables. *Journal of Business Finance & Accounting, 21*, 491–510.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA, USA: Morgan Kaufman Publishers.

Rujoub, M. A., Cook, D. M., & Hay, L. E. (1995). Using cash flow ratios to predict business failures. *Journal of Managerial Issues, 7*, 75–90.

Salchenberger, L. M., Cinar, E., & Lash, N. A. (1992). Neural networks: A new tool for predicting thrift failures. *Decision Sciences, 23*, 899–916.

Serrano-Cinca, C. (1996). Self organizing neural networks for financial diagnosis. *Decision Support Systems, 17*, 227–238.

Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications, 28*, 127–135.

Sinkey, J. F. (1975). A multivariate statistical analysis of the characteristics of problem banks. *The Journal of Finance, 30*, 21–36.

Tam, K. Y. (1991). Neural network models and the prediction of bank bankruptcy. *Omega, 19*, 429–445.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. *Management science, 38*, 926–947.

Tomczak, J. M., & Zięba, M. (2015). Probabilistic combination of classification rules and its application to medical diagnosis. *Machine Learning, 101*, 105–135.

Tomczak, S. (2014a). Comparative analysis of liquidity ratios of bankrupt manufacturing companies. *Business and Economic Horizons (BEH), 10*, 151–164.

Tomczak, S. (2014b). Comparative analysis of the bankrupt companies of the sector of animal slaughtering and processing. *Equilibrium. Quarterly Journal of Economics and Economic Policy, 9*, 59–86.

Tomczak, S. (2014c). The early warning system. *Journal of Management and Financial Sciences, 7*, 51–74.

Wilcox, J. W. (1973). A prediction of business failure using accounting data. *Journal of Accounting Research, 11*, 163–179.

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

Winakor, A., & Smith, R. F. (1935). Changes in financial structure of unsuccessful industrial companies. *Bureau of Business Research, Bulletin, 51*.

Zavgren, C. (1983). The prediction of corporate failure: the state of the art. *Journal of Accounting Literature, 2*, 1–38.

Zhang, G., Hu, M. Y., Patuwo, B. E., & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research, 116*, 16–32.

Zhang, Y., Wang, S., & Ji, G. (2013). A rule-based model for bankruptcy prediction based on an improved genetic ant colony algorithm. *Mathematical Problems in Engineering, 2013*.

Zięba, M., Tomczak, J. M., Lubicz, M., & Świątek, J. (2014). Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing, 14*, 99–108.

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