Economic Modelling 84 (2020) 165–176

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.journals.elsevier.com/economic-modelling

Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets

Martin Zoričák [a], Peter Gnip [b], Peter Drotár [b,*], Vladimír Gazda [a]

[a] Department of Finance, Faculty of Economics, Technical University of Kosice, Bozeny Nemcovej 32, 04200, Kosice, Slovak Republic
[b] Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Letna 9, 04200, Kosice, Slovak Republic

ARTICLE INFO

2010 MSC:
00-01
99-00

Keywords:
Bankruptcy
Imbalanced learning
Anomaly detection
Annual reports

ABSTRACT

Bankruptcy prediction is still important topic receiving notable attention. Information about an imminent bankruptcy threat is a crucial aspect of the decision-making process of managers, financial institutions, and government agencies. In this paper, we utilize a newly acquired dataset comprising financial parameters derived from the annual reports of small- and medium-sized companies. The data, which reveal the true ratio between bankrupt and non-bankrupt companies, are severely imbalanced and only contain a small fraction of bankrupt companies. Our solution to overcome this challenging scenario of imbalanced learning was to adopt three one-class classification methods: a least-squares approach to anomaly detection, an isolation forest, and one-class support vector machines for comparison with conventional support vector machines. We provide a comprehensive analysis of the financial attributes and identify those that are most relevant to bankruptcy prediction. The highest prediction performance in terms of the geometric mean score is 91%. The results are validated on two datasets from the manufacturing and construction industries.

1. Introduction

The recent financial crisis showed the increasing vulnerability of firms involved in complex business relations, relations with financial institutions, obligations toward tax agencies, etc. The threat of financial contagion is rising with the growing complexity of the economy. The latter experience brought evidence of the fragile financial stability of numerous firms. These companies are prone to turbulent financial shocks with their origins in the external environment. Even though many studies have been devoted to bankruptcy prediction, a general methodology that would enable a firm to identify business partners in financial distress has not yet been proposed.

The uniqueness of the bankruptcy prediction problem can be found in the nature of the data that are the subject of analysis. The majority of studies are based on a variety of financial ratios that are derived from annual financial statements. The annual financial statements usually consist of two documents –- the balance sheet and income statement. The first contains information regarding the assets, liabilities, and owners' equity, whereas the income statement considers the costs, revenues, and eventual profit or loss. Because the frequency of data is annual, the information in the financial ratios is condensed and may conceal important fluctuations between two reporting periods. The quality of data is usually determined by the type of companies included in the analysis. In general, larger firms or firms listed on the stock exchange are more likely to disclose more information (Firth, 1979), thereby allowing a more meaningful analysis of their current financial condition. On the other hand, the accounting records of small- and medium-sized companies (SME) are not that complete and precise; therefore, the financial condition of the company may not be completely reflected by the records. Small- and medium-sized companies are integral parts of the economy and secure large share of overall employment (De Wit and De Kok, 2014). Financial reporting of small- and medium-sized companies is subject of standardization according IFRS in many jurisdictions (Chand et al., 2015). Nevertheless, quality of financial reporting of small- and medium-sized companies is relatively low (Chen et al., 2011). The reason might be that, they are, in general, private companies, thus there is no pressure form outside of a company. There are, however, banks and business partners who need

* Corresponding author.
E-mail addresses: martin.zoricak@tuke.sk (M. Zoričák), peter.gnip@student.tuke.sk (P. Gnip), peter.drotar@tuke.sk (P. Drotár), vladimir.gazda@tuke.sk (V. Gazda).

https://doi.org/10.1016/j.econmod.2019.04.003
Received 18 October 2018; Received in revised form 8 February 2019; Accepted 3 April 2019
Available online 15 April 2019
0264-9993/

to assess financial stability of a company. Usually they have to relay on publicly accessible data of potential business partner. Aim of our study is to provide tool for predicting bankruptcy as accurate as possible on authentic available data.

Historical experience led to the introduction of some empirical ad hoc rules for financial distress assessment. This includes the so-called golden balance rule, which states that fixed assets should be covered by fixed capital (equity and long-term debt). The second rule describes the optimal capital structure. Modigliani and Miller (Modigliani and Miller, 1958) studied the optimal ratio between debt and equity in relation to the minimal cost of capital. Later, increasing computational power and a rise in disposable financial datasets allowed the application of more advanced statistical methods. A seminal study by Altman (1968) considered multivariate discrimination analysis to quantify critical values of the so-called Z-score to identify the threat of bankruptcy. Other significant work from that period is that of Beaver (1966) and Tamari (1966). Other studies of bankruptcy used a variety of statistical procedures such as linear discriminant analysis, logistic regression, and factor analysis. Survey of these methods is provided by Dimitras et al. (Dimitras et al., 1996). Owing to their simplicity, these methods are uncritically used by many regardless of changing economic conditions (e.g., different phases of an economic cycle, different inflation levels, factors specific of a national currency, etc.) (Nwogugu, 2007).

In last two decades, popularity of bankruptcy predictions methods has shifted from statistical to intelligent ones. Authors (Kumar and Ravi, 2007) compared papers employing methods from those two groups and conclude that intelligent methods outperform statistical methods. Intelligent methods are able to take into account large number of attributes and to assess more complex relationships among them.

From a machine-learning point of view, bankruptcy prediction is a binary classification task. However, considering a real-world scenario, even in the most burdensome times the number of bankrupt companies constitutes only a fraction of all operating companies. In fact, our data show that only two percent (or less) of all considered companies went bankrupt. This means that one of the two classes (i.e., bankrupt) is underrepresented compared to the other class (i.e., non-bankrupt). Most of the conventional machine-learning approaches would be severely affected by this distribution of data and could not be used.

There are two main approaches to deal with imbalanced datasets. The first approach attempts to create the more balanced distribution of classes through some preprocessing techniques. The preprocessing techniques fail into three categories: undersampling, oversampling and hybrid representing the combination of these two. For highly imbalanced datasets with only few minority samples the oversampling is the recommended approach (Haixiang et al., 2017). Apparent disadvantage of the oversampling is creation of the new minority samples that poses additional demands on computational resources. The second approach to cope with imbalanced data utilizes learning algorithms that are capable to successfully learn from imbalanced data. This includes ensemble classifiers, cost-sensitive learning, algorithmic classifier modification and other methods for imbalanced learning such as one-class learning and outlier detection. Particularly, one-class learning is suggested as the preferred approach for extremely imbalanced datasets in Raskutti and Kowalczyk (2004) and Haibo He and Garcia (2009). A several comprehensive reviews of imbalanced learning were published so we recommend those for the more interested reader (Haixiang et al., 2017; Krawczyk, 2016; Haibo He and Garcia, 2009). We employ one-class methods since these were recommended by some authors for strongly imbalanced scenarios such as ours (Haibo He and Garcia, 2009).

Authors (Zhou, 2013) in their study analyzed six sampling methods for imbalanced data of corporate bankruptcy. They concluded, that for low number of bankrupt companies in sample SMOTE achieved best results. Among five different prediction models SVM achieved best performance. In paper Kim et al. (2015) authors propose to use GMBoost on imbalanced datasets. They argue this method brings best results than AdaBoost or CostBoost. Further, they used SMOTE to re-sample date which lead to additional increase in predictive capacities. One of the latest study (Veganzones and Severin, 2018) focuses on the level of imbalance in bankruptcy prediction data. According their findings, if only less than 20% training data represents bankrupt companies prediction ability is significantly lowered. To overcome this problem, they employed multiple oversampling and undersampling methods. Their findings confirmed, that SVM is the least sensitive technique to imbalanced data. However, in case of strong imbalance the predictive performance is severely affected. None of mentioned studies focuses specifically on the small- and medium-sized companies data.

Another promising approach is application of hazard duration models to bankruptcy prediction. A wide variety of the survival models offered investigators an ample space to precise model specifications under various conditions. The basic answer of the continuous vs. discrete-time model dilemma in SME bankruptcy prediction was resolved Gupta et al. (2015) indicating the discrete-time model superiority. This opinion is supported by the presence of the data discretized into relatively short annual time series, i.e., conditions preventing the proper application of the continuous-time models. However, as the most of the works on bankruptcy also this study deals with balanced or almost balanced data. The studies focusing on severely imbalanced data are quite scarce. Another limitation is that many previous studies mostly focus on bigger companies or companies indexed on stack exchange. The analysis of small- and medium-sized companies start to appear only recently and there are still open issues to investigate.

With this study we aim at the area of bankruptcy prediction that is less investigated: small- and medium-sized companies in highly imbalanced scenario. We collected an extensive dataset consisting of thousands of limited liability companies from two business areas: construction and manufacturing. The dataset contains 20 financial ratios that are derived from these companies' annual reports. The uniqueness of the dataset lies in the number of companies we included, the realistic imbalanced distribution of bankrupt and non-bankrupt companies and the fact that we cover small- and medium-sized companies. The proposed methodology is transferable to all companies that need to provide annual reports and not only companies that are indexed with the stock exchange.

The results obtained with the proposed model are very similar on the datasets from both industries, and this conveys some confidence in our results.

The remainder of the paper is organized as follows. In the next section, we briefly review related papers on bankruptcy prediction. Then, we describe the dataset, which was compiled by the authors, is novel, and was not published before. In the fourth section, we provide details of our preliminary statistical analysis of the data, after which we present an in-depth analysis of the importance of different features on the target variable through feature selection techniques. Finally, we apply several one-class classifiers together with conventional support vector machines and identify the best-performing model. The discussion and conclusions are provided in the last section.

2. Literature review of machine-learning methods for bankruptcy prediction

In the last decades, the development of machine learning drew the attention of economists and the field of bankruptcy prediction is no exception. Predicting the bankruptcy of a firm may be approached as a classification problem, which consists of, in general, two classes: bankrupt and non-bankrupt. The popularity and importance of this topic is reflected in a large number of papers summarized in several recent reviews (Kumar and Ravi, 2007; Lin et al., 2012; Sun et al., 2014; Alaka et al., 2018). Despite the ever-increasing variety of intelligent methods, there are four techniques that are discussed in all mentioned reviews: Neural networks (NNs), Decision trees (DTs), Case-based reasoning (CBR), and support vector machines (SVMs). These are the most

Table 1
Distribution of the data among the bankrupt and non-bankrupt classes.

	2013	2014	2015	2016
construction	25/1205	30/1418	20/1749	14/2174
manufacturing	30/4077	30/4450	26/5019	14/5840

frequently used methods that achieve variable results depending on the data under investigation.

The most popular method, which has been in use since the 1990s, is neural networks. The latest review paper by Alaka (Alaka et al., 2018) listed 38 papers in which the authors used some form of NN. Multiple architectures of NN were applied in the field of bankruptcy prediction (e.g., multilayer perceptrons (Iturriaga and Sanz, 2015), back-propagation neural networks (Lee and Choi, 2013), and probabilistic neural networks (Yang et al., 1999)). The conventional NN method belongs to the earliest machine-learning methods used to predict bankruptcy, thus it serves as a benchmark for other ML methods. Although NN models are highly accurate, they are often described as "black-box." This issue was addressed in a study (Olden and Jackson, 2002), in which the authors argued that it is possible to uncover the underlying processes hidden in such a model.

Another frequently used machine-learning method for bankruptcy prediction is DTs. In a seminal paper about this method (Sung et al., 1999), the authors compered DTs with the following methods: discriminant analysis, genetic algorithms, and NN. The results showed that the DT method provides interpretable results. Other studies employing this method are for instance (Lee et al., 2006) and (Yeh and Lien, 2009).

In general, CBR is based on previous experiences - cases that create precedence for solving similar problems in the future. Usually models based on the CBR method use the Euclidean distance and k-nearest neighbor method. CBR is more suited to smaller data samples and is similar to human decision making (Kumar and Ravi, 2007). Even though an initial study (Jo et al., 1997) suggested that CBR is not a suitable method for bankruptcy prediction, several later studies (Chuang, 2013; Ahn and Kim, 2009; Li and Sun, 2008, 2009; Sartori et al., 2016) achieved results showing that the prediction performance was comparable with other ML methods.

SVM gained popularity for bankruptcy prediction in the late 2000s (Lin et al., 2012). Several papers (Min and Lee, 2005; Shin et al., 2005; Min et al., 2006; Ding et al., 2008; Chaudhuri and De, 2011; Wang and Ma, 2012; Li et al., 2015) compared the prediction accuracy of SVM with that of NN (and other methods) and the results of all the aforementioned studies suggest that the performance of SVM is superior. The advantages of using this method are, however, offset by the non-transparency of the model, which may be confusing for an audience unfamiliar with machine learning (Alaka et al., 2018).

These baseline methods can be combined in several ways in order to boost the accuracy or to overcome certain shortcomings of individual classifiers. There are two basic approaches for combining different classifiers – the ensemble and hybrid methods. The ensemble approach divides the initial problem into smaller sub-problems, which are solved by individual classification algorithms. The results of the base classifiers are then combined. Multiple bankruptcy prediction models applied an ensemble approach (e.g., Liao et al. (2014); Zhu et al. (2017)). The hybrid approach combines different classification techniques sequentially (Azayite and Achchab, 2016; Ding et al., 2008).

As in many other domains, nature inspired algorithms were successfully used also for bankruptcy prediction (Gordini, 2014; Wang et al., 2017).

In practice, the number of bankrupt companies is noticeably smaller than the number of non-bankrupt companies. However, this fact is frequently neglected in many papers and balanced data are considered. Methods for bankruptcy should definitely take into account this imbalance in order to prevent errors of type I and II, according to which a non-bankrupt company is evaluated as bankrupt and vice versa. A few studies have already considered the issue of imbalanced data for bankruptcy prediction (Haibo He and Garcia, 2009; Sun et al., 2018; Zhou, 2013; Li and Sun, 2012).

3. Data

The dataset consists of thousands of records of business entities operating in the Slovak republic (member state of EU, Schengen area, and Eurozone) during the years 2010–2016. Each company is characterized by 20 financial attributes derived from the company's annual report. The financial attributes and their descriptive statistics are listed in Table 2. According to the EU classification (Commission of the European Communities, 2003) the majority of analyzed firms are small- or medium-sized companies; precisely, it is 94.49% for manufacturing and 99.09% for construction.

It should be noted that, especially small companies do not maintain their accounting records precisely and this may result in the occurrence of some data outliers. However, these imperfections naturally exist in

Table 2
Descriptive statistics of analyzed financial ratios.

Category	Financial Ratio	Construction				Manufacturing			
		Q25	Median	Q75	Mean	Q25	Median	Q75	Mean
Activity	Total Asset Turnover (TAT)	131.05	229.64	431.44	20,165.56	149.94	233.34	378.56	2809.57
	Asset Turnover Days (ATD)	29.25	72.75	146.81	4454.34	29.64	56.43	98.42	843.73
	Days Total Receivables Outstanding (DTR)	41.81	94.42	195.24	6273.94	40.61	75.54	146.18	849.40
	Inventory Turnover Days (ITD)	0.42	6.89	34.39	22,155.36	6.58	26.53	62.42	118.27
Liquidity	Cash Ratio (L1)	0.11	0.47	1.86	13.69	0.07	0.34	1.26	4.74
	Quick Ratio (L2)	0.81	1.31	3.35	16.91	0.63	1.17	2.57	7.12
	Current Ratio (L3)	0.96	1.43	3.61	17.82	0.94	1.49	3.17	7.82
Return Of Assets (ROA)	−0.23	2.64	11.60	−10.15	0.06	3.75	13.03	−7.20	
Profitability	Return On Equity (ROE)	0.40	13.93	38.77	2.72	0.86	13.20	37.23	−78.45
	Return On Sales (ROS)	−0.06	1.68	6.81	−277.76	0.11	2.40	8.02	118.54
	Return On Investment (ROI)	2.51	15.39	35.57	19.28	10.16	25.98	50.43	28.55
	Labor-to-Revenue Ratio (LRR)	0.00	5.43	15.85	17.60	3.74	12.27	23.67	22.59
	Wages to added ratio (WAR)	0.00	29.61	63.38	14.06	15.26	48.99	72.80	27.81
Solvency	Debt-to-Assets Ratio (DA)	0.35	0.71	0.94	7.61	0.33	0.61	0.87	5.94
	Debt-to-Equity Ratio (DE)	0.22	1.52	6.36	63.33	0.32	1.23	4.22	58.06
	Financial Leverage (FL)	1.14	2.11	5.16	19.86	1.27	2.05	4.13	17.55
	Debt To Income Ratio (DIR)	25.34	63.79	87.66	83.93	29.06	57.59	81.62	78.32
	Debt Service Coverage Ratio (DCR)	2.06	12.14	43.75	−1053.86	5.92	19.22	50.40	73.83
	Asset Coverage Ratio (ACR)	0.54	1.24	2.80	69.70	0.71	1.27	2.55	7.20
	Bank Liabilities To Debt Ratio (BL)	0.00	0.00	0.00	3.85	0.00	0.00	5.15	5.61

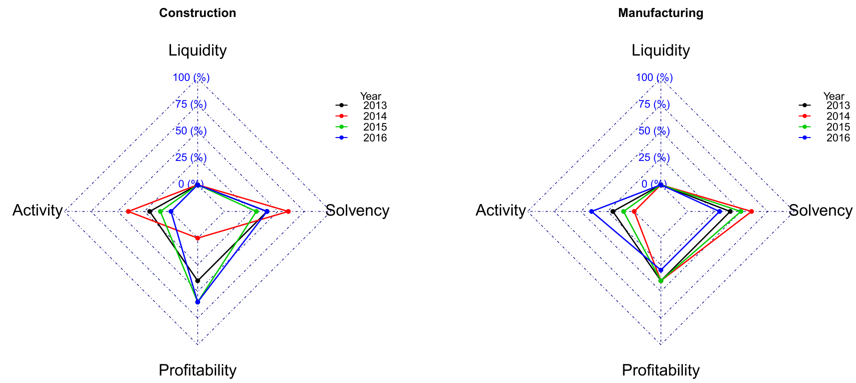


Fig. 1. Most important groups of financial ratios based on the Vote method. Result of the Vote method (FS analysis): the 10 most important financial ratios divided into four categories. The percentage represents the number of financial ratios within a given category that belong to the TŤop 10,´ i.e., the importance of a particular category. The significance of a particular financial ratio is assessed for one to three years prior to bankruptcy, therefore it may occur more than once.

real-world data; thus, any proposed methodology should be able to process these types of data.

The dataset contains data from the annual report of the company three years prior to the year of evaluation R. We consider four different years of evaluation R: 2013, 2014, 2015, and 2016. The economical attributes are available three years prior to each evaluation year i.e., $R-1, R-2, R-3$.

Two types of companies are included in the dataset: bankrupt and non-bankrupt companies. Bankruptcy is, in general, the consequence of financial distress, a situation in which a firm is not able to fulfill its obligations. Although the definition of bankruptcy differs among jurisdictions, two distinct statuses prevail. In the first, all company assets are liquidated in order to fulfill as much debt as possible and the company is no longer a going concern. The second status is in the form of reorganization, which involves the settlement of debt repayment between a company and its creditors while the company continues to exist. All companies operating in the Slovak republic within the time frame 2010 to 2016 are included. Naturally, the number of bankrupt companies is significantly smaller than the number of non-bankrupt companies. The concrete distribution of the data is presented in Table 1. The non-bankrupt class severely outrepresents the bankrupt company class. The fraction of bankrupt companies is less than 2%.

The data included in our analysis originate from two business areas construction and manufacturing. These data are part of a larger dataset that also covers other industries such as services, agronomy, and retail. However, we did not consider these other industries in this study because of the small number of samples (< 10) in their respective minority classes.

4. Preliminary statistical analysis

We obtained initial insight into the data by employing the t-distributed stochastic neighbor embedding (tSNE) (Maaten and Hinton, 2008) approach to dimensionality reduction and visualization of the data. This allows the data to be visualized in the form of a two-dimensional map. In this regard, tSNE was shown to provide improved visualizations than other methods in the case of two-dimensional maps (Maaten and Hinton, 2008) and it was successfully used in different domains (Akcay et al., 2018; Kim and Cho, 2018). tSNE is an extension of stochastic neighbor embedding (Hinton and Roweis, 2002), which basically converts high-dimensional distances between data points in Euclidean space to conditional probabilities that represent similarities. The aim of tSNE is to find a low-dimensional data representation that minimizes the mismatch between the conditional probabilities of data points in the high-dimensional space and low-dimensional dataspace. tSNE achieves this through minimization of a single Kullback-Leibler divergence between the joint probability distribution in the high-dimensional and low-dimensional space. The cost function of tSNE is optimized by using an improved gradient descent procedure, where the improvement lies in adding 12 additional penalties and so-called early exaggeration.

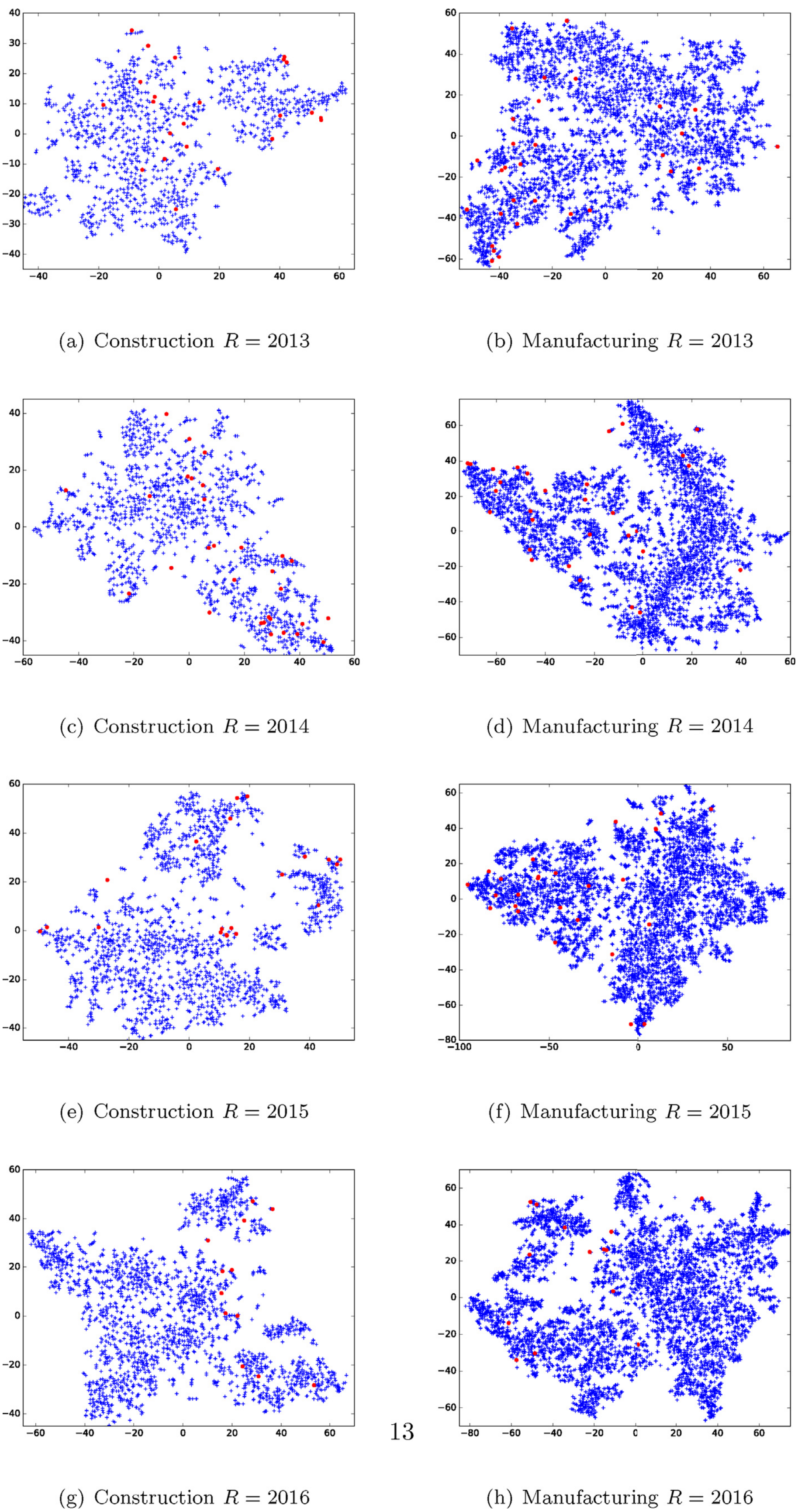
The visualization of both of the datasets pertaining to the construction and manufacturing industries is provided in Fig. 2. The figures depict the tSNE map of data for a particular year R by considering financial variables of three years prior to the evaluation year ($R-3$ & $R-2$ & $R-1$) i.e., 60 variables. All maps share several patterns. First, it is possible to isolate areas containing only non-bankrupt companies. Therefore, we hypothesize that certain rules can be induced to characterize some of the non-bankrupt companies even though the mapped pattern changes slightly for different evaluation years, which means that the attributes vary through the years. This is expected because some of the variables also depend on the specific situation in the market or in the business area, which changes from year to year. Another observation is that the data points representing bankrupt companies lie within clouds of non-bankrupt companies and as such it is difficult to isolate a bankrupt company by using a simple linear classifier. Last, even though some of the bankrupt companies tend to occupy an outlying position, no typical outliers are visible.

The fact that identification of a firm heading toward bankruptcy is not a straightforward process is confirmed by the large number of studies of this topic. Possible reasons why signs of bankruptcy are not more visible in advance are that a firm may intentionally conceal facts about its decline in its annual reports or the process of decay may be shorter than the reporting frequency.

5. Analysis of feature importance

Twenty features can be derived from the annual report. We take into account the three years prior to the evaluation year, which yield 60 financial attributes altogether. Observing and tracking all 60 attributes is not a trivial task for a human observer. Moreover, we assume that some features are more closely related to the target variable than others. Therefore, it is convenient to have smaller group of features that are relevant for bankruptcy prediction. These features can be used as initial indicators of potential bankruptcy.

Basically, there are two approaches that allow the formulation of a smaller group of significant features: feature extraction and feature selection (Garca et al., 2016). Feature extraction transforms existing features into new low-dimensional feature space. Examples are the visu-



(a) Construction $R = 2013$

(b) Manufacturing $R = 2013$

(c) Construction $R = 2014$

(d) Manufacturing $R = 2014$

(e) Construction $R = 2015$

(f) Manufacturing $R = 2015$

(g) Construction $R = 2016$

(h) Manufacturing $R = 2016$

Fig. 2. tSNE visualization of construction and manufacturing datasets for different evaluation years considering data from all three years prior to evaluation year.

alizations that are presented in the previous section. Even though this is very useful for visualization purposes, and there are other advantages such as preventing the course of dimensionality, because a feature extraction approach creates new features. The apparent drawback is the difficulty to link new features to original ones and as such the new features are not a direct representation of the known financial attributes. Another approach, feature selection, focuses on identification of the most relevant features for the prediction of the target variable and removes the irrelevant features. Feature selection (FS) is a very lively area of research to which a wide variety of methods can be applied. In general, FS methods can be divided into filter, wrapper, and embedded techniques (Bolon-Canedo et al., 2015). We apply FS to identify the most relevant features. Obtaining a smaller, compact group helps the human observer to more effectively analyze and evaluate results. Moreover, once we know which features are the most important indicators of bankruptcy we can focus on particular financial attributes and determine the probable cause of potential bankruptcy.

The “no free lunch” theorem also applies to feature selection methods, i.e., no single method exists that would dominate in terms of performance over others. In order to provide robust feature selection we employ four filter and one wrapper FS techniques. Each of these techniques is based on a different theoretical background therefore they provide diverse opinions on the selection of the most important features (Ang et al., 2016). Particularly, we used three supervised filter FS techniques: tree-based FS (Geurts et al., 2006), fisher score FS, reliefF method (Robnik-Šikonja and Kononenko, 2003), unsupervised lap score FS method (He et al., 2005), and recursive feature elimination (RFE) based on an SVM estimator (Guyon et al., 2002).

We applied five FS methods to the data on a per year basis, i.e., FS was applied to data from year $R-3$ to $R-1$ prior to the evaluation year R. Every evaluation year is treated separately. The features selected for the construction industry are provided in Table 3 and those for the manufacturing industry in Table 4. The ten most important features are listed for each method. These are ordered in the order of importance with one exception being the RFE FS, which, rather than providing the ranking of features, only a group of selected features is obtained. Additionally to the five FS methods, we also selected features based on the vote of all FS methods (ensemble FS) per particular year (last column in Tables 3 and 4). The voting procedure is as follows. The each feature that was selected by concrete FS methods obtains one point. The scores from all FS methods were summed, and the features obtaining the highest score were selected by the ensemble FS, i.e., these are the most frequently selected features per year. In the case of an equal score, ties were broken by random decision. As ensemble FS allows us to analyze the opinion of multiple FS methods at once, it is more robust to divergence that can be experienced by single FS. The feature set selected by ensemble FS enables us to identify a group of features that are selected in at least three out of four evaluated years for both the manufacturing and construction datasets. These are DIR_{R-1}, BL_{R-1}, ROS_{R-1}, and BL_{R-3}. The fact that these features were selected multiple times independently for both businesses means that this selection was made with a high degree of confidence. The features that were selected at least for two evaluation years and both considered businesses by ensemble FS are ROA_{R-1}, DA_{R-1}, and ROI_{R-1}. A possible reason why the list of most important features changes over time is that other information (e.g., non-financial data or information pertaining to the economy as a whole) is not taken into consideration. Another possible reason is the different cause of bankruptcy for the firms analyzed in a particular year. The last row of the last column contains the features that were selected the most frequently for the entire dataset by all methods through all evaluated years. The seven features (DIR_{R-1}, BL_{R-1}, ROS_{R-1}, BL_{R-3}, ROA_{R-1}, DA_{R-1}, ROI_{R-1}) are the same for the construction and manufacturing industries. Going beyond the first ten features shown in the table reveals that WAR_{R-1} has the same score TAT_{R-2}, but since the ties are broken randomly TAT_{R-2} was selected by chance. Similarly, BL_{R-2} and TAT_{R-2}, both of which were selected for the manufacturing dataset but not for construction, occupy the next position after the top ten features. Based on this, we can conclude that the attributes that were selected as significant are highly similar for both datasets, thereby validating the achieved results.

Regarding the differences in the selected features, we need to consider the issue of feature selection (in)stability (Kalousis et al., 2007). It was already proved several times that applying FS methods to slightly changed data yields different results (Fakhraei et al., 2014; Drotar et al., 2015). Some methods provide relatively stable results, although the output of others can vary to a greater extent. Therefore,

Table 3
Selected features. Construction dataset.

year	tree	fisher	reliefF	lap	svm	ensemble choice
2013	ROA_{R-1}, ROI_{R-1},	ROA_{R-1}, ROI_{R-1},	BL_{R-3}, BL_{R-2},	DTR_{R-3}, DCR_{R-3},	LRR_{R-2}, ROA_{R-1},	BL_{R-1}, ROI_{R-1}
	DA_{R-1}, ACR_{R-1},	BL_{R-1}, ROI_{R-3},	BL_{R-1}, DIR_{R-3},	DIR_{R-1}, ATD_{R-3},	ROS_{R-1}, LRR_{R-1},	ROA_{R-1}, ROI_{R-3}
	WAR_{R-1}, BL_{R-1},	ROA_{R-3}, DA_{R-3},	DA_{R-3}, ITD_{R-3},	DA_{R-1}, TAT_{R-1},	$L2_{R-3}, L3_{R-3}$,	ACR_{R-1}, DA_{R-1},
	ROS_{R-1}, ROE_{R-1},	DIR_{R-3}, ACR_{R-1},	DCR_{R-1}, WAR_{R-2},	ATD_{R-1}, TAT_{R-3},	TAT_{R-3}, TAT_{R-2},	ROS_{R-1}, BL_{R-3}
	ROI_{R-3}, ROE_{R-3},	DIR_{R-2}, BL_{R-2},	ROI_{R-3}, FL_{R-2},	TAT_{R-2}, ATD_{R-2},	ITD_{R-1}, ROI_{R-1},	TAT_{R-2}, TAT_{R-3}
2014	DIR_{R-1}, ROA_{R-1},	ACR_{R-1}, DTR_{R-2},	BL_{R-1}, BL_{R-3},	DTR_{R-1}, TAT_{R-1},	ROA_{R-3}, DCR_{R-1},	DIR_{R-1}, BL_{R-1},
	ROS_{R-1}, ACR_{R-1},	DIR_{R-1}, ATD_{R-2},	BL_{R-2}, ROS_{R-1},	ATD_{R-1}, DIR_{R-2},	DIR_{R-1}, ROI_{R-1},	FL_{R-1}, TAT_{R-2},
	FL_{R-2}, ATD_{R-1},	TAT_{R-2}, ROS_{R-1},	WAR_{R-3}, ROE_{R-1},	DA_{R-2}, DA_{R-1},	DA_{R-1}, ITD_{R-1},	DA_{R-1}, ITD_{R-1}
	ITD_{R-1}, WAR_{R-1},	ACR_{R-2}, BL_{R-1},	ACR_{R-3}, WAR_{R-1},	TAT_{R-2}, ATD_{R-2},	WAR_{R-2}, LRR_{R-1},	ACR_{R-3}, ATD_{R-2}
	BL_{R-1}, ROE_{R-2},	ACR_{R-3}, BL_{R-3},	ACR_{R-2}, DIR_{R-1},	TAT_{R-3}, ATD_{R-3},	ROI_{R-3}, ITD_{R-2},	TAT_{R-2}, BL_{R-3},
2015	DIR_{R-1}, ROS_{R-1},	ROS_{R-1}, DIR_{R-1},	DA_{R-1}, ROS_{R-1},	TAT_{R-2}, DTR_{R-2},	ROA_{R-3}, DCR_{R-1},	DIR_{R-1}, ROS_{R-1},
	WAR_{R-1}, ATD_{R-1},	LRR_{R-1}, WAR_{R-1},	LRR_{R-1}, ROI_{R-1},	DCR_{R-3}, ATD_{R-2},	$DIR_{R-1}, L1_{R-1}$,	BL_{R-3}, WAR_{R-1},
	LRR_{R-2}, DTR_{R-1},	ROS_{R-2}, ROA_{R-1},	ROA_{R-1}, BL_{R-3},	ACR_{R-1}, DIR_{R-3},	ROI_{R-3}, ITD_{R-3},	ROS_{R-2}, LRR_{R-1},
	ITD_{R-1}, WAR_{R-3},	ATD_{R-1}, BL_{R-3},	BL_{R-2}, BL_{R-1},	TAT_{R-3}, DA_{R-2},	ITD_{R-2}, ROS_{R-3},	BL_{R-1}, ROA_{R-1},
	BL_{R-1}, DA_{R-3},	DA_{R-1}, ROI_{R-1},	DIR_{R-1}, ROS_{R-2},	FL_{R-3}, DE_{R-3},	$L2_{R-3}, L3_{R-3}$,	ROI_{R-1}, ATD_{R-3},
2016	DIR_{R-1}, DE_{R-1},	DE_{R-1}, DE_{R-2},	DIR_{R-1}, DIR_{R-2},	TAT_{R-1}, TAT_{R-3},	ACR_{R-3}, ROS_{R-3},	DIR_{R-1}, DIR_{R-2}
	ROA_{R-1}, FL_{R-2},	DIR_{R-1}, WAR_{R-1},	BL_{R-3}, BL_{R-1},	ATD_{R-1}, DCR_{R-3},	$ITD_{R-1}, L2_{R-3}$,	ROA_{R-1}, WAR_{R-1},
	DTR_{R-1}, WAR_{R-1},	ROA_{R-1}, DIR_{R-2},	ROI_{R-1}, BL_{R-2},	DTR_{R-3}, ATD_{R-3},	$L3_{R-3}, DCR_{R-1}$,	BL_{R-3}, ROE_{R-2},
	DIR_{R-2}, LRR_{R-1},	ROE_{R-2}, BL_{R-3},	ROA_{R-1}, FL_{R-2},	ACR_{R-2}, DTR_{R-1},	DIR_{R-1}, ITD_{R-3},	LRR_{R-1}, ROI_{R-1},
	DCR_{R-1}, ROE_{R-2},	DA_{R-2}, LRR_{R-2},	LRR_{R-2}, ROI_{R-2},	LRR_{R-1}, ACR_{R-1},	$ROI_{R-1}, L2_{R-1}$,	DCR_{R-1}, LRR_{R-2}
best per FS	WAR_{R-1}, ROA_{R-1},	ROA_{R-1}, BL_{R-3},	BL_{R-2}, BL_{R-1},	TAT_{R-3}, TAT_{R-2},	$ROI_{R-1}, L2_{R-3}$,	DIR_{R-1}, BL_{R-1},
	ROS_{R-1}, BL_{R-1},	DIR_{R-1}, WAR_{R-1},	DIR_{R-1}, BL_{R-3},	ATD_{R-2}, ATD_{R-1},	$L3_{R-3}, ITD_{R-1}$,	ROA_{R-1}, ROI_{R-1},
	DIR_{R-1}, ACR_{R-1},	ROS_{R-1}, BL_{R-1},	ROA_{R-1}, ROS_{R-1},	ATD_{R-3}, DCR_{R-3},	DCR_{R-1}, DIR_{R-1},	ROS_{R-1}, WAR_{R-1},
	DTR_{R-1}, ATD_{R-1},	ROI_{R-1}, ACR_{R-1},	ROI_{R-1}, FL_{R-2},	TAT_{R-1}, ACR_{R-1},	DIR_{R-3}, ITD_{R-2},	DA_{R-1}, ATD_{R-1},
	FL_{R-2}, ITD_{R-1},	DIR_{R-2}, LRR_{R-1},	DIR_{R-3}, DA_{R-1},	DA_{R-2}, DTR_{R-1},	LRR_{R-1}, ITD_{R-2},	BL_{R-3}, LRR_{R-1},

Table 4
Selected features. Manufacturing dataset.

year	tree	fisher	reliefF	lap	svm	best per year
2013	ROA_{R-1}, ROI_{R-1},	ROA_{R-1}, ROI_{R-1},	LRR_{R-3}, BL_{R-2},	DTR_{R-3}, DCR_{R-3},	LRR_{R-2}, ROA_{R-1},	BL_{R-1}, ROI_{R-1},
	DA_{R-1}, ACR_{R-1},	BL_{R-1}, ROI_{R-1},	$BL_{R-1}, L2_{R-3}$,	$L2_{R-1}, ATD_{R-3}$,	ROS_{R-1}, LRR_{R-1},	ROA_{R-1}, ROI_{R-3},
	WAR_{R-1}, BL_{R-1},	ROA_{R-3}, DA_{R-3},	DA_{R-3}, ITD_{R-3},	DA_{R-1}, TAT_{R-1},	$L2_{R-3}, L3_{R-3}$,	ACR_{R-1}, ROS_{R-1},
	ROS_{R-1}, ROE_{R-1},	$L2_{R-3}, ACR_{R-1}$,	DCR_{R-1}, WAR_{R-2},	ATD_{R-3}, TAT_{R-3},	TAT_{R-3}, TAT_{R-2},	TAT_{R-3}, TAT_{R-2},
	ROI_{R-1}, ROE_{R-3},	$L2_{R-2}, BL_{R-2}$,	ROI_{R-1}, FL_{R-2},	TAT_{R-2}, ATD_{R-2},	ITD_{R-1}, ROI_{R-1},	BL_{R-2}, DA_{R-1}
2014	$BL_{R-1}, L2_{R-1}$,	BL_{R-2}, WAR_{R-2},	LRR_{R-2}, BL_{R-1},	ROS_{R-2}, DTR_{R-1},	$L2_{R-1}, L3_{R-1}$,	DIR_{R-1}, BL_{R-3},
	BL_{R-2}, ROS_{R-1},	BL_{R-1}, BL_{R-1},	BL_{R-1}, LRR_{R-1},	ATD_{R-3}, TAT_{R-1},	ROS_{R-3}, DCR_{R-2},	BL_{R-2}, ROS_{R-1},
	ACR_{R-2}, DE_{R-2},	ROA_{R-1}, ROI_{R-1},	BL_{R-2}, FL_{R-1},	$ROS_{R-1}, L2_{R-1}$,	$L1_{R-2}, L2_{R-3}$,	BL_{R-1}, ROS_{R-3},
	ITD_{R-1}, FL_{R-1},	$L2_{R-1}, ROS_{R-1}$,	ACR_{R-1}, DE_{R-2},	ROS_{R-3}, ATD_{R-3},	$L3_{R-2}, DCR_{R-1}$,	LRR_{R-2}, LRR_{R-1},
	BL_{R-1}, LRR_{R-3},	LRR_{R-1}, LRR_{R-2},	ITD_{R-3}, FL_{R-2},	$TAT_{R-2}, L2_{R-2}$,	$DA_{R-3}, L1_{R-1}$,	FL_{R-1}, FL_{R-3},
2015	ROA_{R-1}, BL_{R-1},	ROA_{R-1}, LRR_{R-2},	BL_{R-1}, BL_{R-1},	ROS_{R-3}, ATD_{R-2},	TAT_{R-1}, ACR_{R-1},	DIR_{R-1}, BL_{R-1},
	$LRR_{R-2}, L2_{R-1}$,	BL_{R-1}, ITD_{R-2},	LRR_{R-3}, ROI_{R-1},	TAT_{R-2}, DTR_{R-2},	$L2_{R-1}, DA_{R-2}$,	ROA_{R-1}, BL_{R-3},
	$L2_{R-2}, ROI_{R-1}$,	$BL_{R-2}, L2_{R-1}$,	BL_{R-2}, WAR_{R-1},	$ROS_{R-2}, L2_{R-2}$,	$L1_{R-2}, WAR_{R-2}$,	WAR_{R-1}, BL_{R-2},
	ROS_{R-2}, ROE_{R-2},	ROA_{R-2}, FL_{R-2},	ROA_{R-1}, FL_{R-3},	DE_{R-3}, ROA_{R-2},	DA_{R-1}, ROS_{R-1},	LRR_{R-2}, ROI_{R-1},
	BL_{R-1}, WAR_{R-1},	WAR_{R-1}, BL_{R-1},	$FL_{R-1}, L2_{R-1}$,	TAT_{R-3}, TAT_{R-1},	$L1_{R-1}, DCR_{R-2}$,	BL_{R-2}, TAT_{R-1}
2016	DA_{R-1}, DTR_{R-1},	DA_{R-1}, LRR_{R-3},	DA_{R-1}, BL_{R-1},	ATD_{R-3}, TAT_{R-3},	ROA_{R-3}, DCR_{R-2},	ROS_{R-1}, DIR_{R-1},
	$L2_{R-1}, ROS_{R-1}$,	TAT_{R-3}, ITD_{R-3},	BL_{R-1}, BL_{R-2},	$DTR_{R-3}, L2_{R-3}$,	ROS_{R-3}, ROS_{R-1},	ROS_{R-3}, DTR_{R-3},
	WAR_{R-2}, ROA_{R-1},	$ROS_{R-3}, L2_{R-1}$,	$L2_{R-1}, LRR_{R-1}$,	ROS_{R-3}, WAR_{R-1},	$L1_{R-2}, DTR_{R-3}$,	DA_{R-1}, DTR_{R-1},
	BL_{R-1}, FL_{R-2},	ATD_{R-3}, DTR_{R-1},	ROS_{R-1}, ROI_{R-2},	LRR_{R-3}, TAT_{R-2},	ITD_{R-3}, TAT_{R-1},	DA_{R-2}, LRR_{R-3},
	ROE_{R-1}, DA_{R-2},	ROS_{R-1}, DTR_{R-3},	ROA_{R-2}, ITD_{R-1},	DA_{R-2}, ROA_{R-3},	DTR_{R-1}, ROI_{R-1},	TAT_{R-3}, ITD_{R-3}
best per FS	ROS_{R-1}, BL_{R-3},	BL_{R-1}, BL_{R-2},	BL_{R-1}, BL_{R-2},	TAT_{R-2}, ATD_{R-3},	$L1_{R-2}, ROS_{R-1}$,	DIR_{R-1}, BL_{R-1}
	ROA_{R-1}, BL_{R-1},	ROA_{R-1}, DIR_{R-1},	BL_{R-3}, DIR_{R-1},	TAT_{R-3}, TAT_{R-1},	DCR_{R-2}, TAT_{R-1},	ROS_{R-1}, ROA_{R-1},
	DIR_{R-1}, ROI_{R-1},	ROS_{R-1}, BL_{R-3},	ITD_{R-2}, FL_{R-1},	ROS_{R-3}, ATD_{R-2},	$L1_{R-1}, ROS_{R-3}$,	BL_{R-2}, BL_{R-3},
	ROE_{R-1}, DA_{R-1},	LRR_{R-2}, ROI_{R-1},	LRR_{R-1}, LRR_{R-2},	DIR_{R-1}, DIR_{R-2},	DA_{R-1}, DIR_{R-3},	ROI_{R-1}, ROS_{R-3},
	WAR_{R-1}, DE_{R-2},	DA_{R-1}, ITD_{R-2},	FL_{R-2}, ROS_{R-1},	ATD_{R-1}, ROS_{R-2},	DCR_{R-1}, TAT_{R-2},	DA_{R-1}, TAT_{R-2},

in order to obtain exactly the same selection on different data the stability of the FS method would have to be 100% (and the data would have to comprise exactly the same pattern).

The majority of the most important features belong to the categories Profitability and Solvency and each of the analyzed industries has one feature in the Activity category. On the other hand, the financial ratios of the liquidity category are not considered as important from the overall perspective. The financial ratios from the Solvency category imply external capital sources e.g., the debt and financial ratios from the Profitability category imply a lower total return, which is probably insufficient to cover liabilities. The representation of different categories between the most relevant features is visualized in Fig. 1.

Another observation is that the most frequently selected features are derived from the data one year prior to the evaluation year. Actually, approximately one half (51% for construction business; 46% for manufacturing business) of the selected features are from the year $R-1$. This confirms our hypothesis that the signs of potential bankruptcy are the most escalated the year before the firm actually faces bankruptcy. In the following section on prediction accuracy, we further extend this analysis and evaluate the extent to which the data from different years preceding the evaluation year affect the prediction performance.

6. Bankruptcy prediction

Our main goal is to develop a prediction model capable of predicting the financial bankruptcy of a company as accurately as possible and as efficiently as possible, considering that it is equally important to avoid numerous false positives i.e., to identify financially healthy companies as being non-bankrupt. Because of the strongly imbalanced nature of our datasets, the choice of the selected methodology reflects the nature of the data. Detecting the bankrupt company is difficult when overlapping with cluster of non-bakcrupt companies data points. Conventional machine learning methods can easily fail in this scenario, since they have the tendency to classify all samples as a majority class. This is clearly visible in recent study on bankruptcy prediction on imbalanced data (Veganzones and Severin, 2018). As the imbalance increases the prediction performance quickly decreases. Therefore, we employ one-class classification methods, which are known to perform successfully in strongly imbalanced scenarios (Domingues et al., 2018) and several authors recommend to use these if data are severely imbalanced (Haibo He and Garcia, 2009). One class classification methods use only samples from the majority class to train the model. The anomalous samples in the test dataset are labeled as bankrupt.

6.1. Prediction methods

We compare three one-class classifiers: One Class SVM (OCSVM) (Schölkopf et al., 2001), Isolation Forest (Liu et al., 2012) (IF) and Least-Squares Anomaly Detection (Quinn and Sugiyama, 2014) (LSAD). Additionally, we also adopt support vector machines (SVMs) for comparison. We selected SVM since it is recommended as one of the most accurate predictors for bankruptcy prediction (Alaka et al., 2018). Moreover, SVM allows sample weighting so it can be better adapted to imbalanced data and provide competitive alternative for other investigated methods.

6.1.1. SVM and OCSVM

The OCSVM is built on famous Vapnik's idea of support vector machines Vapnik (1995). The starting assumption is that outliers occupy low-density region of the data space and kernel model can be used to characterize high density regions. The goal is to find function f that is able to identify points lying outside the region containing points from majority class. The strategy proposed in Schölkopf et al. (2001) is to map the data into the feature space corresponding to kernel and to separate them from origin with maximum margin. This can be achieved by solving quadratic programming task.

Let us first define training data as $x_1, x_2, \ldots, x_l \in X$ where $l \in \mathcal{N}$ is the number of observations. Additionally, let Φ be the map that maps X into inner product space F so the image of Φ is determined by evaluating the kernel $k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}))$. To separate data from the origin through the hyper-plane the quadratic program that needs to be solved is

$$\min_{w \in F, \xi \in \mathcal{R}, \rho \in \mathcal{R}} \frac{1}{2}\|w\|^2 + \frac{1}{\nu l}\sum_i \xi_i - \rho \tag{1}$$

$$s.t. (w \cdot \Phi(\mathbf{x_i})) \geq \rho - \xi_i, \xi_i \geq 0 \tag{2}$$

where $\nu \in (0, 1]$ characterizes the fraction of support vectors and outliers. The w and ρ are a weight vector and an offset parameterizing a hyper-plane in the feature space associated with the kernel. It can be shown Schölkopf et al. (2001) that results proven for binary classification through SVM Schölkopf et al. (2000) are also valid for single class classifier. Then, assuming $\rho \neq 0$ holds for equations (1) and (2), ν represents upper bound for fraction of outliers and lower bound on fraction of SVM. Moreover, if the data are separable and generated independently form distribution P not containing discrete components and the kernel is analytic and non-constant, ν equals to the fraction of outliers and fraction of SVs.

For w and ρ that solve quadratics programming problem in 1 is the decision function

$$f(\mathbf{x}) = sgn((w \cdot \Phi(x)) - \rho) \tag{3}$$

positive for the most examples x_i, while regularization term $\|w\|$ is still small. The tradeoff is controlled by variable ν. As long as Φ is implicit the above optimization problem can be solved by its dual form

$$\min_{a} \frac{1}{2} \sum_{i,j} a_i a_j k(x_i, x_j) \tag{4}$$

$$s.t. \qquad 0 \le a_i \le \frac{1}{\nu l} \sum_i a_i = 1 \tag{5}$$

Then ρ can be determined as

$$\rho = (w \cdot \Phi(x_i)) = \sum_j a_j k(x_i, x_j) \tag{6}$$

We improved the performance OF classifier by conducting a grid search over the grid (*degree*, γ, μ) defined by the product of the sets $degree = [1, 2, 3]$, $\gamma = [0.01, 0.1, 1, 5]$, $\mu = [0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9]$, where *degree* is the degree of the polynomial kernel of the SVM, μ is the upper bound on the fraction of training errors and a lower bound of the fraction of support vectors, and γ is the kernel coefficient.

6.1.2. LSAD

The idea of LSAD is based on assumption similar to OCSVM but use the different loss function that makes LSAD faster and easier to train at no cost in the prediction performance. LSAD is extended application of the least squares probabilistic classification Quinn and Sugiyama (2014); Sugiyama (2010). Assume class labels $y_i \in Y$ corresponding to observations X and let $y_i \in \{1, \ldots, c\}$ to be set of possible classes. Our aim is to estimate class conditional probabilities $p(y \mid x)$. To estimate $p(y = i \mid x)$ for each $i \in Y$ we can construct $q(y = i \mid \mathbf{x}, \theta_i = \theta_i^T \Psi(\mathbf{X}))$, where $\theta_i = (\theta_{i,1}, \ldots, \theta_{i,B})^T \in \mathcal{R}$ for B parameters. Considering case when classes $\{c + 1, c + 2, \ldots\}$ are represented only in test data but not in the training data, we need to assign value to estimate $\hat{p} = (y = * \ x)$ for some test data $\mathbf{x}$. The $y = *, * \in Y$ denote anomaly class. In this case conditional probability of an outlier can be estimated with

$$q(y = * \mid x, \theta_*) = 1 - \theta_*^T \Psi(X). \tag{7}$$

This is equal to searching for θ_*, such that 7 is close to zero when $\mathbf{x}$ lies inside the region containing points from majority class and zero otherwise. To achieve this we need to minimize loss function

$$l_*(\theta_*) = \frac{1}{2} \int (1 - \theta_*^T \Psi(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x} + \frac{\alpha}{2} \|\theta_*\|^2. \tag{8}$$

It can be shown that 8 is minimized by [13].

$$\hat{\theta}_* = (\Psi^T \Psi + \alpha \mathbf{I_B})^{-1} \sum_{j \in Y} \Psi^T \mathbf{m}_j = \sum_{j \in Y} \hat{\theta}_j \tag{9}$$

and therefore

$$q(y = * \mid \mathbf{x}, \hat{\theta}_1, \ldots, \hat{\theta}_S) = 1 - \sum_{j \in Y} q(y = * \mid \mathbf{x}, \hat{\theta}_j). \tag{10}$$

The parameter α is used to regularize and to increase the sensitivity to outliers.

In the case of LSAD we search through the parameters $\alpha = [0.01, 0.1, 1, 2, 3, 5, 10]$ and $\sigma = [0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 1, 2, 5, 10]$. Here, α controls the sensitivity to outliers and σ determines the smoothness of the boundary.

6.1.3. Isolation forest

The isolation forest is unsupervised non-parametric approach to anomaly detection (Liu et al., 2012). In contrast to the previous two methods IF differs in terms of the approach that is used to separate data. It does not employ any distance or density measure but rather attempt to isolate anomalies in data. The IF builds the ensemble of proper binary trees called isolation trees. The data samples with the short average length on the isolation tree are anomalies. The anomaly detection is two step process. First, the isolation trees are constructed from $\overline{X}$ by recursive partitioning. $\overline{X}$ is obtained by subsampling input data X by selection without replacement. In second, evaluation step, the anomaly score is estimated from the average path length $h(x)$ (Liu et al., 2012). The single path length $h(x)$ is obtained by counting the number of edges from the root node to a terminating node as sample x traverses through an isolation tree. So the anomaly score a of individual sample x is defined as

$$a(x, \tau) = 2^{-\frac{E(h(x))}{c(\tau)}}, \tag{11}$$

where $E(h(x))$ is the average value computed from all trees in ensemble and $c(\tau)$ is average path length of unsuccessful searches for set of τ instances (Liu et al., 2012). This is equivalent to unsuccessful search in binary search tree, therefore

$$c(\tau) = \begin{cases} 2H(\tau - 1) - 2(\tau - 1)/n & for \quad \tau > 2, \\ 1 \quad for \quad \tau = 2, \\ 0 \quad otherwise. \end{cases} \tag{12}$$

For the IF classifier we experimented with parameters $Nestimators = [100, 200, 300, 400, 500]$, $contamination = [0.02, 0.05, 0.1, 0.2, 0.3, 0.4]$ and $MaxSamples = [256, 512, 1024, 2048]$.

6.2. Empirical results

In the case of one-class classifiers, we validate the results by dividing the majority class into training data (80%) and test data (20%) and iterate the experiment 1,000 times with a random split for each loop. The final result is the average of all 1,000 loops. For the one-class classifier all minority data are holdout and used only for the testing phase. In the case of two-class SVM, all data, i.e., including the minority class, are divided into 80% training and 20% testing samples, after which the procedure is the same as for the one-class classifiers.

In the case of missing data, these are imported on a per feature basis to replace the missing value with the mean value of a particular feature. Afterwards, the data are scaled to have zero mean and unit variance.

As this is a severely imbalanced dataset, conventional accuracy measures cannot be used. Therefore, we decided to use the geometric mean (GM) score and Area Under the Receiver Operating Characteristic Curve (ROC AUC). Both measures take into account the accuracy of prediction on both classes, thereby preventing the result from being dominated by the accuracy of one class. The GM is the squared root of the product of the sensitivity and specificity and it is defined as

$$GM = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}}. \tag{13}$$

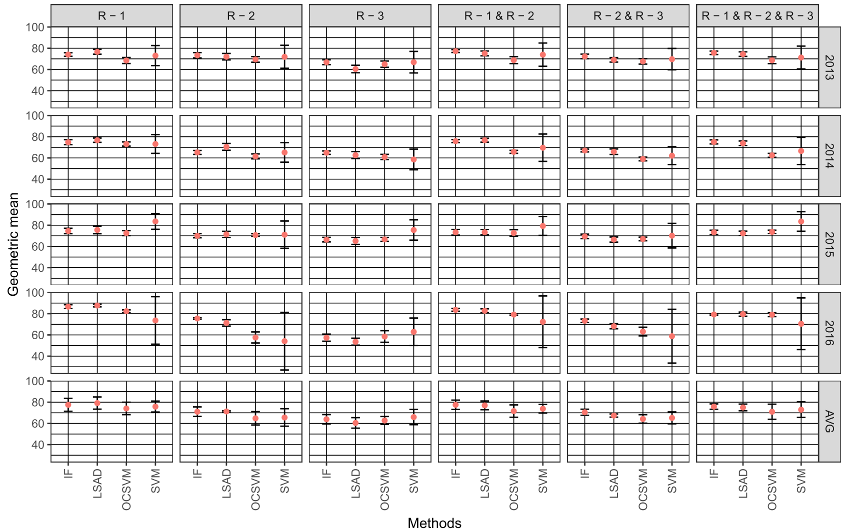


Fig. 3. GM score of methods IF, LSAD, OCSVM, SVM for evaluation years 2013–2016. The data are from one (R-1) to three (R-3) years prior to bankruptcy and their combinations, respectively. Industry dataset.

Here, TP and TN represent the number of true positives and true negatives, respectively. Similarly, FP denotes the number of false positives and FN the false negatives. Note that, in contrast to the accuracy score, the value of GM would be reduced to zero if the sensitivity score of one of the classes was equal to zero.

The ROC curve is determined by plotting TP rate against the FP rate at different threshold levels. The ROC AUC score is then computed as area under ROC.

One of our goals is to identify the most crucial time frame for bankruptcy prediction. Therefore, we evaluate the GM score separately based on data from one to three years prior to bankruptcy ($R-1$, $R-2$, $R-3$) and then on the combination of the data from multiple years, i.e., $R-1$ & $R-2$, $R-2$ & $R-3$, and $R-1$ & $R-2$ & $R-3$.

The GM scores for all four classification methods for both the industry and construction datasets are provided in Figs. 3 and 4, respectively. Similarly, the ROC AUC scores are presented in Figs. 5 and 6.

The highest prediction performance is achieved when only data from year $R-1$ are used for prediction. The GM and AUC ROC score for years $R-2$ and $R-3$ decrease recognizably, where the GM/ROC AUC scores for $R-3$ are as low as 50%–60%. This is consistent with the results presented in the previous section, where the features from year $R-1$ appeared to be the most significant. We can confirm our initial hypothesis that the data from one year prior to bankruptcy are the most indicative of upcoming financial problems. This observation is valid for both the manufacturing and construction data. Interestingly, the combination of multiple years does not improve the classification score. This can be explained by reasoning that either the classifiers were not able to take advantage of the data diversity from multiple years or, more probably, that the dominant information about bankruptcy is contained in data from year $R-1$ and that data from previous years do not contribute any new information for the predictor.

One-class classification proved to be an efficient approach for bankruptcy prediction on imbalanced data. Of the three one-class

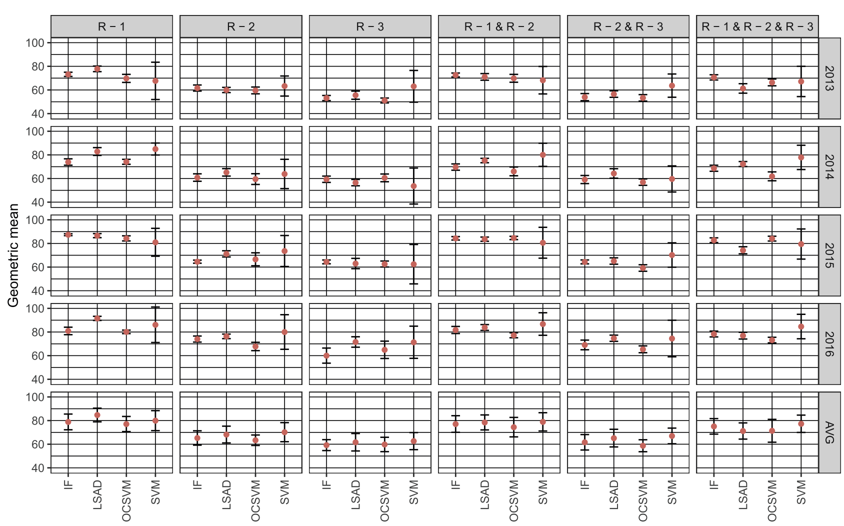


Fig. 4. GM score of methods IF, LSAD, OCSVM, SVM for evaluation years 2013–2016. The data are from one (R-1) to three (R-3) years prior to bankruptcy and their combinations, respectively. Constructions dataset.

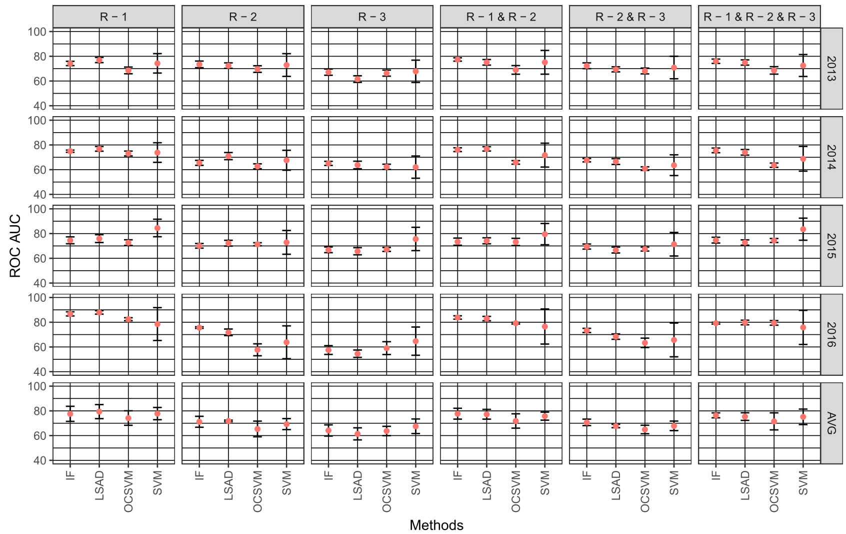


Fig. 5. ROC AUC score of methods IF, LSAD, OCSVM, SVM for evaluation years 2013–2016. The data are from one (R-1) to three (R-3) years prior to bankruptcy and their combinations, respectively. Industry dataset.

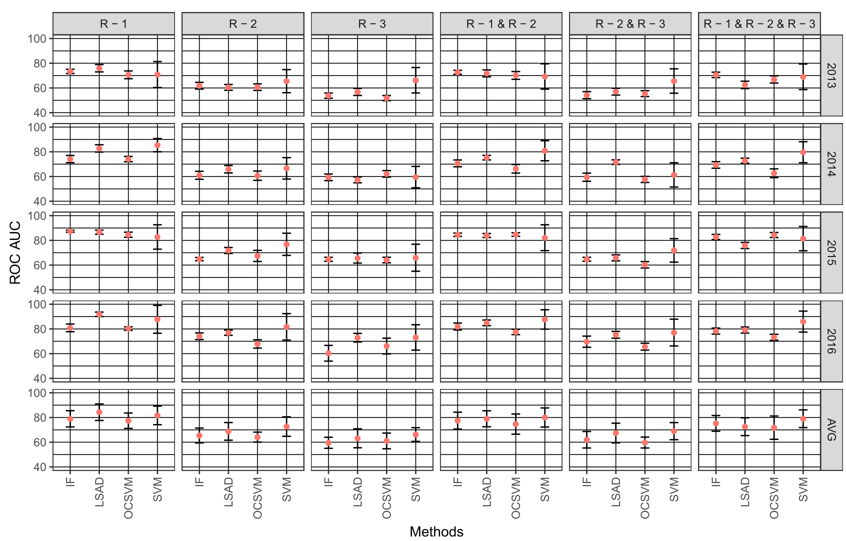


Fig. 6. ROC AUC score of methods IF, LSAD, OCSVM, SVM for evaluation years 2013–2016. The data are from one (R-1) to three (R-3) years prior to bankruptcy and their combinations, respectively. Constructions dataset.

methods used in this paper, LSAD obtained the highest prediction scores to outperform the other predictors in the majority of experiments. The highest prediction scores were also obtained by LSAD: $GM = 91,54\%$ ($ROCAUC = 91,83\%$) for construction data and $GM = 87,76\%$ ($ROCAUC = 87,92\%$) for manufacturing data. The SVM that was used as a baseline state-of-the art method displays quite competitive results in terms of average performance; however, as the standard deviation of the results is extremely high (more than 10% in some experiments), it cannot be recommended as a reliable approach according to this evaluation.

7. Conclusions

Knowledge of an upcoming bankruptcy is a crucial aspect of the decision-making process of the imperiled company itself as well as of other institutions interacting with the company. In this paper, we proposed a classification model to predict bankruptcy of small- and medium-sized companies, based on data from the annual report of the companies. These data are frequently available in publicly accessible databases or can be obtained through web scraping; thus, the proposed model can be built on these data and form the decision support system. We used a new dataset that reflects the authentic imbalanced distribution of bankrupt and non-bankrupt companies in two different areas of industry: construction and manufacturing. The proposed model based on one-class LSAD achieves a prediction score from 76% to 91%, depending on the evaluation year. Other than the classification model, we also conducted a detailed analysis of the financial parameters used for prediction. Knowing the most representative parameters provides an additional level of information to support the decision-making process since the responsible authorities can only focus on the relevant parameters.

The results for both industrial areas were very similar, thereby providing some indication that the model is applicable to other areas such as machinery and telecommunications. However, validation of the model for additional areas is a topic for further research.

There are several research directions for future work. We still need to investigate another approaches for prediction on imbalanced data such as cost sensitive learning and resampling strategies. We plan to employ also this approaches and then utilize the obtained knowledge for the development of new methods for classification on severely imbalanced datasets. Topic of imbalanced learning has been around for some time, so there are already several established methods. However, there is lack of feature selection methods for imbalanced data, so we believe that this is also topic that deserves more attention. Additionally, similarly to many other paper we focused on standard financial attributes and use these to build classification models. The introduction of new features can further improve the classification performance.

Acknowledgement

This work was supported by the Slovak Research and Development Agency under contract No. APVV-15-0358.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.econmod.2019.04.003.

References

Ahn, H., Kim, K.-J., 2009. Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. Appl. Soft Comput. 9, 599–607.
Akcay, S., Kundegorski, M.E., Willcocks, C.G., Breckon, T.P., 2018. Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. IEEE Trans. Inf. Forensics Secur. 13, 2203–2215, https://doi.org/10.1109/TIFS.2018.2812196.
Alaka, H.A., Oyedele, L.O., Owolabi, H.A., Kumar, V., Ajayi, S.O., Akinade, O.O., Bilal, M., 2018. Systematic review of bankruptcy prediction models: towards a framework for tool selection. Expert Syst. Appl. 94, 164–184.
Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Financ. 23, 589–609.
Ang, J.C., Mirzal, A., Haron, H., Hamed, H.N.A., 2016. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE ACM Trans. Comput. Biol. Bioinform 13, 971–989, https://doi.org/10.1109/TCBB.2015.2478454.
Azayite, F.Z., Achchab, S., 2016. Hybrid discriminant neural networks for bankruptcy prediction and risk scoring. Procedia Comput. Sci. 83, 670–674.
Beaver, W.H., 1966. Financial ratios as predictors of failure. J. Account. Res. 4, 71–111.
Bolon-Canedo, V., Sanchez-Marono, N., Alonso-Betanzos, A., 2015. Feature Selection for High Dimensional Data. Springer, London, GB.
Chand, P., Patel, A., White, M., 2015. Adopting international financial reporting standards for small and medium-sized enterprises. Aust. Account. Rev. 25, 139–154, https://doi.org/10.1111/auar.12067, https://onlinelibrary.wiley.com/doi/abs/10.1111/auar.12067, https://onlinelibrary.wiley.com/doi/pdf/10.1111/auar.12067.
Chaudhuri, A., De, K., 2011. Fuzzy support vector machine for bankruptcy prediction. Appl. Soft Comput. 11, 2472–2486.
Chen, Feng, Hope, Ole-Kristian, Li, Qingyuan, Wang, Xin, 2011. Financial reporting quality and investment efficiency of private firms in emerging markets. Account. Rev. 86, https://doi.org/10.2308/accr-10040.
Chuang, C.-L., 2013. Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction. Inf. Sci. 236, 174–185.
Commission of the European Communities, 2003. Commission Recommendation of 6 May 2003 (2003/361/ec) Concerning the Definition of Small and Medium-Sized Entreprises Official Journal No. L 124/36. http://data.europa.eu/eli/reco/2003/361/oj.
De Wit, G., De Kok, J., 2014. Do small businesses create more jobs? new evidence for europe. Small Bus. Econ. 42, 283–295.
Dimitras, A., Zanakis, S., Zopounidis, C., 1996. A survey of business failures with an emphasis on prediction methods and industrial applications. Eur. J. Oper. Res. 90, 487–513. https://doi.org/10.1016/0377-2217(95)00070-4.
Ding, Y., Song, X., Zen, Y., 2008. Forecasting financial condition of Chinese listed companies based on support vector machine. Expert Syst. Appl. 3081–3089.
Domingues, R., Filippone, M., Michiardi, P., Zouaoui, J., 2018. A comparative evaluation of outlier detection algorithms: experiments and analyses. Pattern Recogn. 74, 406–421. https://doi.org/10.1016/j.patcog.2017.09.037.
Drotar, P., Gazda, J., Smekal, Z., 2015. An experimental comparison of feature selection methods on two-class biomedical datasets. Comput. Biol. Med. 66, 1–10.
Fakhraei, S., Soltanian-Zadeh, H., Fotouhi, F., 2014. Bias and stability of single variable classifiers for feature ranking and selection. Expert Syst. Appl. 41, 6945–6958.
Firth, M., 1979. The impact of size, stock market listing, and auditors on voluntary disclosure in corporate annual reports. Account. Bus. Res. 9.
Garca, S., Luengo, J., Herrera, F., 2016. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowl. Base Syst. 98, 1–29. https://doi.org/10.1016/j.knosys.2015.12.006.
Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42.
Gordini, N., 2014. A genetic algorithm approach for smes bankruptcy prediction: empirical evidence from Italy. Expert Syst. Appl. 41, 6433–6445. https://doi.org/10.1016/j.eswa.2014.04.026.
Gupta, Jairaj, Gregoriou, Andros, Healy, Jerome, 2015. Forecasting bankruptcy for SMEs using hazard function: To what extent does size matter? Rev. Quant. Financ. Acc. 45, 845–869. https://doi.org/10.1007/s11156-014-0458-0 ISSN 1573-7179.
Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. J. Mach. Learn. Res. 46, 389–422.
Haibo He, Garcia, E., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21.
Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: review of methods and applications. Expert Syst. Appl. 73, 220–239.
He, X., Cai, D., Niyogi, P., 2005. Laplacian score for feature selection. In: Proceedings of the 18th International Conference on Neural Information Processing Systems NIPS'05. MIT Press, Cambridge, MA, USA, pp. 507–514.
Hinton, G., Roweis, S., 2002. Stochastic neighbor embedding. Adv. Neural Inf. Process. Syst. 15, 833–840.
Iturriaga, F.J.L., Sanz, I.P., 2015. Bankruptcy visualization and prediction using neural networks: a study of us commercial banks. Expert Syst. Appl. 42, 2857–2869.
Jo, H., Han, I., Lee, H., 1997. Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. Expert Syst. Appl. 13, 97–108.
Kalousis, A., Prados, J., Hilario, M., 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. 12, 95–116.
Kim, M.-J., Kang, D.-K., Kim, H.B., 2015. Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. Expert Syst. Appl. 42, 1074–1082.
Kim, T.-Y., Cho, S.-B., 2018. Web traffic anomaly detection using c-lstm neural networks. Expert Syst. Appl. 106, 66–76. https://doi.org/10.1016/j.eswa.2018.04.004.
Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Prog. Artif. Intell. 5, 221–232. https://doi.org/10.1007/s13748-016-0094-0.
Kumar, P.R., Ravi, V., 2007. Bankruptcy prediction in banks and firms via statistical and intelligent techniques - a review. Eur. J. Oper. Res. 180, 1–28.
Lee, S., Choi, W.S., 2013. A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. Expert Syst. Appl. 40, 2941–2946.
Lee, T., Chiu, C., Chou, Y., Lu, C., 2006. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Comput. Stat. Data Anal. 50, 1113–1130.
Li, H., Sun, J., 2008. Ranking-order case-based reasoning for financial distress prediction. Knowl. Base Syst. 21, 868–878.
Li, H., Sun, J., 2009. Predicting business failure using multiple case-based reasoning combined with support vector machine. Expert Syst. Appl. 36, 10085–10096.
Li, H., Sun, J., 2012. Forecasting business failure: the use of nearest-neighbour support vectors and correcting imbalanced samples - evidence from the Chinese hotel industry. Tourism Manag. 33, 622–634.
Li, X., Wang, F., Chen, X., 2015. Support vector machine ensemble based on Choquet integral for financial distress prediction. Int. J. Pattern Recognit. Artif. Intell. 29.
Liao, J.-J., Shih, C.-H., Chen, T.-F., Hsu, M.-F., 2014. An ensemble-based model for two-class imbalanced financial problem. Econ. Modell. 37, 175–183. https://doi.org/10.1016/j.econmod.2013.11.013.
Lin, W.Y., Hu, Y.H., Tsai, C.F., 2012. Machine learning in financial crisis prediction: a survey. IEEE Trans. Syst., Man, Cybern., Part C (Appl. Rev.) 42, 421–436.
Liu, F.T., Ting, K.M., Zhou, Z.-H., 2012. Isolation-based anomaly detection. ACM Trans. Knowl. Discov. Data 6, 3:1–3:39.
Maaten, L. v. d., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.
Min, J., Lee, Y., 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Syst. Appl. 28, 603–614.
Min, S.-H., Lee, J., Han, I., 2006. Hybrid genetic algorithms and support vector machines for bankruptcy prediction. Expert Syst. Appl. 31, 652–660.
Modigliani, F., Miller, M.H., 1958. The cost of capital, corporation finance and the theory of investment. Am. Econ. Rev. 48, 261–297.
Nwogugu, M., 2007. Decision-making, risk and corporate governance: a critique of methodological issues in bankruptcy/recovery prediction models. Appl. Math. Comput. 185, 178–196.
Olden, J.D., Jackson, D.A., 2002. Illuminating the black box a randomization approach for understanding variable contributions in artificial neural networks. Ecol. Model. 154, 135–150.
Quinn, J.A., Sugiyama, M., 2014. A least-squares approach to anomaly detection in static and sequential data. Pattern Recogn. Lett. 40, 36–40.
Raskutti, B., Kowalczyk, A., 2004. Extreme re-balancing for svms: a case study. SIGKDD Explor. Newsl. 6, 60–69. http://doi.acm.org/10.1145/1007730.1007739.
Robnik-Šikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. 53, 23–69.

Sartori, F., Mazzucchelli, A., Gregorio, A.D., 2016. Bankruptcy forecasting using case-based reasoning: the creperie approach. Expert Syst. Appl. 64, 400–411. https://doi.org/10.1016/j.eswa.2016.07.033.
Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural Comput. 13, 1443–1471.
Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L., 2000. New support vector algorithms. Neural Comput. 12, 1207–1245. https://doi.org/10.1162/089976600300015565.
Shin, K.-S., Lee, T.S., Kim, H.-j., 2005. An application of support vector machines in bankruptcy prediction model. Expert Syst. Appl. 28, 127–135.
Sugiyama, M., 2010. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. IEICE Trans. Info Syst. E93.D, 2690–2701, https://doi.org/10.1587/transinf.E93.D.2690.
Sun, J., Lang, J., Fujita, H., Li, H., 2018. Imbalanced enterprise credit evaluation with DTE-SBD: decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Inf. Sci. 425, 76–91.
Sun, J., Li, H., Huang, Q.-H., He, K.-Y., 2014. Predicting financial distress and corporate failure: a review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. Knowl. Base Syst. 57, 41–56.
Sung, T.K., Chang, N., Lee, G., 1999. Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction. J. Manag. Inf. Syst. 16, 63–85.
Tamari, M., 1966. Financial ratios as a means of forecasting bankruptcy. Manag. Int. Rev. 6, 15–21.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, Berlin, Heidelberg.
Veganzones, D., Severin, E., 2018. An investigation of bankruptcy prediction in imbalanced datasets. Decis. Support Syst. 112, 111–124. https://doi.org/10.1016/j.dss.2018.06.011.
Wang, G., Ma, J., 2012. A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine. Expert Syst. Appl. 39, 5325–5331.
Wang, M., Chen, H., Li, H., Cai, Z., Zhao, X., Tong, C., Li, J., Xu, X., 2017. Grey wolf optimization evolving kernel extreme learning machine: application to bankruptcy prediction. Eng. Appl. Artif. Intell. 63, 54–68. https://doi.org/10.1016/j.engappai.2017.05.003.
Yang, Z., Platt, M., Platt, H., 1999. Probabilistic neural networks in bankruptcy prediction. J. Bus. Res. 44, 67–74.
Yeh, I.-C., Lien, C.-h., 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Syst. Appl. 36, 2473–2480.
Zhou, L., 2013. Performance of corporate bankruptcy prediction models on imbalanced dataset: the effect of sampling methods. Knowl. Base Syst. 41, 16–25.
Zhu, Y., Xie, C., Wang, G.-J., Yan, X.-G., 2017. Comparison of individual, ensemble and integrated ensemble machine learning methods to predict chinas sme credit risk in supply chain finance. Neural Comput. Appl. 28, 41–50.