“Feature selection methods and sampling techniques to financial distress prediction for Vietnamese listed companies”

AUTHORS	Loan Thi Vu Lien Thi Vu Nga Thu Nguyen Phuong Thi Thuy Do Dong Phuong Dao

ARTICLE INFO	Loan Thi Vu, Lien Thi Vu, Nga Thu Nguyen, Phuong Thi Thuy Do and Dong Phuong Dao (2019). Feature selection methods and sampling techniques to financial distress prediction for Vietnamese listed companies. *Investment Management and Financial Innovations*, *16*(1), 276-290. doi:10.21511/imfi.16(1).2019.22
DOI	http://dx.doi.org/10.21511/imfi.16(1).2019.22
RELEASED ON	Monday, 25 March 2019
RECEIVED ON	Friday, 21 December 2018
ACCEPTED ON	Tuesday, 12 March 2019

JOURNAL	"Investment Management and Financial Innovations"
ISSN PRINT	1810-4967
ISSN ONLINE	1812-9358
PUBLISHER	LLC “Consulting Publishing Company “Business Perspectives”
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
40	0	16

BUSINESS PERSPECTIVES

LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10, Sumy, 40022, Ukraine

www.businessperspectives.org

Loan Thi Vu (Vietnam), **Lien Thi Vu** (Vietnam), **Nga Thu Nguyen** (Vietnam), **Phuong Thi Thuy Do** (Vietnam), **Dong Phuong Dao** (Vietnam)

FEATURE SELECTION METHODS AND SAMPLING TECHNIQUES TO FINANCIAL DISTRESS PREDICTION FOR VIETNAMESE LISTED COMPANIES

Abstract

The research is taken to integrate the effects of variable selection approaches, as well as sampling techniques, to the performance of a model to predict the financial distress for companies whose stocks are traded on securities exchanges of Vietnam. A firm is financially distressed when its stocks are delisted as requirement from Vietnam Stock Exchange because of making a loss in 3 consecutive years or having accumulated a loss greater than the company's equity. There are 12 models, constructed differently in feature selection methods, sampling techniques, and classifiers. The feature selection methods are factor analysis and F-score selection, while 3 sets of data samples are chosen by choice-based method with different percentages of financially distressed firms. In terms of classifying technique, logistic regression together with SVM are used in these models. Data are collected from listed firms in Vietnam from 2009 to 2017 for 1, 2 and 3 years before the announcement of their delisting requirement. The experiment's results highlight the outperformance of the SVM model with F-score selection method in a data sample containing the highest percentage of non-financially distressed firms.

Keywords financial distress prediction, feature selection, sampling technique, logistic regression model, Support Vector Machine (SVM)

JEL Classification G32, G33, G38

Received on: 21st of December, 2018
Accepted on: 12th of March, 2019

Loan Thi Vu, Ph.D., Faculty of Banking and Finance, VNU University of Economics and Business, Vietnam.

Lien Thi Vu, Ph.D., Faculty of International Training, Thainguyen University of Technology, Vietnam.

Nga Thu Nguyen, Ph.D., Faculty of Finance and Banking, Thainguyen University of Economics and Business Administration, Vietnam.

Phuong Thi Thuy Do, Ph.D., Faculty of Accounting, Thainguyen University of Economics and Business Administration, Vietnam.

Dong Phuong Dao, MBA, Faculty of Banking and Finance, VNU University of Economics and Business, Vietnam.

INTRODUCTION

According to Beaver (1966) in the first study on financial distress prediction, a firm is considered financially distressed or failed if the company fails to fulfill its financial obligations when mature. Since Beaver's pioneering work, the construction of a warning model has become the center of research in corporate finance worldwide. Traditionally, a financially distressed firm is a company that falls into bankruptcy because of business failure (Beaver, 1966; Altman, 1968; Norton & Smith, 1980; Ohlson, 1980) and remains popular in more relevant and recent research (Zhou et al., 2012; Altman et al., 2016; Liang et al., 2015). Another financial distress measure is known as finance-based definition (Pindado et al., 2008). In this measure, the financial distress of a company may not necessarily put it into bankruptcy (Altman, 1984), thus, a model for financial distress forecasting plays a crucial role in helping firms to avoid bankruptcy (Santoso, 2018).

It is clear that financial distress regardless of its recognition produces huge potential loss to the stakeholders of a company. Therefore, a financial prediction model works as an early warning system that supports the companies' managers to make necessary adjustments in

http://dx.doi.org/10.21511/imfi.16(1).2019.22

their financial management strategies to avoid becoming distressed. It also assists the investors and creditors in their decision-making process and helps the government to provide an alarm to the firms before putting them on the "control" list. Although there are numerous models that have been created and tested, it has been revealed that the performance of a financial distress prediction model varies if different sets of predictors, data samples and classifiers are applied.

The independent variables in a prediction model are mainly accounting ratios and they have been extended to other features outside the accounting reports. In order to obtain the optimal variables for the model, different feature selection methods have been applied to choose the most informative and discriminant ratios. In addition to selection method, there is also evidence that the choice of data sampling affects the prediction model's performance. According to the number of financially distressed companies chosen, the sampling techniques can be divided into either the choice-based sampling or the sampling technique named complete data. Recently, there has been an increasing number of papers that have attempted to draw a comparison between these two selection approaches, but no consistent conclusion can be discerned.

The third determinant of a model's performance is the choice of classifying technique applied to determine whether a firm is financially distressed or not. Supporting by the computer science, the classifiers have been developed from the Univariate model (Beaver, 1966) to the Discriminant Technique model (Altman, 1968), and further to the Logistic Regression model (Ohlson, 1980) and Data Mining models. Although the comparison between different models has been taken widely, there is no consistent answer for the best classifier, which presents its superiority in all data samples.

Based on the factors that influence the model's performance, most of the relevant research focuses on improving the model's accuracy by selecting the optimal set of predictors and an appropriate classifying technique for a particular data sample. However, it can be seen that the combined effects of feature selection methods and sampling techniques have not yet received enough interest from researchers. Therefore, this research aims to build models to predict the financial distress condition of listed firms on securities exchange in Vietnam that focuses on the role of the feature selection method in association with different sampling choices in improving the model's performance. The importance of the study is emphasized as the number of firms financially distressed is increasing in Vietnam, while the number of research projects occurring is limited.

Data are collected from companies listed in the Vietnamese securities market from 2009 to 2017, while a financially distressed company is the one receiving the requirement of being delisted. The analysis results reveal that the model's accuracy is higher as the number of non-financially distressed firms chosen increases. Overall, the SVM models with F-score feature selection outperform the Logistic Regression models.

1. LITERATURE REVIEW ON FINANCIAL DISTRESS PREDICTION MODEL

1.1. Review of predictors and predictor selection

In a financial distress prediction model, the choice of predictors can affect the accuracy level of prediction. While the usefulness of each predictor varies in different models, independent variables in existing studies can be classified into three main groups: accounting ratios, market variables and macroeconomic ratios. Among these groups, the largest one is the accounting ratio group. This includes ratios calculated from companies' financial statements, which are prepared by the companies according to pre-determined accounting principles. It can be perceived that the accounting ratios reflecting companies' financial performance such as liquidity, profitability, business ca-

pacity, capital structure etc. of the firm are favored by researchers. In addition to accounting ratios, market variables are also used in an ex-ante model as they contain information on expected future cash flows, which are relevant to the likelihood of being financial distressed (Rees, 1995; Back et al., 1996; Beaver et al., 2005). In existing papers, the market variables that have discriminate power can be listed as the company's stock price (Beaver et al., 2005; Christidis & Gregory, 2010; Tinoco et al., 2013), the lagged cumulative stock residual return (Tinoco et al., 2013), and the company's market capitalization (Beaver et al., 2005; Tinoco et al., 2013). The third main source of predictors consists of macroeconomic ratios as the number of default firms increases during economic downturns (Koopman & Lucas, 2005). The proxies for macroeconomic ratios can be: real interest rates; business cycles (Bhattacharjee & Han, 2014), Gross Domestic Product (GDP), money supply and Consumer Price Index – CPI (Alifiah, 2014). Apart from three main groups of predictors, previous research also tests the roles of corporate governance indicators (Liang et al., 2015) and industry specifications (Sayari & Mugan, 2016) such as board structure, ownership structure in the prediction models.

A literature review shows that there is a great number of predictors that can be utilized in a financial distress prediction model. According to Zhou et al. (2012), there are 500 different variables that can be found in 128 papers and the predictive power of each variable changes in different papers (Sayari & Mugan, 2016). As stated by Powell (2007), the high dimensionality problem can be raised if too many variables are used for data analysis. Therefore, reducing the number of total variables by retaining only informative and discriminative predictors is crucial to improve a prediction model's performance.

Feature selection, defined as the approach for selecting the optimal set of predictors, has been applied broadly in existing papers. It is also designed to produce better performance, reduce the cost of processing a model, as well as to obtain better understanding of the company's operation (Guyon & Elisseeff, 2003). In previous articles, variable selection techniques are recognized as expert recommendation and statistical methods (Lin et al., 2014). As examples of using the expert recommendation method, studies taken by Alifiah (2014) and Liang et al. (2015) select variables, which are useful in at least ten previous papers or factors appearing more than 3 times in 127 relevant models for the model's indicators. On the other hand, filter-based feature and the wrapper-based feature selection method are categorized into the statistical methods for variable selection. These methods are considered to be computationally efficient when they apply into a large number of independent variables (Blum & Langley, 1997; Guyon & Elisseeff, 2003).

The filter-based selection method includes the t-test, factor analysis, and stepwise regression, which assess the relevance of the variables according to pre-determined indices. The proposed criteria for this method can be Fisher score (Yjlmaz, 2013), Laplacian score (Wan et al., 2015) and F-score (Chen & Lin, 2003). Among those criteria, F-score is considered to be the simplest (Song et al., 2017). In contrast, a wrapper method evaluates the variables based on their usefulness through a process that requires a lot of data processing. According to Kittler (1978), wrapper techniques can be listed as sequential forward selection or sequential backward selection. In other papers of Kohavi and John (1997) and Goldberg (1989), it can be recognized as randomized hill climbing or genetic algorithms.

1.2. Review of sampling technique and classifiers

In addition to the discussion of feature selection methods, there is also a disagreement on the data selection for the model. Zmijewski (1984) was the first researcher to discuss two data selection techniques in building a financial distress prediction model: the choice-based sampling technique and the complete data sampling technique. The choice-based sampling technique or stratified random sampling is used when the available distressed companies and only a part of the non-financially distressed companies are kept in the sample. The non-distressed firms are chosen randomly or by some criteria such as industry or company size. This sampling technique has been applied widely by Beaver (1966), Altman (1968), Liang et al. (2015), Mine and Hakan (2006), Geng et al. (2014), Lin et

http://dx.doi.org/10.21511/imfi.16(1).2019.22

al. (2014), Shaonan et al. (2015) and Mselmi et al. (2017). According to Shaonan et al. (2015), "choice-based technique successfully remedies the potential problem of extremely low frequency rate of bankruptcy events in the population". Opponents of this method state that the significant difference in financial distress contribution in the sample in comparison with that in the population may lead to biased estimation of parameters (Zmijewski, 1984; Shaonan et al., 2015). In contrast to the previous approach, the latter technique brings all available non-financially distressed firms to the data sample. For example, Ohlson (1980) brings entire records of the 2,050 non-distressed companies and 105 failed companies to the data set. A similar approach has been applied in the works of Bharath and Shumway (2008), and Kim and Sohn (2010). Supporters of complete data sampling technique argue that the rate of financially distressed companies should be representative of the population in a sample (Ohlson, 1980) and the biased parameters can be decreased as the likelihood of distressed firms in the sample approaches that of the population (Zmijewski, 1984). However, because of the great number of non-distressed firms compared to the number of distressed firms in the sample, this technique requires a huge amount of computation that may lead to a class imbalance problem and degradation in the final prediction performance (Liang et al., 2015).

Unquestionably, the classifier which is used to discriminate a company in the data sample according to their selected predictors plays a significant role in increasing the level of accuracy. With the development of statistical and soft computing techniques, a significant number of financial distress prediction models with various classifiers have been constructed and many of them can obtain impressive levels of accuracy. Zhou et al. (2012) summarize the related empirical researches and divide these techniques into 2 groups: traditional classifiers and modern classifiers.

Beaver (1966), Altman (1968) and Ohlson (1980) are the authors who construct financial distress prediction models with traditional classifiers. Beaver (1966) was the pioneer in presenting the Univariate model, while Altman (1968) introduced the Multiple Discriminant Analysis (MDA) model that identifies a discriminant function known as the Z-score model. Until the 1980s, the MDA model was the dominant model regarding research on financial distress prediction (Balcaen & Ooghe, 2006). Beyond this, however, the domination of MDA model decreased due to the introduction of the Logistic Regression model by Ohlson (1980). This model has overtaken the MDA as the dominant model as it does not require any assumptions of normal distribution and equal covariance which are considered drawbacks of the MDA model. In addition to traditional models, the development of Artificial Intelligence (AI) and Data Mining has created modern classifiers such as Decision Tree (DT), Neural Network (NN), and Support Vector Machines (SVM).

There have been a number of studies on performance comparison between models with different classifiers. Ugurlu (2006) discovered that the Logit model provided a better accuracy level and overall fit than the MDA model. The same conclusion was also made in the study of Pindado et al. (2008). Recent studies taken by Lin et al. (2011, 2014) assert that the SVM model outperforms not only traditional models, but also other data mining models. Another paper produced by Gepp and Kumar (2015) concluded that the DT model is a superior classifier compared with the Logistic Regression model.

2. AN OVERVIEW OF VIETNAMESE SECURITIES MARKET

Vietnamese securities market was established by the opening of Hochiminh stock exchange on July 20, 2000 before the launching of Hanoi Stock Exchange on March 8, 2005. After over 20 years of operation, Vietnamese securities market has played an extremely important role in capital raising for Vietnamese enterprises, as well as fostering the capitalization process of state owned companies. Beginning with only 2 listed companies with the market capitalization of nearly USD 49.3 billion, in 2016, that the number of listed companies is 678 with the market capitalization contributing 33% to GDP, 114 times higher than that in 2000. The securities market is now well organized with the existence of stocks, bonds and derivatives. At the end of 2018, there are 1,558 listed companies in the market with the mar-

ket capitalization amounting to 82.2% of country's GDP. Vietnamese securities market has shown the attraction to foreign investors as foreign investors number increases by 47.4% in 2017. Vietnames government has set the target to transfer the securities market from frontier to emerging market in 2019.

From 2009, the State Securities Commission of Vietnam started to require a company to be delisted because of its financial distress. Specifically, a company is delisted if it incurs losses in 3 consecutive years or having accumulated loss bigger than its equity. The number of delisted companies increases from 6 companies in 2010 to the peak of 31 companies in 2013 and slightly decreases to 27 companies in 2017. Although the delisting requirement can improve the quality of listing stocks, an increasing number of delisted company may affect the market belief from investors. Therefore, investors in Vietnam should be supported by a financial distress prediction model for stock selection, while a company's managers also need this model to make necessary adjustments that can help company to avoid being financially distressed.

2.1. Research design

The main objective of the research is to construct the financial distress prediction models that take into account the combined effects of feature selection and sampling methods for companies listed in Vietnamese securities market. There are two main steps conducted to fulfill the research objective. In the first step, a number of models with different sets of predictors, data and classifiers are designed. In the next step, a comparison is taken to find the most effective model. The recognition of a firm's financial distress follows the finance-based definition, which emphasizes the independence of financial distress and its outcomes. A company is considered to be financially distressed when it is required to be delisted in Hanoi's securities market or Ho Chi Minh's securities market as it suffers losses in 3 consecutive years or its accumulated loss rises above the company's equity.

2.2. Feature selection methods

In order to highlight the analysis results, feature selection procedure is applied into 2 different sets of variables (variable set 1 and variable set 2) chosen from empirical analysis. The variable set 1 (see Table A1 in Appendix A) is taken mainly from the research of Geng et al. (2014) with an additional variable, which measures the ownership structure of the company in Vietnam. Those predictors cover different features such as solvency, profitability, operational capacity of an enterprise's financial performance. The variable set 2 (see Table A2 in Appendix A) is originated from the paper of Lin et al. (2014), because it is the result of a comprehensive selection method that integrates expert recommendations and wrapper approach.

In this paper, factor analysis and the F-score method are applied to minimize the number of predictors in order to increase the level of accuracy of the model. Factor analysis is performed to explore the "variables that seem(s) to be doing the best job in predicting financial distress" or are the most informative in the model with the application of VARIMAX for rotation. The factor analysis is also applied to detect the multicollinearity among the predictors. The selection procedure is based on a number of criteria. Firstly, Bartlett's test of sphericity should be significant to ensure the appropriateness of independent indicators for factor analysis. Second, the most informative variables should have factor loading above 0.5, the eigenvalue bigger than 1 and the communality greater than 0.8. The results of factor analysis are presented in Table 3 for models 1.1, 1.2 and 1.3. After conducting factor analysis, stepwise regression combined with binary logistic regression will provide the significant ratios that can act as predictors in the model.

F-score is a simple filter selection method that can be used together with any of the SVM models. F-score measures the discrimination of two variables set according to below function:

$$F_{(i)} = \frac{\left(\overline{x}_i^{(+)} - \overline{x}_i\right)^2 + \left(\overline{x}_i^{(-)} - \overline{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n+}\left(x_{k,i}^{(+)} - \overline{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n-}\left(x_{k,i}^{(-)} - \overline{x}_i^{(-)}\right)^2}, \quad (1)$$

where $\overline{x}_i$ – average of feature *i* of the whole feature, $\overline{x}_i^{(+)}$ – average of positive feature, $\overline{x}_i^{(-)}$ – average of positive feature.

Features with higher F-score should be chosen in the model because of having higher discrimination ability. First introduced by Chen and Lin

http://dx.doi.org/10.21511/imfi.16(1).2019.22

(2003), there are 2 steps in this feature selection. In the first step, the F-score of every feature is calculated before setting a threshold to remove the feature with *F*-score below that of the threshold and retain those with an *F*-score above it. In the next step, the data are again split randomly into new training data and testing data.

2.3. Sampling techniques

The researchers use the choice-based sampling technique to choose the firms used in the sample. In this method, the non-distressed firms are chosen randomly with their number equal to that of the number of distressed firms. However, because of the concern about the biased parameters produced from the inconsistent distress rates in the sample and population, three data samples were created with increasing numbers of non-financially distressed firms. Data sample 1 consists of 68 distressed firms and 68 non-distressed firms, data sample 2 includes 68 distressed firm and 136 non-distressed firms, while the number of non-distressed firms (204) is triple the number of distressed firms in data sample 3. The increase in the number of non-distressed firms, as well as the data size from data sets 1 to 3, reduces the inconsistency between the distress rates of the sample and that of the population. By choosing these data samples, the authors expect to discover the relationship between the rate of a distressed firm in the sample and the prediction model's accuracy.

2.4. Classification techniques

A classification technique is used to train the data for constructing the classifying function using selected independent variables. The function then applies to testing data set for determining the model's accuracy. This study uses logistic regression and the machine learning algorithm SVM as classifiers.

The logistic regression tries to compute the likelihood of being "financially distressed" for a listed firm. In the below function, the dependent variable *Y* receives value of 1 or 0. The former describes the company's financial distress, while the latter denotes the condition of non-financial distress:

$$P(Y=1)=\frac{e^{-(\propto+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}}{1+e^{-(\propto+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}}. \quad (2)$$

With the support of computer software such as SPSS, a company is considered to be financially distressed if $P(Y)>0$ or non-distressed if $P(Y)<0$. The logistic regression model is chosen as it is a traditional classifier that exhibits high level of prediction accuracy in recent studies of Ugurlu (2006) and Pindado et al. (2008). In addition, the logistic regression model does not require the assumption of normality, as well as equal covariance of independent variables.

SVM which is known as a type of machine learning classifier establishes a hyperplane that separates two groups of companies according to their financial performance. Especially, this classifier identifies the optimized hyperplane with largest margin for companies' separation.

$$\begin{aligned}&\min\frac{1}{2}w^Tw+C\sum_{i=1}^{n}\xi_i,\\&subject\ to\ y_i\left(w^T\varnothing\left(x_i\right)+b\right)\geq1-\xi_i\\&\xi_i\geq0\end{aligned} \quad (3)$$

According to Hsu et al. (2016), in order to construct the optimized hyperplane, the parameters *C* and γ should be determined by grid search. In this study, the radial basis function (RBF) $K\left(x_i,x_j\right)=\exp\left(-\gamma x_i-x_j^{\ 2}\right),\gamma>0$ incorporating C and γ is established. As the two parameters are selected, data training should be performed again. With the support of LibSVM tool, before conducting the classification techniques, the data are separated into 2 sets for training and predicting. SVM is a machine learning technique, which should be applied in comparison with logistic model, a traditional classifier. Stated by Sánchez et al. (2016), SVM model works well with small sample size and it also outperforms other data mining classifiers (Liang et al., 2015; Lin et al., 2011, 2014).

There are 12 models constructed with combinations of the different data sampling techniques, feature selection and classification methods. According to Table 1, 6 models from 1.1 to 1.6 apply factor analysis into 3 data sets with different sample size and use logistic regression as a classification technique. Meanwhile, 6 models from 2.1 to 2.6 determine *F*-score as a feature selection method to data sets 1, 2 and 3 and use SVM as classifier.

Table 1. Description of the models

Feature selection	Data set			Variable set	Classification technique
	Data set 1	Data set 2	Data set 3		
Factor analysis	Model 1.1	Model 1.2	Model 1.3	Set 1	Logistic regression
	Model 1.4	Model 1.5	Model 1.6	Set 2	
F-score	Model 2.1	Model 2.2	Model 2.3	Set 1	SVM
	Model 2.4	Model 2.5	Model 2.6	Set 2	

3. DISCUSSION OF ANALYSIS RESULTS

3.1. Feature selection results

3.1.1. Factor analysis

The factor analysis is applied for 3 sets of data samples. In the first step, Bartlett's test and Kaiser-Meyer-Olkin (KMO) are run in order to assess the overall significance of the correlation matrix and the sample adequacy. If the value of KMO is under the range from 0.5 to 1, the factor analysis is considered to be appropriate for the data. In addition, Bartlett's test is important to make sure of the significant correlation between variables. In the next step, the VARIMAX is used for rotation to select the factors according to their eigenvalue, factor loading and communality. A suitable factor must have eigenvalue bigger than 1 along with factor loading and communality greater than 0.5 and 0.8, respectively. As shown in Table A3 and Table A4 in Appendix A, according to those criteria, the number of selected variables decreases quite dramatically in each variable set.

After conducting factor analysis, there is no multicollinearity found among the predictors and the redundant variables are also removed. In the next step, the significance test of all informative variables is performed to ensure the validity and the significance of the prediction model by running stepwise regression for logistic regression models. The results of stepwise regression procedures show that the independent variables are the same for all different models in one prediction time. However, the coefficient of each predictor varies in each different data sample. Tables 2 and 3 describe the coefficient of each significant predictor of 6 models for 1 year, 2 years and 3 years before the distress event.

According to Table 2, models constructed 1 year before the financial distress event emphasize the importance of current liabilities on total assets ratio, net profit over average current assets ratio and net profit on average fixed assets. Surprisingly, the negative sign in the coefficient of total liabilities on total assets ratio gives rise to the concern about the parameters produced by model 1.1 as there should be a positive relationship between this ratio and the financial distress probability.

Table 2. Stepwise selection results for logistic regression models – variable set 1

	Variable	Coefficient		
		Model 1.1	Model 1.2	Model 1.3
	Year 1			
1	*TL/TA*	–0.98	0.204	1.981
2	*CA/CL*	–0.46	–0.718	–0.845
3	*(CA–I)/CL*	–0.379	–0.321	–0.224
4	*CL/TA*	3.146	2.255	2.166
5	*NP/ACA*	–1.66	–3.677	–2.778
6	*NP/AFA*	–1.498	–1.818	–1.592
7	*CS/APA*	–0.947	–0.643	–0.882
8	*NP/NOS*	–0.059	–0.131	–0.256
9	*NA/NOS*	–0.042	–0.039	–0.054
	Year 2			
1	*TL/TA*	–1.176	1.118	3.386
2	*CA/CL*	–0.245	–0.655	–0.707
3	*(CA–I)/CL*	1.754	0.039	0.089
4	*TL/TSE*	0.058	0.062	0.04
5	*CL/TA*	6.052	3.823	2.462
6	*NP/SR*	–0.626	–1.216	–0.967
7	*NP/ACA*	–2.263	–5.747	–1.326
8	*NP/AFA*	–0.964	–1.463	–1.139
9	*CS/APA*	–0.963	–0.732	–0.781
10	*NP/NOS*	0.209	–0.14	–0.356
11	*NA/NOS*	0.046	0.043	0.056
	Year 3			
1	*TL/TA*	7.724	3.938	8.34
2	*CA/CL*	–2.551	–0.143	–0.149
3	*(CA–I)/CL*	2.458	0.355	0.045
4	*TL/TSE*	1.414	0.554	1.204
5	*CL/TA*	–1.738	3.076	4.244
6	*EBIT/IE*	–0.067	–0.017	–0.007
7	*NP/SR*	–1.174	–0.261	–0.163
8	*EBIT/ATA*	–23.436	–5.844	–1.979
9	*NP/AFA*	–0.584	–4.105	–11.289
10	*CS/APA*	–1.019	–0.518	–0.707
11	*CA/TA*	–0.739	–1.786	–5.498

http://dx.doi.org/10.21511/imfi.16(1).2019.22

The number of significant and discriminative independent variables increases to 11 variables in 2-year prediction models from 9 variables in 1-year prediction models, because total liabilities over total shareholders' equity ratio and net profit over sale revenue ratios are included. There are consistencies in the signs of variables' coefficients that can be found among models. The negative coefficient sign of total liabilities/total assets ratio and positive coefficient sign of net assets/number of ordinary shares at the end of year ratio threaten the application of model 1.1. The 3-year model has the same number of significant variables as the 2-year model before the financial distress event. However, whilst there are 8 overlapping ratios between the models, 3 of them differ. The variables of *NP/SR*, *NP/ACA*, and *NA/NOS* are exclusive to the 2-year model. They are replaced in the 3-year model by *EBIT/IE*, *EBIT/ATA* and *CA/TA*. There is a significant difference in the order of coefficients among these models. For example, the net profit/average fixed assets ratio obtains the highest coefficient in model 1.3, but a very small coefficient in model 1.1.

According to Table 3, models in year 1 contain 8 significant variables, while 9 is the number of significant variables in the remaining models using variable set 2. In addition, the variables chosen are nearly the same for 3 models in different time of prediction. Most of selected variables reflect the solvency, profitability, and asset development of the companies. However, while the coefficient sign of total assets growth ratio is supposed to be positive, it is found to be negative in all 3 models for 1 year before the distress event.

Table 3. Stepwise selection results for logistic regression models – variable set 2

	Variable	Coefficient		
		Model 1.4	Model 1.5	Model 1.6
	Year 1			
1	(CA–I)/CL	–0.132	–0.147	–0.154
2	IE/E	0.407	0.519	0.475
3	IE/SR	0.920	0.631	0.533
4	TA(t)/TA(t–1)	0.118	–0.335	–0.352
5	RE/TA	–0.206	–0.063	–0.066
6	OIBT/TA	–1.152	–0.893	–0.938
7	GP/NS	–0.119	–0.341	–0.468
8	TL/TA	0.162	0.363	0.440

Table 3 (cont.). Stepwise selection results for logistic regression models – variable set 2

	Variable	Coefficient		
		Model 1.4	Model 1.5	Model 1.6
	Year 2			
1	WC/SR	–0.141	–0.248	–0.170
2	IE/SR	0.41	0.353	0.316
3	CF/TL	–0.079	–0.183	–0.292
4	CF/E	–0.308	–0.423	–0.545
5	RE/TA	–0.307	–0.274	–0.223
6	GP/NS	–0.175	–0.284	–0.228
7	TA(t)/TA(t–1)	–0.300	0.152	0.441
8	NP/ASE	–0.267	–0.380	–0.499
9	TL/TA	0.448	0.350	0.281
	Year 3			
1	WC/SR	–0.141	–0.245	–0.226
2	IE/SR	0.241	0.248	0.325
3	CF/TL	–0.308	–0.217	–0.301
4	CF/E	–0.143	–0.147	–0.161
5	RE/TA	–0.207	–0.132	–0.341
6	GP/NS	–0.281	–0.358	–0.235
7	TA(t)/TA(t– 1)	–0.267	–0.275	0.454
8	NP/ASE	–0.180	–0.185	–0.143
9	TL/TA	0.236	0.240	0.289

3.1.2. F-score selection method results

In addition to factor analysis, the other filter feature selection is applied in SVM by calculating the F-score of each variable. Using LIBSVM, the variables with *F*-score bigger than 0.3, 0.04 and 0.03 for data sets 1, 2 and 3, respectively, are selected by the program. Table 7 presents the predictors chosen according to their *F*-scores for 3 data sets.

As shown in Table 4 and Table 5, the results of the filter selection process show that the smallest number of selected variables is found in model with the smallest sample size. However, there is not much difference in the orders of selected predictors according to their *F*-scores among each three models. For example, the net profit on the number of ordinary shares ratio receives the highest F-score in model 2.2 and 2.3 and it also gets the second highest *F*-score in model 2.1. Similarly, acid test ratio obtains the highest F-score in all 3 models 2.4 to 2.6. According to predictor group, the variables selected by F-score mainly reflect the capital expansion capacity, profitability and operational capacity of a company.

Table 4. Features selected in SVM models – variable set 1

No.	Model 2.1		Model 2.2		Model 2.3	
	Variable	F-score	Variable	F-score	Variable	F-score
1	NA/ NOS	0.686	NP/NOS	0.844	NP/NOS	0.774
2	NP/ NOS	0.510	EBIT/ATA	0.688	NP/ASE	0.696
3	TA(t)/ TA(t–1)	0.449	NP/ATA	0.608	NP/ATA	0.549
4	EBIT/ ATA	0.439	NP/ASE	0.418	NP/ACA	0.381
5	NP/ATA	0.431	NP/ACA	0.395	TL/TSE	0.361
6	TL/TA	0.429	TL/TA	0.352	EBIT/ATA	0.355
7	NP/AFA	0.335	TL/TSE	0.347	NA/NOS	0.243
8	NP/ACA	0.307	NP/SR	0.327	NP/SR	0.233
9	–	–	NA/NOS	0.216	TL/TA	0.232
10	–	–	CA/CL	0.205	MBI/ATA	0.121
11	–	–	CL/TA	0.165	CA/CL	0.117
12	–	–	MBI/ATA	0.113	NP/AFA	0.107
13	–	–	NP/AFA	0.108	CL/TA	0.074
14	–	–	CS/APA	0.099	(CA–I)/ CL	0.073
15	–	–	(CA–I)/CL	0.094	CS/APA	0.055
16	–	–	CA/TA	0.043	MBI(t)/ MBI(t–1)	0.038

Table 5. Features selected in SVM models – variable set 2

No.	Model 2.4		Model 2.5		Model 2.6	
	Variable	F-score	Variable	F-score	Variable	F-score
1	(CA-I)/ CL	0.701	TL/TA	0.873	TL/TA	0.881
2	TA(t)/ TA(t–1)	0.525	(CA–I)/ CL	0.711	(CA-I)/ CL	0.522
3	IE/SR	0.354	TA(t)/ TA(t–1)	0.638	TA(t)/ TA(t–1)	0.412
4	CF/TA	0.424	CF/TA	0.432	CF/TA	0.286
5	CF/TL	0.446	CF/TL	0.408	CF/TL	0.271
6	CF/E	0.644	CF/E	0.364	CF/E	0.571
7	RE/TA	0.350	CRI	0.706	CRI	0.783
8	OIBT/ TA	0.322	OIAT/S	0.338	OIAT/S	0.618
9	GP/NS	0.455	RE/TA	0.223	RE/TA	0.429
10	NI/E	0.427	OIBT/ TA	0.212	OIBT/ TA	0.406
11	–	–	GP/NS	0.651	GP/NS	0.399
12	–	–	NI/E	0.554	NI/E	0.273

3.2. Comparison of models' performance

3.2.1. Logistic regression model's performance

Using different sets of predictors as results of factor analysis and the stepwise regression procedure, the overall significance of 3 models 1.1, 1.2, and 1.3 is tested. The overall significance is tested by running the Omnibus test, –2 Log likelihood, and the Hosmer and Lemeshow test. A significant model obtains a small –2 Log likelihood, a significant Omnibus test and an insignificant Hosmer and Lemeshow test. As presented in Table A5 and Table A6 in Appendix A, models with different data sets are significant for 3 years of prediction.

Table 6 presents the level of accuracy of 3 logistic models 1-3 years prior to the financial distress event. In terms of the sample size, the level of accuracy increases from model 1.1 to model 1.3 using variable set 1 and from model 1.4 to model 1.6 using variable set 2. Therefore, the smaller the percentage rate of financially distressed firms in the models, the higher the model's accuracy. In terms of prediction time, there is a slight increase of the model's accuracy as the time of prediction progresses. In general, the accuracy rates of all three models are quite high with the highest level of 86% belonging to model 1.3 that makes a prediction 3 years in advance.

Table 6. Logistic regression classification accuracy (%)

Model	1-year ahead	2-year ahead	3-year ahead
Model 1.1	64.70	66.20	73.50
Model 1.2	72.50	72.50	76.90
Model 1.3	84.60	85.30	86.00
Model 1.4	63.00	62.00	66.17
Model 1.5	66.23	68.00	75.61
Model 1.6	76.20	82.5	84.00

The performance of a model can be assessed further by looking at their Type I error and Type II error. A Type I error is detected when a financially distressed firm is classified as non-financially distressed, while a Type II error is an error whereby a non-financial distressed firm is predicted to be financially distressed by the model. A Type I error should receive more attention from the huge potential losses that can be brought to the model's users. A Type I error varies as the time of prediction increases. However, Type I error decreases dramatically as the sample size increases. Surprisingly, Type I error stays the same for model 1.3 in all prediction years (Table 7).

http://dx.doi.org/10.21511/imfi.16(1).2019.22

Table 7. Type I errors of logistic regression models (%)

Model	1-year ahead	2-year ahead	3-year ahead
Model 1.1	21	24	19
Model 1.2	25	24	35
Model 1.3	15	15	15
Model 1.4	34	28	23
Model 1.5	22	25	27
Model 1.6	20	19	17

3.2.2. *SVM model's performance*

From the original number of variables, according to the F-score calculated of each variable as shown in Table 4, 5 predictors are chosen for models 2.1 to 2.6. The function to build the hyperplane is related to value of C and gamma. C and gamma are selected through the work of grid search (see Table A7 in Appendix A). Using the optimal hyperplane constructed by the choice of C and gamma, each SVM model classifies a firm in testing data into a distressed group and a non-distressed group. The accuracy levels of 6 SVM models at different points of prediction time can be summarized in Table 8.

As can be seen, there is an increase in the model's accuracy level as the time of prediction progresses. The level of accuracy that can be reached peaks at nearly 87% in model 1.3 for 3 years before the distress event. In terms of the data set used, the model's accuracy improved as the rate of distressed firms decreased.

Table 8. Summary of classification – SVM models (%)

Model	1-year ahead	2-year ahead	3-year ahead
Model 2.1	77.45	79.41	79.41
Model 2.2	77.45	82.35	85.29
Model 2.3	83.09	83.09	86.76
Model 2.4	77.60	80.80	79.00
Model 2.5	80.00	81.74	83.20
Model 2.6	84.80	85.70	84.40

Compared to the Type I error calculation in the logistic regression model with the same data set and time of prediction, Type I error in the SVM model is extremely low (Table 9). In particular, there is no Type I error that can be detected in model 2.2 and model 2.3 and 2.3 for 3 years of prediction.

Table 9. Type I error of SVM models (%)

Model	1-year ahead	2-year ahead	3-year ahead
Model 2.1	1	4	1
Model 2.2	0	0	0
Model 2.3	1	0	0
Model 2.4	4	3	0
Model 2.5	2	2	2
Model 2.6	0	2	1

CONCLUSION

The study is conducted with the intention to consider the combined effects of 3 factors: predictor choices, sampling technique, and classification techniques to the performance of an ex-ante model for listed firms in the securities market of Vietnam. In terms of predictors selection, the factor analysis is used in logistic regression models, while *F*-score method is calculated to select the predictors in SVM models. Regarding the sampling techniques, three data sets with different numbers of non-distressed firms are created. Each prediction model is constructed from 1 to 3 years before the firm receives the requirement of being delisted from the securities exchanges because of poor financial performance.

There are 12 models are constructed from 2 different original variable sets: set 1 is taken from the study of Geng et al. (2014), while set 2 comes from the paper of Lin et al. (2014). Although factor analysis and F-score all reduce the number of variables in all models dramatically, the predictors choices made by these two approaches are quite different. From those selected variables, the performance of each models

is assessed by computing the accuracy level and Type I error. The analysis results show that the model's performance increases in data set containing larger numbers of non-distressed companies or smaller distress rates. This finding is similar with that of the paper from Zmijewski (1984), and Shaonan et al. (2015). Regarding to classifier, it can be seen that logistic models underperform SVM models. This detection is also supported by Lin et al. (2011, 2014) and Kumar (2015) in their related researches. In addition, the SVM model combined with the *F*-score selection method, which was applied in the biggest sample size, outperforms other models because of the highest level of accuracy and smallest Type I error. Therefore, for future research, the complete sampling should be selected to test whether the reduction in the inconsistency between the financial distress rate of the sample and that of the population can further improve the financial distress prediction model.

REFERENCES

1. Alifiah, M. (2014). Prediction of financial distress companies in the trading and services sector in Malaysia using macroeconomic variables. *Procedia - Social and Behavioral Sciences,* 129, 90-98.
2. Altman, E. I. (1968). Financial Ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance, 23,* 589-609.
3. Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2016). Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model. *Journal of International Financial Management & Accounting, 28*(2), 131-171. https://doi.org/10.1111/jifm.12053
4. Altman, E. I. (1984). A further empirical investigation of the bankruptcy cost question. *The Journal of Finance, 39*(4), 1067-1089. https://doi.org/10.1111/j.1540-6261.1984.tb03893.x
5. Back, B., Laitinen, T., & Sere, K. (1996). Neural Networks and Bankruptcy Prediction: Funds Flows Accrual Ratios and Accounting Data. *Advances in Accounting, 14,* 23-37.
6. Balcaen, S., & Ooghe, H. (2006). 35 Years of Studies on Business Failure: An Overview of the Classic Statistical Methodologies and Their Related Problems. *The British Accounting Review, 38*(1), 63-93. https://doi.org/10.1016/j.bar.2005.09.001
7. Beaver, W. H. (1966). Financial ratios as predictors of failures. *Journal of Accounting Research, 4,* 71-111.
8. Beaver, W. H., McNichols, M. F., & Rhie, J.-W. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies, 10,* 93-122.
9. Bharath, S., & Shumay, T. (2008). Forecasting Default with the Merton Distance to Default Model. *The Review of Financial Studies, 21*(3), 1339-1369. https://doi.org/10.1093/rfs/hhn044
10. Bhattacharjee, A., & Han, J. (2014). Financial distress of Chinese firms: Microeconomic, macroeconomic and institutional influences. *China Economic Review, 30,* 244-262. https://doi.org/10.1016/j.chieco.2014.07.007
11. Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*(1-2), 245-271.
12. Chen, Y. W., & Lin, C. J. (2003). Combining SVMs with various feature selection strategies. In *NIPS 2003 feature selection challenge* (pp. 1-10).
13. Christidis, A., & Gregory, A. (2010). *Some new models for financial distress prediction in the UK* (Discussion paper No. 10/04). Xfi centre for finance and investment. http://dx.doi.org/10.2139/ssrn.1687166
14. Geng, R., Bose, I., & Chen, X. (2014). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research, 241*(1), 236-247. http://dx.doi.org/10.1016/j.ejor.2014.08.016
15. Gepp, A., & Kumar, K. (2015). Predicting Financial Distress: A Comparison of Survival Analysis and Decision Tree Techniques. *Procedia Computer Science, 54,* 396-404.
16. Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Reading, MA: Addison-Wesley.
17. Hsu, C., Chang, C., & Lin, C. (2016). A Practical Guide to Support Vector Classification. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
18. Kim, H. S., & Sohn, S. Y. (2010). Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research, 201,* 838-846.
19. Kittler, J. (1978). Feature set search algorithms. In C. H. Chen (Ed.), *Pattern Recognition and Signal Processing* (pp. 41-60). Netherlands: Sijthoff and Noordhoff, Alphen aan den Rijn.
20. Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1), 273-324.
21. Koopman, S. J., & Lucas, A. (2005). Business and default cycles for credit risk. *Journal of Applied Econometrics, 20,* 311-323.

http://dx.doi.org/10.21511/imfi.16(1).2019.22

22. Liang, D., Tsai, C., & Wu, H. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems, 73,* 289-297.

23. Lin, F., Liang, D., Yeh, C., & Huang, J. (2014). Novel feature selection methods to financial distress prediction. *Expert Systems with Applications, 41*(5), 2472-2483.

24. Mselmi, N., Lahiani, A., & Hamza, T. (2017). Financial distress prediction: The case of French small and medium-sized firms. *International Review of Financial Analysis, 50,* 67-80. https://doi.org/10.1016/j.irfa.2017.02.004

25. Norton, C., Smith, L., & Ralph, E. (1980). A Comparison of General Price Level and Historical Cost Financial Statements in the Prediction of Bankruptcy: A Reply. *The Accounting Review, 55*(3), 516-521. Retrived from https://www.jstor.org/stable/246414

26. Ohlson, D. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109-131.

27. Pindado, J., Rodrigues, L., & Torre, C. (2008). Estimating financial distress likelihood. *Journal of Business Research, 61,* 995-1003.

28. Powell, W. B. (2007). Approximate Dynamic Programming: Solving the Curses of Dimensionality. Hoboken, NJ: Wiley-InterScience.

29. Rees, W. P. (1995). *Financial analysis.* London: Prentice-Hall.

30. Sánchez, L., García, V., Marqués, A., & Sánchez, J. (2016). Financial distress prediction using the hybrid associative memory with translation. *Applied Soft Computing, 44,* 144-152.

31. Santoso, N., & Wibowo, W. (2018). Financial Distress Prediction using Linear Discriminant Analysis and Support Vector Machine. *Journal of Physics Conference Series, 979*(1). Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/979/1/012089

32. Sayari, N., Mugan, C. S. (2016). Industry specific financial distress modeling. *BRQ Business Research Quarterly, 20*(1), 45-62. http://dx.doi.org/10.1016/j.brq.2016.03.003

33. Shaonan, T., Yan, Y., & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance, 52*(C), 89-100.

34. Song, Q., Jiang, H., Zhao, X., & Wu, X. (2017). Combination of minimum enclosing balls classifier with SVM in coal-rock recognition. *PLoS One, 12*(9). https://doi.org/10.1371/journal.pone.0184834

35. Tinoco, M. H., Holmes, P., & Wilson, N. (2018). Polytomous response financial distress models: The role of accounting, market and macroeconomic variables. *International Review of Financial Analysis, 59,* 276-289. https://doi.org/10.1016/j.irfa.2018.03.017

36. Ugurlu, M., Aksoy, H. (2006). Prediction of corporate financial distress in an emerging market: the case of Turkey. *Cross Cultural Management: An International Journal, 13*(4), 277-295. https://doi.org/10.1108/13527600610713396

37. Wan, J. W., Yang, M., & Chen, Y. J. (2015). Discriminative cost sensitive Laplacian score for face recognition. *Neurocomputing, 152,* 333-344.

38. Yjlmaz, E. (2013). An Expert System Based on Fisher Score and LS-SVM for Cardiac Arrhythmia Diagnosis. *Computational and Mathematical Methods in Medicine,* 1-6. http://dx.doi.org/10.1155/2013/849674

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

40. Zhou, L., Lai, K., & Yen, J. (2012). Empirical models based on features ranking techniques for corporate financial distress prediction. *Computers & Mathematics with Applications, 64*(8), 2484-2496.

APPENDIX A

Table A1. List of independent variables – variable set 1

Source: Adapted from Geng, Bose, and Chen (2014).

Variables	Description
TL/TA	Total liabilities/total assets
CA/CL	Current assets/current liabilities
(CA-I)/CL	(Current assets-inventory)/current liabilities
TL/TSE	Total liabilities/total shareholders' equity
CL/TA	Current liabilities/total assets
NOCF/CL	Net operating cash flow/current liabilities
EBIT/IE	Earnings before interest and tax (EBIT)/interest expense
(SR–SC)/SR	(Sales revenue–sales cost)/sales revenue
NP/SR	Net profit/sales revenue
EBIT/ATA	Earnings before income tax/average total assets
NP/ATA	Net profit/average total assets
NP/ACA	Net profit/average current assets
NP/AFA	Net profit/average fixed assets
NP/ASE	Net profit/average shareholders' equity
MBI/ATA	Main business income/average total assets
SR/ACA	Sales revenue/average current assets
SR/AFA	Sales revenue/average fixed assets
MBC/AI	Main business cost/average inventory
MBI/ABAR	Main business income/average balance of accounts receivable
CS/APA	Cost of sales/average payable accounts
MBI(t)/MBI(t–1)	Main business income of this year/main business income of last year
TA(t)/TA(–1)	Total assets of this year/total assets of last year
NP(t)/NP(t–1)	Net profit of this year/net profit of last year
CA/TA	Current assets total assets
FA/TA	Fixed assets/total assets
SE/FA	Shareholders' equity/fixed assets
CL/TL	Current liabilities/total liabilities
NP/NOS	Net profit/number of ordinary shares at the end of year
NA/NOS	Net assets/number of ordinary shares at the end of year
NICCE/NOS	Net increase in cash and cash equivalents/number of ordinary shares at the end of year
CR/NOS	Capital reserves/number of ordinary shares at the end of year
SO	1 if having state ownership and 0 otherwise

Table A2. List of independent variables – variable set 2

Source: Adapted by Lin et al. (2014).

Variables	Description
TL/TA	Total liabilities/total assets
CA/CL	Current assets/current liabilities
(CA–I)/CL	Acid test
NCI	No-credit interval
TA(t)/TA(t–1)	Total assets growth
WC/TA	Working capital/total assets
WC/SR	Working capital/sales
IE/E	Interest expenses/equity
MVE/TD	Market value equity/book value of total debt
IE/SR	Interest expense/salerevenue
CF/TA	Cash flow/total assets
CF/TL	Cash flow/total liabilities
CF/E	Cash flow/equity
CRI	Cash re-investment ratio
OIAT/S	Operating income after tax per share
RE/TA	Retained earnings/total assets
OIBT/TA	Operating income before tax/total assets
OI/E	Operation income per employee
GP/NS	Gross profit/net sales
NI/E	Net income/equity

http://dx.doi.org/10.21511/imfi.16(1).2019.22

Table A3. Results of factor analysis – variable set 1

Variables	Communality	Eigenvalue	Factor loading	Variables	Communality	Eigenvalue	Factor loading
Model 1.1 (data set 1)							
TL/TA	0.83	6.02	0.76	NP/AFA	0.87	1.64	0.71
CA/CL	0.89	3.32	0.91	NP/ASE	0.86	2.83	0.82
(CA–I)/CL	0.89	2.87	0.92	CS/APA	0.88	1.35	0.78
CL/TA	0.91	1.84	0.62	NP(t)/NP(t–1)	0.98	3.22	0.98
NP/SR	0.93	1.31	0.91	CA/TA	0.81	1.17	0.85
EBIT/ATA	0.87	1.13	0.91	CL/TL	0.81	1.09	0.86
NP/ATA	0.93	1.09	0.94	NP/NOS	0.85	1.08	0.84
NP/ACA	0.83	1.72	0.79	NA/NOS	0.81	3.01	0.63
Model 1.2 (data set 2)							
TL/TA	0.83	6.02	0.76	NP/ACA	0.98	1.77	0.79
CA/CL	0.89	3.32	0.91	NP/AFA	0.82	1.53	0.71
(CA–I)/CL	0.89	2.87	0.92	NP/ASE	0.86	1.83	0.82
TL/TSE	0.87	2.35	0.63	MBI/ATA	0.83	1.38	0.78
CL/TA	0.91	1.83	0.62	CS/APA	0.88	1.35	0.78
NP/SR	0.93	1.32	0.91	MBI(t)/MBI(t–1)	0.93	1.38	0.86
EBIT/ATA	0.87	1.13	0.91	NP/NOS	0.85	1.08	0.84
NP/ATA	0.93	1.09	0.94	NA/NOS	0.81	1.03	0.63
Model 1.3 (data set 3)							
TL/TA	0.91	18.46	0.82	NP/AFA	0.87	2.71	0.74
CA/CL	0.95	11.58	0.96	NP/ASE	0.87	2.47	0.80
(CA–I)/CL	0.91	10.07	0.94	MBI/ATA	0.80	2.28	0.65
TL/TSE	0.83	7.58	0.84	SR/ACA	0.90	1.90	0.88
CL/TA	0.86	5.88	0.60	MBI/ABAR	0.81	1.28	0.56
EBIT/IE	0.83	4.72	0.85	CS/APA	0.90	1.03	0.87
NP/SR	0.94	3.93	0.92	MBI(t)/MBI(t–1)	0.94	1.02	0.85
EBIT/ATA	0.89	3.49	0.91	CA/TA	0.86	1.44	0.90
NP/ATA	0.92	3.20	0.92	NP/NOS	0.86	1.22	0.87
NP/ACA	0.99	3.13	0.78	–	–	–	–

Table A4. Results of factor analysis – variable set 2

Variables	Communality	Eigenvalue	Factor loading	Variables	Communality	Eigenvalue	Factor loading
Model 1.1 (data set 1)							
(CA-I)/CL	0.87	1.31	0.85	RE/TA	0.81	1.77	0.64
IE/E	0.88	1.09	0.99	OIBT/TA	0.85	1.54	0.68
IE/SR	0.98	2.35	0.97	GP/NS	0.87	1.38	0.70
TA(t)/TA(t–1)	0.91	1.84	0.74	TL/TA	0.93	1.53	0.76
Model 1.2 (data set 2)							
WC/SR	0.82	1.76	0.80	RE/TA	0.86	1.52	0.59
IE/SR	0.93	1.34	0.82	GP/NS	0.90	1.29	0.63
CF/TL	0.84	2.60	0.92	TA(t)/TA(t–1)	0.92	1.13	0.65
CF/E	0.86	1.59	0.69	NP/ASE	0.90	1.28	0.71
	–	–	–	TL/TA	0.84	1.24	0.55
Model 1.3 (data set 3)							
WC/SR	0.81	1.99	0.66	RE/TA	0.81	1.17	0.83
IE/SR	0.90	1.57	0.68	GP/NS	0.89	1.64	0.87
CF/TL	0.83	2.83	0.78	TA(t)/TA(t–1)	0.91	1.48	0.79
CF/E	0.85	1.24	0.55	NP/ASE	0.92	1.63	0.57
–	–	–	–	TL/TA	0.83	1.59	0.69

Table A5. Model's overall significance – variable set 1

Prediction time	Omnibus tests		Nagelkerke R-Square	Hosmer and Lemeshow test	
	Chi-square	Sig.		Chi-square	Sig.
Model 1.1					
1-year ahead	51.50	0.00	0.708	1.58	0.99
2-year ahead	77.18	0.00	0.905	4.00	0.85
3-year ahead	76.53	0.00	0.901	0.49	1.00
Model 1.2					
1-year ahead	65.54	0.00	0.849	8.35	0.40
2-year ahead	74.84	0.00	0.766	4.82	0.77
3-year ahead	66.02	0.00	0.346	6.79	0.56
Model 1.3					
1-year ahead	87.66	0.00	0.704	2.67	0.95
2-year ahead	88.62	0.00	0.709	1.74	0.98
3-year ahead	123.15	0.00	0.882	5.57	0.69

Table A6. Model's overall significance – variable set 2

Prediction time	Omnibus tests		Nagelkerke R Square	Hosmer and Lemeshow test	
	Chi-square	Sig.		Chi-square	Sig.
Model 1.4					
1-year ahead	61.319	0.00	24.205	3.413	0.906
2-year ahead	87.696	0.00	65.791	3.042	0.932
3-year ahead	72.418	0.00	29.024	3.735	0.880
Model 1.5					
1-year ahead	81.332	0.00	45.298	4.539	0.806
2-year ahead	8.214	0.00	95.730	2.829	0.945
3-year ahead	46.950	0.00	122.080	4.539	0.806
Model 1.6					
1-year ahead	20.598	0.00	65.415	4.539	0.806
2-year ahead	123.931	0.00	65.259	3.735	0.88
3-year ahead	46.950	0.00	154.569	2.829	0.945

Table A7. Summary of C and gamma

Model	Parameter	1-year ahead	2-year ahead	3-year ahead
Model 2.1	C	256	256	256
	Gamma	0.000	0.500	0.000
Model 2.2	C	64	1	256
	Gamma	0.031	0.008	0.000
Model 2.3	C	4	1	256
	Gamma	0	0.002	0.008

http://dx.doi.org/10.21511/imfi.16(1).2019.22