Adv. Math. Fin. App., 2022, **7(**3), P. 811-835

Advances in Mathematical Finance & Applications
www.amfa.iau-arak.ac.ir
Print ISSN: 2538-5569
Online ISSN: 2645-4610
Doi: 10.22034/AMFA.2021.1905055.1461

Applied-Research Paper

Experimental Comparison of Financial Distress Prediction Models Using Imbalanced Data Sets

Seyed Behrooz Razavi Ghomi[a], Alireza Mehrazin[a,*], Mohammadreza Shoorvarzi[a], Abolghasem Masih Abadi[b]

[a]*Department of Accounting, Neyshabur Branch, Islamic Azad University, Neyshabur, Iran*
[b]*Department of Accounting, Sabzevar Branch ,Islamic Azad University, Sabzevar, Iran*

ARTICLE INFO

Article history:
Received 2020-07-22
Accepted 2021-03-01

Keywords:
Imbalanced Data Sets.
Financial Distress Prediction Models.
Grid search optimization.
Tuning parameters.
Financial ratios.

ABSTRACT

From machine learning perspective, the problem of predicting financial distress is challenging because the distribution of the classes is extremely imbalanced. The goal of this study was comparing the performance of financial distress prediction models for the imbalanced data sets with different proportions. In this study, the data of the previous year before financial distress was used for 760 company year for the time period of 2007-2017. Besides using traditional classifications such as logistic regression, linear discriminant analysis, artificial neural network, and the classification models of least square support vector machine with four kernel functions, random forest and the Knn algorithm, the measures of the area under the curve and Friedman and Nemenyi tests were also utilized to determine the average rank and the difference significance of the Auc of the models. For selecting the models´ optimal parameters, the combined method of grid search optimization and cross validation was used. The results of this experimental study showed that for the balanced and imbalanced datasets with lower proportions, the best performance was for the random forest. For more imbalanced datasets, the best performance belonged to the least square support vector machine with sigmoid, radial, and linear kernel functions; performance of Knn algorithm had no significant difference from the other models and the performance of the artificial neural network was average or appropriate. Also, the performances of the linear logistic regression and linear discriminant analysis were weaker than other nonlinear models.

1 Introduction

Recent financial crises have made the financial systems vulnerable more than before. Firms with various sizes are suffering financial problems. Those problems influence financial institutions, stockholders, managers, employees, and governments; thus, it is important to predict financial distress of the firms. In turn, this critical corporate issue has become a major research area in the corporate finance field although different financial distress prediction models have been suggested using a variety of variables and prediction methods [10]. The datasets of the financial distress consist of the same number of financially distressed and healthy firms. But, it has been mostly ignored in the real world

* Corresponding author. Tel.: +989155518842
E-mail address: mehrazeen@gmail.com

conditions since the number of healthy firms is high in comparison to the financially distressed companies. Thus, scholars face imbalanced datasets in financial distress prediction. The issue of imbalanced data has been examined from two perspectives. First, acknowledging that financial distress prediction models can be used for an imbalanced data set as a representative of the real society, the very low frequency of the financially distressed companies reduces the predictability of the models. The second is the proposed corrective techniques to manipulate the imbalanced data sets and improve the accuracy of the financial distress prediction models. Although these views provide a basis for understanding this issue, the key question remain as to whether the imbalanced distribution of the classes disrupts the predictability of the models. The proportion of the data set may be imbalanced in financial distress prediction. Because in reality, the bankruptcy rate of the companies does not follow a specific rule. Datasets may present multiple imbalanced class levels that contain different proportions of financially distressed firms because of the irregular bankruptcy rates in the population, the rare bankrupt firms, and the lack of accessibility to these firms' information [84]. To assess whether a fiaaiii ll distress prddittion mllll 's rr ddictinn aapaii lity is nnMMMM rrrr essigg tee imlll ccced proportion that significantly impairs the performance of the model becomes important. In addition, financial distress is an important issue that has a social cost.

Although the issue of the imbalanced data has attracted the attention of the researchers ([59,73]), to date, the field of financial distress prediction lacks a perspective on the different proportions of imbalanced data sets and performance loss in the financial distress prediction models. The imbalanced data set is important in the field of financial distress prediction because in the real world, the number of financially distressed companies is much lower than that of healthy companies. As a result, the misclassification cost of the financially distressed companies will increase. Since the misclassification cost of financially distressed companies is higher than that of healthy companies, it is important to compare and select the best models for the financial distress prediction. So far, there has been a lot of research on financial distress prediction with balanced data sets and the number of the healthy and financially distressed companies has been equal; but, in the field of financial distress prediction, the issue of the impact of the imbalanced data sets with different proportions on the performance of a wide range of financial distress prediction models has not been considered. Although the research methods are improved to greatly enhance the prediction accuracy, the research objects of most literature merely focus on the balanced credit risk evaluation ([29,57]), which is different from the actual situation. Recently, the study on the class-imbalanced credit risk evaluation has become more and more popular. These studies generally considered that the number of risky companies was much smaller than those of normal corporations.

If the samples of different classes were selected by the equal proportion, the performance of the models would be not only overestimated, but unsuitable for the actual situation ([80,22]). Therefore, to fill such a gap in this study, the performance of financial distress prediction models based on 6 scenarios from imbalanced data sets with different proportions of 50 / 50%, 60 / 40%, 70 / 30%, 80 /20%, 90 /10%, and 95 / 5% were investigated. In other words, the effect of increasing the proportion of imbalanced classes on the performance of financial distress prediction models was studied. For this means, 9 financial distress prediction models based on the area under the ranking curve and significance in ranking financial distress prediction models using Friedman [33] and Nemenyi [68] tests were examined. Financial distress prediction models included linear discriminant analysis, logistic regression, and least square support vector machine with four different kernel functions, random forest, and the nearest neighbor algorithm. Besides, setting optimal parameters or hyper parameter has

important effect on the performance of the financial distress prediction algorithm. The selection of the inappropriate parameters of financial distress prediction models can lead to the over fitness and under fitness issues. Since little attention has been paid to the issue of selecting optimal parameters when creating financial distress prediction models and the default parameters have been mainly used, the combined method of grid search optimization and cross validation with number 5 was used in this study. This is one of the most widely used methods for selecting optimal parameters. The reason for using this method was its high computational speed and confidence about searching the whole parameter space rather than an approximate and exploratory search. There is also no need to specify a parameter to select the optimal parameters. Choosing a model that has high predictability for the imbalanced data sets is important for the investors and creditors because improving financial distress prediction models will imrr vve iss sstrr ,, rr dditors add otee∆ stkkooolrrr '' ccciii ,,,, ll timtt ely laadigg to tee inrr essdd welfare of society.

2 Literature Review

2.1 Imbalanced Data Set

Classification performance of an algorithm is affected when the under study dataset is highly imbalanced ([15,18,6,21]). Traditionally, classification algorithms are driven to increase the predictive accuracy of the derived classifiers. However, maximizing the overall accuracy may not be the best approach in case of an imbalanced dataset. While maximizing the overall accuracy, a classifier focuses on the majority class as it has the higher weight in the data. As a result, the classifier can achieve a high degree of accuracy on the majority class and by extension on the overall dataset while performing poorly on the minority set. It is worth noting that identifying the minority instances is often of greater importance as in the case of rare diseases.

In machine learning, classifiers are derived to minimize misclassification errors and thereby maximize prediction accuracy [15]. The underlying assumption in these classification methods is that the dataset under study has a roughly balanced number of instances per available class. In other words, it assumes that the prior probabilities of the target classes are similar [36]. However, in many real-world domains, such as medical diagnostics, most of the classification data tends to be skewed towards negative class value. Data is said to be imbalanced if at least one of the target variable values has a significantly smaller number of instances when compared to the other values. Class imbalance is especially prevalent in the models used to detect rare but important diseases such as autism spectrum disorder ([72, 73]). However, in many real-world domains, such as medical diagnostics, most of the classification data tend to be skewed towards the negative class value. Data is said to be imbalanced if at least one of the target variable values has a significantly smaller number of the instances when compared to the other values. Class imbalance is especially prevalent in the models used to detect rare but important diseases such as autism spectrum disorder ([82, 83]).

2.2 Financial Distress Prediction Models Using Imbalanced Data Sets

The first researcher who studied bankruptcy was Fitzpartrick [32]. He adopted the univariate ratio to analyze the distinction between the successful enterprises and failed firms. With the development of economic globalization, the risk management gained for attention. By the 1960s, the focal research rr lll mmff tee finnnii ll maaagmnttt wss wwwto timll y ddd ccrrr ately vvll ttt e cmmiiii ss' rikk. Aftrr that, the research on credit risk evaluation officially kicked off. Beaver followed the univariate analysis method to compare the distinction of financial ratios of the failure firms with those of normal companies

[11]. Altman used multivariate discriminant analysis more than once to build the discrimination model by multiple financial indicators [4]. Ohlson first applied regression model (logistic regression) to evaluate business credit risk [70]. Then, Zmijewski applied probit regression model instead of logistic regression [94]. These typical methods belong to statistical theories and econometrical models. After the 1990s, because the demand for risk management in the emerging credit market was increasing day by day and the artificial intelligence was rising, the risk management was brought into the new era. Since hypothesis conditions are relaxed, these intelligence models, such as artificial neural network (ANN) [46], support vector machine (SVM) [72], decision tree (DT) [91], and so on, are more suitable for the credit risk evaluation than the traditional statistical and econometrical models. Therefore, these intelligence algorithms were widely applied for credit risk evaluation. As a typical intelligence algorithm, ANN was first adopted by Odom and Sharda for bankruptcy prediction.

They built a back propagation neural network with a three-layer and compared it with multivariate discriminant analysis. The result showed that ANN can effectively improve the prediction accuracy [69]. Then, the credit risk evaluation widely adopted artificial intelligence models to make the research [24, 45]. As in many other domains, nature-inspired algorithms were also successfully used for bankruptcy prediction [87]. 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 in them. Methods for bankruptcy should definitely take into account this imbalance in order to prevent type I and II errors, 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 [87,55]. We explore the predictive capacity of bankruptcy models in imbalanced datasets. Data and their characteristics are the most crucial elements of any prediction model [5]. Conceptually, a dataset presents a class imbalance if it contains unequal distributions between classes. However, literature generally accepts that a dataset is imbalanced when one class significantly outnumbers another [44]. Although imbalanced datasets appear frequently in the classification field (as bankruptcy predictions), classification models tend to expect equal misclassification costs because it is the prevailing scenario. The models are designed to optimize the overall accuracy; they do not take into account the relative distribution of each class [61]. Such degradation of prediction, caused by the imbalanced distributions, occurs during the learning phase (training set); classification models tend to concentrate on the accurate classifications of the majority class, while ignoring the minority class; because classification rules maximize the overall prediction accuracy.

That is, the decision (classification) boundary of the majority class tends to invade the decision boundary of the minority class [47]. As a result, in the prediction phase (test set), models often are biased toward the majority class; they accurately classify the majority classes but frequently misclassify the minority classes. Thus, the issue of imbalanced datasets originates from the learning phase, where tee clssii firrs' prdditt inn rrr frr mccce is ii tturddd, sscccially with rggrrd to tee mirrr ity ll ass. Although the imbalanced dataset issue has received attention by the scientific community to date, the bankruptcy prediction has lacked insights into the relationships among the imbalance proportion and the loss of performance [52, 73]. In the case of bankruptcy prediction and due to the limited instances in the minority class, an imbalance scenario is the representative of the domain since bankrupt firms are rare. Moreover, bankruptcy prediction domain presents two additional peculiarities that make it a rather a challenging task. On the other hand, samples are described by the financial attributes because they give g ' igg ff tee firm'' finccc ill situtt i... Hwwvrr, evnn tggggh ccch ttt a prssett mjj or vvvttt ggss to predict bankruptcy and their importance may not be neglected, the fact that they can be manipulated

[76,19] may lead to a distortion that can be detrimental to the bankruptcy models. On the other hand, firms that follow similar paths in their deterioration may have different outputs. It is not uncommon that some firms acquire a sort of ability to allow them to survive more easily than others, while apparently their financial situation suggests no difference [25] in which some data points may appear as the (valid) examples for the bankrupt or non-bankrupt firms. In Weiss and Provost [88], it was found that the naturally occurring class distributions in the 25 data sets looked at, often did not produce the best-performing classifiers. It was shown that the optimal class distribution should contain between 50% and 90% minority class examples within the training set.

Wilson and Sharda [90] provide a primary example: They use three sample proportions on the training set, including a balanced sample (composed of 50% failed and 50% non-failed firms) and two imbalanced proportions (20% failed and 80% non-failed firms; 10% failed and 90% non-failed firms) to analyze the prediction performance of the discriminant analysis and neural network methods. These authors find that prediction methods achieve better results, especially in the failed firms, when the training set presents a balanced sample. Therefore, in this study, a comprehensive study of the imbalanced data set and its effect on the performance of financial distress prediction models was conducted. The complexity of financial data in financial distress prediction may be exacerbated as an important factor in reducing the predictability of the models by adding the issue of imbalanced data was examined. Therefore, in this study, the complexity of financial distress prediction data in the context of imbalanced data sets. Since financially distressed companies are representatives of the minority class, the consequences of the misclassification cost of financially distressed companies include capital loss and the contagion effects. Misclassification of financially distressed companies can lead to the decline of the whole economy and its components, including employment and economic well-being [10]. In addition, some of them have compared the performance of these different prediction models [9].

2.3 Background and Literature Review

Ever since Beaver [11] and Altman [4] first studied bankruptcy prediction, the classic paradigm of the sample selection for the bankruptcy models has been to choose balanced samples with available financial information, in which the proportions of the bankrupt and non-bankrupt firms are equal. Balanced samples can be produced using a popular technique known as the paired sample, in which data containing firms that eventually failed are usually paired according to the size, industry, or age criteria with the firms that did not fail [46]. This sample selection strategy provides a clear advantage, because it avoids class bias during the learning phase. The classifier maximizes the overall prediction regardless of the class distribution. However, the strategy also has a serious drawback for it does not represent the real-world proportion. Zmijewski [84] demonstrates that if failed and non-failed proportions do not represent the real world population, the sample selection bias may occur, leading to the underestimations of failed firms and overestimations of non-failed firms.

Moreover, Ooghe and Joos [71] claim that samples of failing and non-failing firms should be the representatives of the whole population of the firms so that failure prediction models can be used in a predictive context, especially in the failed firms, when the training set presents a balanced sample. In addition, Mc Kee and Greenstein [62] investigated the capacity of three bankruptcy prediction methods in five highly imbalanced data sets and showed that the imbalanced sample distribution in the learning phase causes poor classification performances, especially for the bankrupt firms. Another frequently used machine-learning method for bankruptcy prediction is DTs. In a seminal paper about this method the authors compared DTs with discriminant analysis, genetic algorithms, and NN methods. Their

results showed that the DT method provides interpretable results [66], Other studies employing this method belong to Lee et al. [53] and Yeh and Lien [92]. In general, CBR is based on previous cases that created precedence for solving similar problems in the future. The models based on the CBR method mostly use the Euclidean distance and k-nearest neighbor method. CBR is more suited to smaller data samples and is similar to human decision making [51]. Even though an initial study [43] suggested that CBR is not a suitable method for the bankruptcy prediction, several later studies [1, 22, 55, 77] achieved results showing that the prediction performance was comparable with other ML methods. SVM gained popularity for the bankruptcy prediction in the late 2000s [56]. Several papers [20,30,64,65] compared the prediction accuracy of SVM with that of NN (and other methods) and the results of all those aforementioned studies suggested that the performance of SVM was 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 the machine learning [2].

These baseline methods can be combined in several ways in order to boost the accuracy or to overcome certain shortcomings of the individual classifiers. There are two basic ensemble and hybrid approaches for combining different classifiers. 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 [58]. The hybrid approach combines different classification techniques sequentially [8]; [30]. In contrast, Desai et al. [28], found that the neural network significantly outperformed linear discriminant analysis. Baesens et al. [9] used seventeen techniques, including two well-known methods such as logistic regression and linear discriminant analysis, and more advanced techniques such as least square supporting vector machine for eight sets of imbalanced data sets, although more complex techniques such as radial functions of the least squares of the support vector machine and neural network performed well in terms of AUC criteria; but, simpler classifiers such as linear discriminant analysis and logistic regression also performed well; however, inconsistent results were mostly observed in comparing financial distress prediction models. Aiming to compare tree decision models and multiple discriminant analysis, Gepp et al. [34] selected a sample of 200 companies with a proportion of 71% healthy companies and 29% financially distressed companies. They wanted to predict financial distress of the companies using step-wise variable selection technique. Their results showed that the decision tree model with 87.6% accuracy outperformed the multiple discriminant analysis with 84.5% accuracy. Kim [48] examined the financial distress prediction models of a sample including 66 companies with a proportion of 50% healthy companies and 50% financially distressed companies by the step-wise variable selection to predict the financial distress of the companies.

The accuracy of the support vector machine was 95.95%, artificial neural network´s was 91.6%, multiple discriminant analysis´ was 72.6. %, and logistic regression´s was 80%. De Anders et al. [26] selected a sample of 122 companies, including 61 healthy companies and 61 financially distressed companies with the aim of improving financial distress prediction models, using the neural network and multiple discriminant analysis. Results showed the 76.03% accuracy of artificial neural network and 74.87% accuracy of multiple discriminant analysis. Arieshanti et al. [7] selected a sample of 240 companies with a proportion of 53% healthy companies to 47% financially distressed companies with the aim of comparing financial distress prediction models using stepwise variable selection method. Their results showed 70.42% accuracy of the support vector machine model and 71% accuracy of the artificial neural network. With the aim of presenting a new predictive variable selection method, Zhou et al. [93] selected a sample of 2010 companies with a proportion of 50% healthy companies to 50%

financially distressed companies using genetic algorithm method. Results showed the 89.42% accuracy of the support vector machine, 93.27% accuracy of the artificial neural networks, 77.88% accuracy of the multiple discriminant analysis, and 81.73% accuracy of the logistic regression. Iturriaga and Sans [42] selected a sample of 772 companies with the proportion of 50% healthy companies to 50% financially distressed companies to improve the artificial neural network using combined variable selection method of Gini coefficient. The results showed 75.6% accuracy of artificial neural network, 50.67% accuracy of the decision tree, 71.72% accuracy of the multiple differential analysis, and 73.99% accuracy of the logistic regression.

Rezaei and Tolaminejad [74] used ant colony optimization algorithm as a model for predicting financial distress and compared it with multiple discriminant analysis and Logit methods. After selecting 15 predictor variables for 130 companies including 40 financially distressed companies and 90 healthy companies from 2005 to 2010, the experimental findings showed the highest accuracy of Logit, ant colony optimization algorithm, and multiple discriminant analysis, respectively. Stayesh et al. [79] examined the usefulness of random forest classifiers and the method of relief variable selection to select the optimal prediction variables. Experimental findings of 95 healthy companies and 95 bankrupt companies in Tehran Stock Exchange from 2001 to 2016 indicated the better performance of the random forests than the logistic regression. In addition, research findings showed the usefulness of the variable selection method in predicting financial distress. In other words, if the relief variable selection method is used, the mean accuracy significantly will increase while the first and second order error decrease. Namazi et al. [67] examined and compared the usefulness of different methods of selecting predictive variables in predicting financial distress of the listed companies in Tehran Stock Exchange. In this regard, 98 healthy companies and 98 financially distressed companies were selected and the performances of variable selection methods, including t-test, multiple discriminant analysis, factor analysis, relief, and support vector machine were examined and compared. The used classifiers also included artificial neural networks, support vector machine, and AdaBoost (boosting). Fallahpour and Eram [31] examined financial distress prediction using ant colony algorithm. Used classifiers included artificial neural networks, support vector machine, and AdaBoost (boosting). Statistical population of the study included 174 healthy and financially distressed companies in TSE. The predictor variables were selected based on the proportions that were the main predictor variables in the prediction model of the previous research. The comparative model used in this study was multiple discriminant analysis. The results indicated that the ant colony algorithm method has a significantly better performance than the multiple discriminant analysis method in financial distress prediction of the companies. Taking a new approach, Botshekan et al. [14] aimed to predict financial distress for selecting effective variables using the opinion of the experts and decision-making algorithms.

For this purpose, they considered 29 random financial proportions for the financially distressed manufacturing companies based on the Article 14 of the Business Law and the same number of the healthy companies from the listed companies of Tehran Stock Exchange since 2006-2016, using the audited financial statements for one, two, and three years before the distress. In the end, they predicted financial distress by the support vector machine. The results showed that the suggested model in one, two and three years before the occurrence of financial distress has a better performance compared to the logistic and Altman regression methods. Rezaei and Javaheri [75] compared the neural network's predictability with the combined method of the genetic algorithm and artificial neural network. For this purpose, they selected a sample of 58 healthy companies and 49 financially distressed companies. In order to compare these methods, determination coefficient, MSE and RMSE were used. The results showed 97.7% accuracy of the artificial neural network and 100% accuracy of the combined method of

the artificial neural network and genetic algorithm in predicting healthy and financially distressed companies. Therefore, the combined method of artificial neural network and genetic algorithm was the best way to predict the financial crisis of the companies. Sarouei [78] examined the revenue of Springit, Zimsky, and Olsen models in the pharmaceutical and textile industries. Their results showed that in all three years, the used models in the textile industry outperformed the others. Ghasemi and Sarlak [35] examined the impact of the financial crisis on the financial transparency and conservatism in the banking industry. To this end, they collected data from 18 banks since 2011-2015. The results of a linear regression analysis showed the correlation and the impact of the financial crisis on the financial transparency and conservatism.

3 Research Methodology

This study was quasi-experimental, using a post-event approach for applied goals. Theoretical foundations of the research were collected from Persian and Latin specialized journals and websites and the financial data were collected from Rahavard Novin and Tadbir Pardaz software and Codal website. In order to predict financial distress, the data of one year before the financial distress since 2007 to 2017 was utilized. For predicting financial distress of the companies, the classification models, including support vector machine with four different kernel functions of random forest, artificial neural network, nearest neighbour, logistic regression, multiple discriminant analysis was used. In order to select the optimal parameters and improve the validity of the model, the results from the combined optimization method of grid search and cross validation of the number 5 was used so that 80% of the total data entered the training process each time and 20% of the total data entered the financial distress prediction test process. Sampling method was linear cross validation; thus, all data were tested. The arrangement of healthy and financially distressed samples was such that the proportion of healthy and financially distressed companies at any given time was equal to the above proportion for the total data. In the following section, for comparing the predictability of financial distress models, the area under the curve and for examining the significance of the performance difference of financial distress prediction models at 95% confidence level for six balanced and imbalanced data sets with different degrees, Friedman and Nemenyi tests were used. All models in this research were performed with Rapid Miner software and Friedman and Nemenyi tests with Xlstat software.

Table 1: Feature of Imbalanced Data Sets

Proportion of healthy to financially distressed companies	Number of total companies	Number of healthy companies	Number of financially distressed companies	Training set size	Test set size
%50-%50	760	380	380	608	152
%60-%40	760	456	304	608	152
%70-%30	760	532	228	608	152
%80-%20	760	608	152	608	152
%90-%10	760	684	76	608	152
%95-%5	760	722	38	608	152

3.1 Data

In this study, six data sets with 760 company years and different proportions, including 50/50%, 60/40%, 70/ 30%, 80/ 20%, 90/ 10%, and 95/ 5% were used, respectively. In order to create an imbalanced data set with the mentioned proportions, the Brown and Mues [17] method was used. First, 380 healthy and 380 financially distressed companies were randomly selected; then, the first 380 financially distressed companies were deleted and the same number of healthy companies were

randomly added so that the total number of the companies for the six imbalanced data sets always remained 760 companies. Table 2 shows the samples selection method in terms of financially distressed and healthy companies.

3.2 Variables

After studying the literature on financial distress prediction of the companies, 64 predictor variables were selected and the dependent variable was determined based on Article 141 of the Business Law, so that if a company was subject to Article 141 of the Business Law in a way that the company's accumulated losses exceeded half its capital, it would be regarded financially distressed otherwise healthy. Table 2 shows the list of 64 independent predictor variables and the ways of their calculation.

Table 2: Financial Ratios in Financial Distress Prediction

Row	Variable name	Row	Variable name	Row	Variable name
1	NI/SE	24	S/FA	47	CL/TL
2	NI/TA	25	S/SE	48	D/NI
3	OCF	26	S/TA	49	EPS
4	OCF/SE	27	SE/TA	50	EBIT/IE
5	OCF/CL	28	SE/TL	51	EBIT/S
6	OCF/IE	29	Size(log TA)	52	EBIT/TA
7	OCF/S	30	TIBL/TL	53	FA/(SE+LTD)
8	OCF/TA	31	TL/TA	54	FA/TA
9	OCF/TL	32	WC/S	55	GP/S
10	OCF/NI	33	WC/TA	56	IE/GP
11	OCF/OI	34	(Ca+STI)/CL)	57	IE/S
12	NI/GP	35	(R+Inv)/TA)	58	Inv/WC
13	OI/S	36	P/S	59	Inv/S
14	OI/TA	37	R/S	60	LTD/SE
15	PIC/SE	38	Ca/CL	61	LTD/TA
16	QA/CL	39	Ca/TA	62	MVE/TA
17	QA/Inv	40	NI/S	63	MVE/TL
18	QA/TA	41	CA/CL	64	MVE/SE
19	R/Inv	42	CGS/Inv		
20	RE/Inv	43	CA/S		
21	RE/SC	44	CA/TA		
22	RE/TA	45	CL/SE		
23	S/Ca	46	CL/TA		

The definitions of the above predictor variables are as follows:
CA: is current assets, NI: is net profit, Ca is cash balance, OI is operating income, CL is current liability, QA is quick assets, PIC is paid capital, R is receivables, EBIT is earnings before interest and taxes, RE is retained earnings, FA is fixed assets, S is income, GP is gross profit, SC is stock capital, IE is financial expenses, SE is stock equity, INV is inventories, STI is short term investments, TA is Total assets, LTD is long-term debts, TL is total debts, MVE is market value equity, WC is working capital, OCF is operating cash flow, D is dividend, TIBL is total interest liabilities.

3.3 Overview of Financial Distress Prediction Models

A large number of classification models have been designed to predict financial distress. None of the

classification models is significantly superior to other models [10]. In this study, nine classification models with different features were used. The two methods of logistic regression and linear discriminant analysis were derived from the concepts of statistical decision theory, although they rely on linear functions but have valid results. The other seven methods included artificial neural networks, random forest, nearest neighbor algorithm, and support vector machine with linear, radial, sigmoid, and multi-nominal kernel functions that focus on the model learning process and have been developed to directly predict the data and rely on nonlinear approaches for testing complex data sets.
Linear discriminant analysis

3.3.1 Linear Discriminant Analysis

The LDA method is among the first ones used to predict bankruptcy [4]. It assumes that class-conditional densities follow Gaussian distributions and that the distributions have a common covariance matrix [86]. When LDA is employed to discriminate between failed and non-failed firms, it needs only to estimate the distributions´ means and their common covariance; LDA creates a discrimination score (z-score) to distinguish two classes by combining explanatory variables on a linear function. The z-score is computed as follows in Eq.1:

$$z = \sum_{i=1}^{n} (xiwi + c) \quad (1)$$

Where xi represents explanatory variables, wi indicates the discriminant weights, and c is a constant. Although LDA assumes Gaussianity on the class-conditional distributions and equal covariance matrices and these assumptions do not hold in corporate failure, it has been widely used for its robustness [10].

3.3.2 Logistic Regression

Ohlson [70] proposed LR to model the posterior probabilities of the classes, using linear functions of the independent variables, while ensuring that they sum to 1 and remain in [0,1] to provide a probabilistic interpretation. Similar to LDA, LR makes use of the log-likelihood ratio to assign a firm to either failed or non-failed classes; the log-ratio takes the form of a linear function. This method allows the use of non-lieerr miii mum likll idddd ɗ sstimtt e firm'' rr bbiii litiss of failure uii gg a lggitt ic function based on the dependent variables, in this case, financial ratios. The LR method takes the following form in Eq.2:

$$z = \frac{1}{1 + e^{-(w0 + wjxj)}} \quad (2)$$

Where, Xi are explanatory variables, Wi are the weights estimated using maximum likelihood estimation, and z is the score for a given firm. Although both statistical methods, LR and LDA, have similar forms in their discriminant functions, the estimation of their parameters is quite different. The LR method makes fewer assumptions than the LDA method and is generally considered, in statistical literature, to be a safer method.

3.3.3 Neural Networks

The NN technique is a mathematical model that emulates the function of a human brain. It is an efficient model for statistical pattern recognition [12], providing a general framework for representing

non-linear functional mapping between sets of input variables and output variables. It is designed by establishing an architecture that connects neurons among layers. In this study, we focus on the multilayer perceptron (MLP), composed of three layers, and including an input layer composed of n neurons for input variables, a hidden layer composed of m neurons, and an output layer. Every neuron in the hidden layer is connected to every neuron in the input and output layers. We estimate the connectivity weights that is, the parameters of the NN representing the relevance of the connections between neurons-by a back propagation learning method. An NN model computes z-score that represents the failure probability of a given firm, as follows in Eq.3:

$$z = g(\sum_{J=0}^{M} Wkjg\,(\sum_{j=0}^{d} wji\ xi)) \tag{3}$$

Where g the activation function, $x_i$are explanatory variables; w_{ij}corresponds to the weight matrix, including the bias term between the input node (i) and the hidden node (j); and corresponds to the weight matrix with bias connecting the hidden node to the output layer. Since Messier and Hansen [63] introduced the NN method to the study of corporate failure, authors may have applied it because of its ability to learn complex nonlinear relationships and its good adoption to data.

3.3.4 Random Forest

The random forest classifier is made of a group of decision trees, in which each classifier is generated using a random vector sampled independently from the input vector [16]. In RF, each tree is built from a bootstrap sample of the data and at each split; a random sample of predictors is examined. In the end, classification is determined by a majority vote for each case over the ensemble of classification trees. When constructing a tree, RF searches for a random subset of the input features (bands) at each splitting node and the tree is allowed to grow fully without pruning. Since only a portion of the input features is used and no pruning is required, random forest is computationally fast and simple with with a good performance.

3.3.5 Support Vector Machines

The machine learning community has widely adapted the SVM, proposed by Boser et al [13] for data classification. An SVM classifier maps training vectors into a higher dimensional space, where it finds a separating hyper plane with a maximal margin. The attractiveness of the SVM method arises largely because there is no need to know the form of the high-dimensional mapping function; it is necessary only to know its inner product, such that any dissimilarity function, even a non-linear function that holds some mild coiii tinn aan ee eee.. Tii s fett ure is kwwwm ss tee kkrr eel trikk”” ([36]; [81]). The classification capacity of the SVM relies on the ability to transform the input space into a more elaborate feature space in which the separability of the classes is enhanced in a margin-maximization condition that increases generalization capability by constraining the structure of the model. An SVM is defined as follows in Eq.4:

$$\begin{aligned} &MIN\ w.b.\ e = \tfrac{1}{2}\ w^t w + c\sum_{i=1}^{N} ei \\ &Subject\ to\ yi\ (w\ \Omega(xi)+b)+ei-1\geq 0 \quad ei\geq 0 \end{aligned} \tag{4}$$

eee r,, Ω (ii) msss trii ning vcctrr s to a high ii msss illll aaaee; w is the weigtt vcctrr ; b is tee ii as term; c is the penalty for the error; and ei is the slack variable [85]. When the optimal hyper plane separation between classes is built, a classification decision is given as follows in Eq.5:

$$f(y) = sign(\sum_{i=1}^{N} yipik(x, xi) + b \quad (5)$$

3.3.6 K-NN Algorithm

The nearest neighbor algorithm classifies the nearest data neighborhood by voting for the majority of k-data with high similarity. In this study, the criterion of similarity of Euclidean distance between data was used as follows in Eq.6:

$$d(xi, xj) = \|xi - xj\| = [(xi - xj)^T (xi - xj)]^{1/2}. \quad (6)$$

3.4 Area Under the Curve

Unfortunately, some of the most common performance appraisal metrics are suitable for evaluating the performance of a balanced data classifier but not for the imbalanced data. Accuracy rates, as the most common measure of performance appraisal in predicting financial distress, do not consider sample distribution. Imagine that 95% of the observations are in one class and the other 5% are in the other class. The prediction method has the accuracy of 95% because it focuses only on the prediction of the majority class. As a performance evaluation metrics, prediction accuracy rate suggests the accurate classifier but in fact, predicting the minority class is ignored. In accordance with the above points, it is necessary to modify the performance evaluation of the model and rely on the evaluation metrics that are not sensitive to the distribution of the sample. In this study, the metrics of the area under the curve was used to compare the performance of financial distress prediction models.

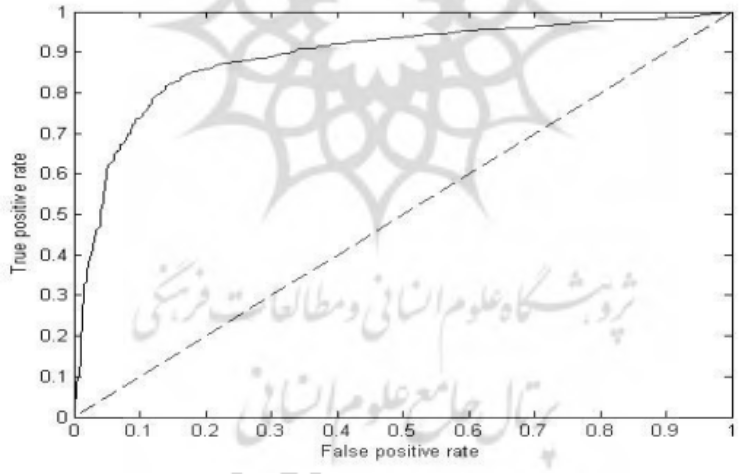

Fig. 1: An example of ROC curves (x axis represents 1-specificity and y axis represents sensitivity)

Many studies have used this measure to evaluate the overall performance of the classification models with imbalanced data for it is insensitive to the cost of misclassification and imbalanced distribution. Area under the curve provides a demonstration of the trade-off between a true positive (failed firms that have been correctly classified) and a false positive (failed firms that have been misclassified) [37]. The area under the curve can be easily used to compare two classifiers. The ROC curve for a classifier should be as far to the top left corner as possible, where its value will be close to 1. In Fig. 1, the classifier with the solid line outperforms the one with the dashed line.

3.5 Statistical Comparison of Financial Distress Prediction Models

ee ssdd rr immm̓i’ s tsst [33] to compare the AUCs of the different classifiers. The Friedman test statistic is based on the average ranked (AR) performances of the classification techniques on each data set, calculated as follow in Eq.7:

$$x_F^2 = \frac{12D}{K(K+1)}\left[\sum_{j=1}^{K} AR_j^2 - \frac{k(k+1)^2}{4}\right], where\ AR_j = \frac{1}{D}\sum_{i=1}^{D} r_i^j. \quad (7)$$

Where, D denotes the number of data sets used in the study, K is the total number of classifiers, and r^j_i is the rank of classifier j on data set i. x^2_F is distributed according to the Chi-square distribution with k-1 degrees of freedom. If the value of x^2_F is large enough, then the null hypothesis that there is no difference between the techniques can be rejected. The Friedman statistic is well suited for this type of data analysis as it is less susceptible to the outliers [33]. The post-hoc Nemenyi test [68] is applied to report any significant difference between individual classifiers. The Nemenyi post-hoc test states that the performances of two or more classifiers are significantly different if their average ranks differ by at least the critical difference (CD), given by Eq.8:

$$CD = q_{a,\infty,K}\sqrt{\frac{K(K+1)}{12D}} \quad (8)$$

In this formula, the value $q_{a,\infty,K}$ is based on the student range statistic [68]. Finally, the results from iii mmm̓i’ s sttt istic add tee Nmnyyyi tttt ooc tssts rre ii sll yydd uii ng a modified version of Demsar [27] significance diagrams [54]. These diagrams display the ranked performances of the classification techniques along with the critical difference to clearly show any techniques which are significantly different to the best performing classifiers.

3.6 Parameters Optimization with Combined Method of Grid Search and Cross Validation

Grid search is a display of search based on a defined subset of the optimal parameters space (hyper parameters). The reasons for using the grid search optimization method are as follows: There may not be enough confidence in using meta-heuristic optimization methods; because meta-heuristic and approximate optimization methods prevent the full search of parameters. Another reason is that computational time to find the values of optimal parameters is lower compared to other more advanced methods [52]; optimal parameters are determined using the minimum value (lower boundary) and the maximum value (upper boundary) and the number of the stages. The performance of each combination of parameters is evaluated using the area under the curve.

The total procedure of selection stages of optimum parameters using the combined method of grid search optimization and cross validation is as follows: 1. k-1 combination of the parameters is selected, 2. data is divided into K proportion in each trial, k-1 proportion is selected as a training set and the remaining proportion is selected as the model test set, 3.every combination of the parameters will be tested k times, 4. the average k of the area under the curve is obtained for each separate combination of the parameters, and 5. the average area under the curve of different parameter combinations are compared and the best parameters are selected. The optimum parameters are the ones which have the

best ranked performance measurement metrics. For increasing the reliability of the test set results and preventing overfitting and underfitting, the combined method of grid search optimization and cross validation with number 5 was used. The sampling method was linear cross validation; so, all companies were tested. For example, if the proportion of financially distressed companies to healthy companies is 90 to 10, the proportion for each section will be also 90 to 10 and at each stage of the experiment, 80% of the companies will belong to the training set and the remaining 20% as the test set. For example, in determining the optimal parameters of gamma, if 12 different values are selected for each parameter, a total of 144 different combinations of parameters will be tested. If cross validation with number 5 is used, a total of 144×5 = 720 separate experiments will exist. Since determining optimal parameters plays an important role in improving the performance of classification models and classifier models are highly sensitive to the value of the parameters, the selection of optimal parameters is very important. Not much attention has been paid to the previous research on the parameters optimization of the classification models; so, in order to fill the existing gap, the optimal parameters of classifier models including least square support vector machine with 4 radials, sigmoid, linear, and polynomial kernel functions, neural network, random forest, and nearest neighbor algorithm with grid search optimization were used.

3.6.1 Parameter Tuning

Although linear discriminant analysis and logistic regression do not require tuning specific parameters, a set of optimal parameters should be tuned for the other classification models. The linear, radial, sigmoid, poly nominal kernel, and least square support vector machine functions were used. Radial, sigmoid, and poly nominal kernel functions have two parameters of C and gamma, and the linear kernel function has parameter C. C and gamma kernel parameter have vital roles in the performance of the support vector machine ([39]; [81]). Therefore, the improper selection of these two parameters can lead to the overfitting and underfitting issues. However, there are few practical guidelines for determining optimal parameters. Hsu et al. [39] suggested practical guidelines for setting the parameters of the support vector machine using the grid search method and cross validation.

The purpose of setting C and gamma parameters is to maximize the accuracy of the classification model for the unobserved data. The exponential growth of both C and gamma parameters is a practical method to set optimal parameters. In this study, the values of 2^{-5},2^{-3} ,2^{-1} ,2, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11}, 2^{13}, and 2^{15} were used for parameter C and the values of 2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-5}, 2^{-7}, 2^{-3}, 2, 2^{3}, and 2^{5} were implemented for the gamma parameter. A total of 121 combinations of parameters were selected. The multilayer perceptron neural network was used with a single layer, the learning rate and the momentum rate were set from 0.1, 0.2, and 0.3; thus, a total of 9 combinations were set for the artificial neural network. In order to determine the optimal parameter number of the random forest trees with gain_ ratio criterion, 28 values in the range of 10 to 1000 and the nearest neighbor algorithm with Euclidean distance criterion, 28 values of k parameter in the range of 1 to 1000 were tuned.

4 Research Results

Tables 3-8 show the area under the curve of the financial distress prediction models for six different imbalanced data sets with different proportions of imbalanced classes. For each proportion of the imbalanced data, Friedman test and p value were shown to be correlated. In case of the existence of significant difference at 95% confidence level, the post-hoc Nemenyi [68] test was conducted for each imbalanced data sets. Financial distress prediction model which has the highest area under the curve

has the highest rank among other financial distress prediction models. For examining the difference significance in ranking financial distress prediction models, the performance measurement metric of the area under the curve was used. In the main imbalanced data sets, the performances of logistic regression and linear discriminant analysis were worse than the random forest, the least square support vector machine with radial and sigmoid kernel functions, and artificial neural network with the significance at the confidence level of 95%. There was low difference between financial distress prediction models. An increase in the ability of the financial distress prediction models, even a small percentage, can have future savings and significant benefits for the investors, creditors, and other users [38]. The higher the predictability of the model, the closer the average rank of the model will be to one. In the figures below, financial distress prediction models have been classified into different groups. By the pair comparison of the groups, the models that were not common between the two groups significantly differed in the statistical significance at 95% confidence level.

Table 3: Dataset AUC with 50/ 50% Proportion

Classification model	Cross Validation					Average	Average rank
	1	2	3	4	5		
SVM (RBF)	93.12%	92.66%	97.40%	96.98%	98.06%	95.6 %	3.8
SVM(Linear)	92.95%	92.68%	97.14%	97.66%	98.11%	95.7%	3.2
SVM(Sigmoid)	96.18%	92.87%	96.17%	97.47%	97.03%	95.9%	3.6
SVM(Polynomial)	92.49%	91.49%	96.74%	96.43%	95.49%	94.5%	6
Neural Network	93.64%	93.52%	95.51%	96.17%	97.74%	95.3%	4.4
LDA	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%	8.5
Random Forest	96.39%	95.92%	98.61%	97.42%	98.02%	97.3%	1.8
KNN	94.22%	93.29%	94.64%	95.73%	96.96%	95%	5.2
Logistic Regression	50.00%	50.00%	50.00%	50.00%	50.00%	50%	8.5

Table 4: Dataset AUC with 60/40% Proportion

Classification model	Cross Validation					Average	Average Rank
	1	2	3	4	5		
SVM (RBF)	97.43%	96.82%	94.82%	97.83%	93.62%	%96.1	3
SVM(Linear)	97.67%	96.11%	94.62%	97.69%	93.36%	%95.9	3.2
SVM(Sigmoid)	97.46%	96.52%	96.17%	97.93%	93.27%	%96.3	3
SVM(Polynomial)	93.45%	92.87%	90.49%	97.44%	88.85%	%92.6	6.4
Neural Network	96.31%	%95.90	95.37%	96.23%	95.56%	%95.9	4.2
LDA	50.00%	50.00%	50.00%	50.00%	50.00%	%50	8.5
Random Forest	97.83%	96.67%	96.92%	98.05%	96.55%	%97.2	1.2
KNN	95.15%	92.46%	94.75%	94.45%	92.65%	%93.9	6.2
Logistic Regression	50.00%	93.29%	50.00%	50.00%	50.00%	%50	8.5

According to the tables 3-8 for the imbalanced data sets, the average ranked performance of the area under the curve of the Friedman test are shown. The best average rank of the area under the curve of the imbalanced data sets belonged to the random forest with the average rank of 1.8. When the data is imbalanced with lower proportions, 60%/70% proportion relates to the healthy companies and 40%/30% proportion belongs to the financially distressed companies, respectively. Random forest had the best average ranked performance of the area under the curve with the Friedman test score of 1.2. Thus, in the case of balanced and imbalanced data sets with lower degrees, the best financial distress prediction models belonged to the random forest with higher consistency compared to the other models. With the increase in the imbalanced data intensity, so that 80%, 90%, and 95% of the companies are

healthy and only 20%, 10%, and 5% of the companies are financially distressed, respectively, the best average rank area under the curve belonged to the least square support vector machine with sigmoid, radial, and linear kernel functions with the average rank of 2.8, 2, and 2.9, respectively. Therefore, it can be assumed that in the case of balanced and imbalanced data with higher proportions, the best predictability was related to the random forest model; also, by increasing the proportion of the imbalanced data, the best predictability was related to the least square support vector machine followed by the sigmoid, radial, and linear kernel functions.

Table 5: Dataset AUC with 70/30% Proportion

Classification model	Cross Validation					Average	Average Rank
	1	2	3	4	5		
SVM (RBF)	97.64%	96.91%	93.31%	97.26%	96.60%	% 96.3	2.8
SVM(Linear)	94.78%	91.73%	94.35%	97.14%	92.82%	% 94.2	5
SVM(Sigmoid)	97.91%	95.79%	94.68%	95.62%	95.51%	% 95.9	3.4
SVM(Polynomial)	93.94%	91.83%	87.35%	94.93%	91.39%	% 91.9	6.8
Neural Network	97.57%	96.12%	93.48%	96.60%	94.07%	% 95.6	4.2
LDA	50.00%	50.00%	50.00%	50.00%	50.00%	% 50	8.5
Random Forest	97.69%	96.95%	94.91%	97.37%	98.03%	% 97	1.2
KNN	95.46%	94.42%	92.67%	97.00%	95.88%	% 95.1	4.6
Logistic Regression	50.00%	50.00%	50.00%	50.00%	50.00%	% 50	8.5

Table 6: Dataset AUC with 80/20% Proportion

Classification model	Cross Validation					Average	Average Rank
	1	2	3	4	5		
SVM (RBF)	92.24%	97.51%	%98.26	89.95%	97.36%	% 95.1	3.4
SVM(Linear)	92.30%	97.06%	97.50%	90.09%	98.48%	% 95.1	3.4
SVM(Sigmoid)	92.60%	97.61%	95.94%	90.41%	98.50%	% 95	2.8
SVM(Polynomial)	89.07%	95.96%	98.50%	83.84%	94.49%	% 92.4	5.4
Neural Network	92.68%	96.71%	97.83%	92.54%	96.34%	%95.2	3
LDA	50.00%	50.00%	50.00%	50.00%	50.00%	% 50	8.5
Random Forest	89.64%	95.58%	98.40%	91.33%	95.31%	% 94.1	4.4
KNN	90.08%	90.58%	96.09%	84.59%	96.50%	% 91.6	5
Logistic Regression	50.00%	50.00%	50.00%	50.00%	50.00%	% 50	8.5

Table 7: Dataset AUC with 90/10% Proportion

Classification model	Cross Validation					Average	Average Rank
	1	2	3	4	5		
SVM (RBF)	97.90%	96.43%	97.76%	98.98%	94.42%	%97.1	2
SVM(Linear)	98.95%	95.60%	92.21%	98.78%	85.99%	%94.3	3.6
SVM(Sigmoid)	99.00%	96.17%	91.44%	98.69%	85.91%	%94.2	4
SVM(Polynomial)	95.57%	93.12%	91.48%	98.88%	81.04%	% 92	5.2
Neural Network	98.62%	95.70%	90.85%	97.71%	89.39%	%94.5	4.6
LDA	50.00%	50.00%	50.00%	50.00%	50.00%	% 50	8.5
Random Forest	96.51%	96.61%	97.47%	98.78%	94.54%	%96.8	2.5
KNN	86.94%	81.39%	92.21%	84.72%	85.38%	%86.1	6.1
Logistic Regression	50.00%	50.00%	50.00%	50.00%	50.00%	% 50	8.5

Table 8: Dataset AUC with 95% -5% Proportion

Classification model	Cross Validation					Average	Average rank
	1	2	3	4	5		
SVM (RBF)	95.30%	94.56%	95.89%	84.12%	98.03%	% 93.6	3.4
SVM(Linear)	96.55%	95.57%	94.29%	90.03%	96.95%	% 94.7	2.9
SVM(Sigmoid)	95.48%	92.93%	94.29%	89.36%	97.64%	% 93.9	4.1
SVM(Polynomial)	81.19%	95.49%	87.33%	75.00%	98.52%	% 87.5	4.7
Neural network	93.57%	87.41%	95.66%	99.66%	99.21%	% 95.1	3.3
LDA	50.00%	50.00%	50.00%	50.00%	50.00%	% 50	8.5
Random forest	95.06%	94.17%	92.81%	99.66%	98.52%	% 96	3.4
KNN	73.69%	93.63%	80.71%	86.74%	84.83%	% 83.9	6.2
Logistic Regression	50.00%	50.00%	50.00%	50.00%	50.00%	% 50	8.5

According to the Fig. 2 for the balanced data set, the area under the curve of the random forest model was significant at 95%, outperforming the logistic regression and linear discriminant analysis models. With the data being imbalanced so that 40% proportion belonged to the financially distressed companies and 60% to the healthy companies as shown in Fig. 3, the average rank of the area under the curve of the least square support vector machine and sigmoid and radial kernel functions improved so that along with the random forest, it outperformed the logistic regression and linear discriminant analysis at 95% confidence level. With the increase in imbalanced data proportion from 70% to 30% according to Fig. 4, it was observed that the average rank of the random forest model is still significantly better than the logistic regression and linear discriminant analysis.

In that scenario, the lowest rank belonged to the support vector machine with poly nominal kernel with the average rank of 6.8 occurring among all imbalanced data sets, so that at 95% confidence level, it performed worse than the random forest. Thus, it can be hypothesized that for the low-intensity balanced and imbalanced data sets, the significant difference of the random forest with the logistic regression and linear discriminant analysis has been maintained. With the gradual increase in imbalanced data proportion from 50% / 50% to 70% / 30%, the area under the curve of the least square support vector machine with radial kernels improved from the average rank of 3.8 to 2.8, and a significant difference was created with logistic regression and linear discriminant analysis. Fig. 5 shows an imbalanced data set with the 80% proportion of healthy companies and 20% proportion of financially distressed companies. In contrast to the first three scenarios, the random forest was replaced by a support vector machine with a sigmoid kernel. It was also observed that by the gradual proportion increase of the imbalanced data from 50% / 50% to 80% /20%, the average rank area under the curve of the artificial neural network improved from 4.2 to 3 and there was a significant difference between the models of support vector machine with sigmoid kernel and artificial neural network and logistic regression and linear discriminant analysis.

According to Fig. 6, with the increase of imbalanced data proportion from 90% to 10%, the best averaged rank of the area under the curve belonged to the least square support vector machine with radial kernel function. Thus, that model and random forest model significantly outperformed logistic regression and linear discriminant analysis. In the pre-scenario of the imbalanced data set with 90%/10% proportion, a significant difference was found between the average rank of the support vector machine model and radial kernel with the imbalanced data sets with the proportion of 70% /30% and the random forest model with the balanced and imbalanced data sets and the proportion of 60%/40% and 70%30% in comparison with the logistic regression and linear differentiation analysis models. Finally, according to Figure 8, in the most imbalanced distribution of the classes, so that 95% of the

data belonged to the healthy companies and 5% belonged to the financially distressed companies, the best average rank of the area under the curve was for the least square support vector machine with the linear kernel function, significantly outperforming the logistic regression and linear discriminant analysis at 95% confidence level.

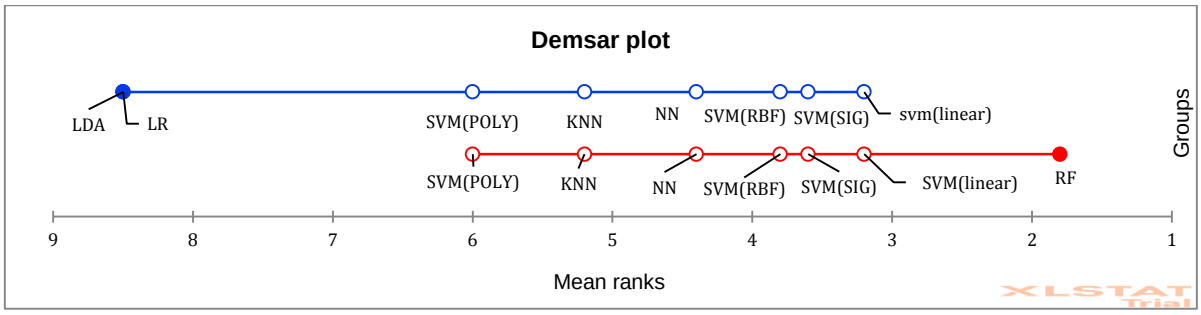

Fig. 2: Average Rank Comparison at a 50/50% Split of Healthy/ Distress Observations.

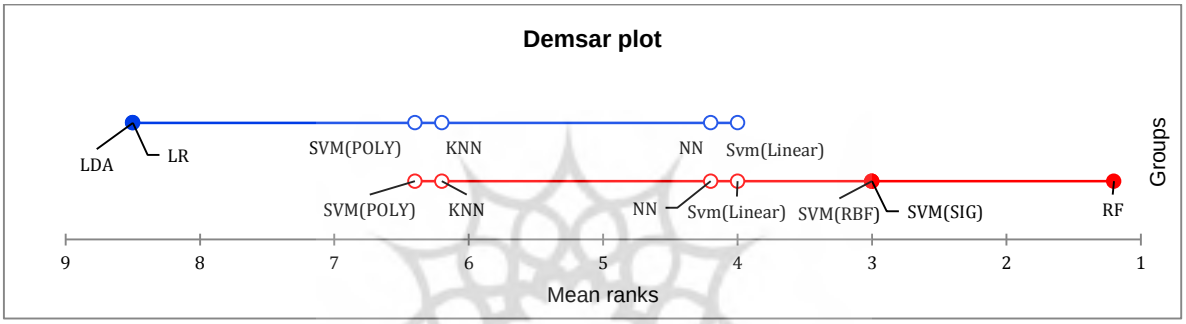

Fig. 3: Average Rank Comparison at a 60/40% Split of Healthy/ Distress Observations.

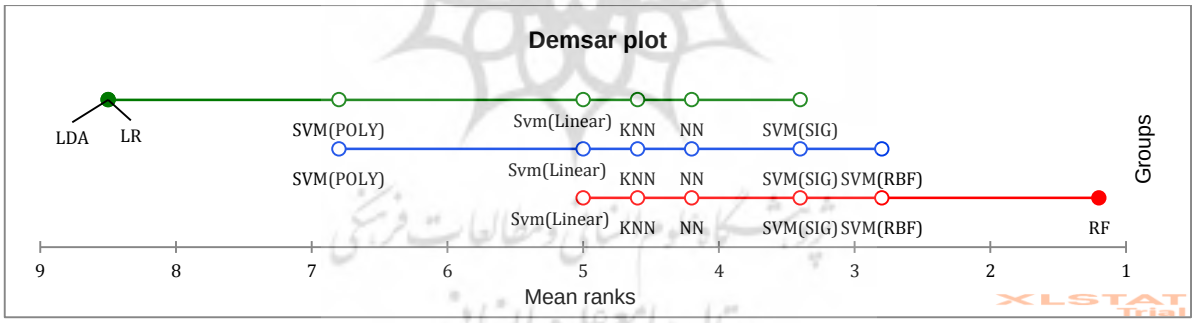

Fig. 4: Average Rank Comparison at a 70/30% Split of Healthy/ Distress Observations.

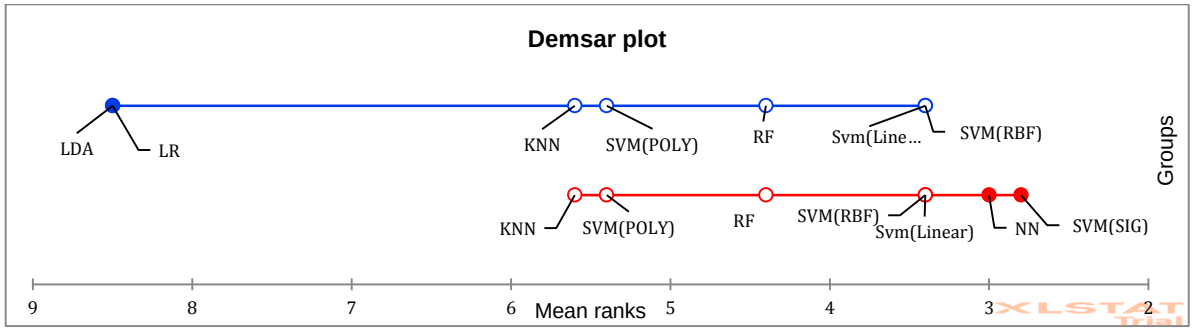

Fig. 5: Average Rank Comparison at a 80/20% Split of Healthy/ Distress Observations.

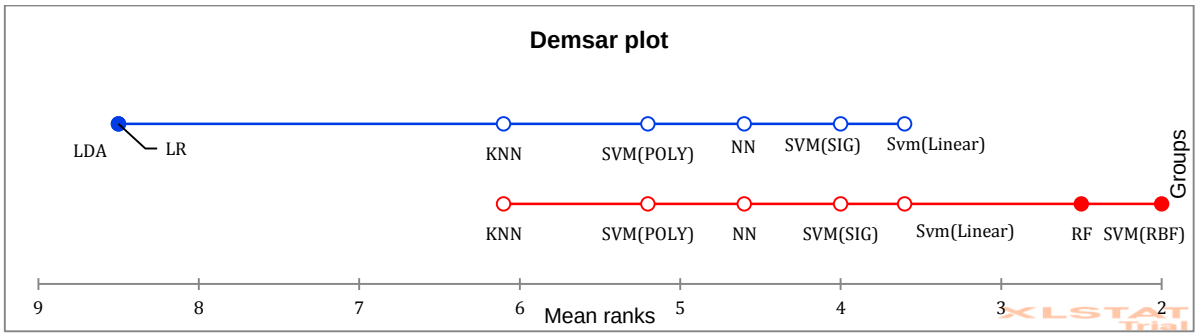

Fig. 6: Average Rank Comparison at a 90/10% Split of Healthy/ Distress Observations

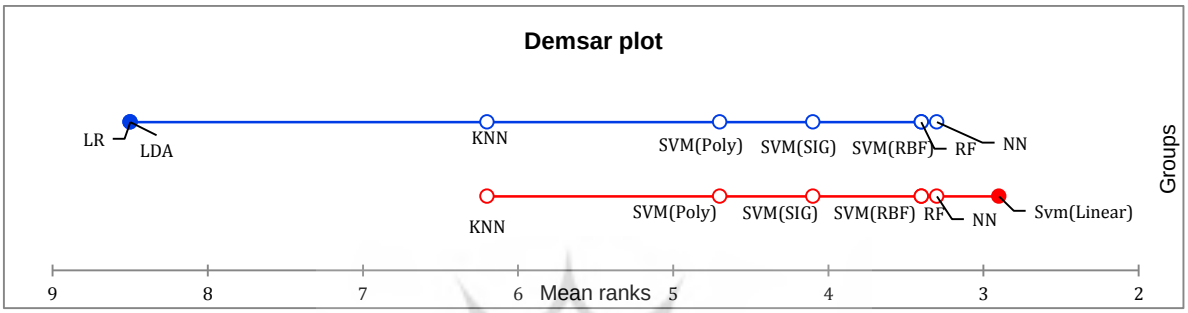

Fig. 7: Average Rank Comparison at a 95/5% Split of Healthy/ Distress Observations

5 Discussions and Conclusion

In summary, it can be is concluded that for the balanced and imbalanced data sets in small scales (the first three scenarios), the random forest and the least square support vector machine with radial and sigmoid kernel functions and for the extreme scales, the imbalanced data sets (three second scenarios) of the least square support vector machine along with the linear and sigmoid kernels significantly outperformed logistic regression and linear discriminant analysis; also, the performances of the statistical and linear models, including logistic regression and linear discriminant analysis were weaker than nonlinear and intelligent models; although no significant difference was observed between the two models of logistic regression and linear discriminant analysis with intelligent and nonlinear models of the nearest neighbor algorithm and the least square support vector machine with poly nominal kernels. In general, the nearest neighbor algorithm showed moderate performance by the increase of imbalanced data intensity, so that it is not statistically significant from the other financial distress prediction models at 95% confidence level. Therefore, it is recommended to use random forest model, artificial neural network, and the least square support vector machine for the imbalanced data sets.

In previous research, the issue of unbalanced data has not been received much attention and mainly the selected sample included the number of healthy and financially distressed companies in equal proportions; but, in practice, the number of healthy companies is more than the number of financially distressed companies. In other words, we are dealing with unbalanced data. The problem of unbalanced data leads to the skewness of classification models towards the majority class, and the ratio of healthy companies to financially distressed companies for specific industries or all companies always varies over different years; so, the classification models should be used which have maximum accuracy in forecasting financially distressed and healthy companies at the same time. According to the above cases, if in a specific industry or all industries in a certain year, the ratio of unbalanced data is 50% to 50%,

60% to 40% and 70% to 30%, it is recommended to use the random forest classification model for financially distressed companies prediction because it is less sensitive than other classification models and has higher accuracy; but, for unbalanced data ratios with higher degrees including 80% to 20% and 90% to 10% and % 95 to 5% , it is suggested to use a the least square support vector machine classification model with radial, sigmoid and linear kernels, respectively. As mentioned above, the issue of imbalanced data causes the decrease of predictability of financially distressed companies and imposes high costs on the investors and creditors. Thus, applied suggestions for the further studies are as follows:

1. It is suggested to use under sampling method that leads to the decreased number of healthy companies as the representatives of the majority, oversampling leading to the increased number of financially distressed companies as the minority class representatives, and combined method that leads to the increase of financially distressed companies and the decrease of healthy companies with a specific percentage. The most important methods of under sampling include clustering-based random under sampling method and the most important oversampling methods, including random oversampling and artificial oversampling of the minority class.
2. Solving the problem of unbalanced data with cost-sensitive learning method: In this approach, the cost of misclassification for financially distressed companies will be determined more than healthy companies and will need to define a matrix for the cost of misclassification. Financially distressed companies will need defining a matrix for misclassification costs. More misclassification costs for financially distressed companies will improve the predictability of financially distressed companies, so it is suggested to increase the accuracy of financially distressed companies' prediction by defining the misclassification cost matrix for financial distress prediction models.
3. Geometric mean criterion: It is suggested to evaluate the performance of the companies in financial distressed prediction models using the geometric mean criterion that simultaneously considers the accuracy of predicting healthy and financially distressed companies. This measure is obtained from the square of multiplying the accuracy of correct prediction of healthy companies by financially distressed companies and is basically a good measure to be used in evaluating the performance of the models with unbalanced data.
4. The power of classification models used in this research is suggested to be used for unbalanced data sets in other areas such as fraud.

References

[1] Ahn, H., Kim, K. J., *Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach*, Applied Soft Computing, 2009, **9**, P.599–607. Doi:10.1016/j.asoc.2008.08.002

[2] Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., Bilal, M, *Systematic Review of Bankruptcy Prediction Models: Towards A Framework for Tool Selection*, Expert Systems with Applications, 2018, **94**, P.164–184. Doi: 10.1016/j.eswa.2017.10.040

[3] Aliabadi, M., Sarraf, F., Darabi, R., *The Power Indexes of the CEO and the Performance of the Company under Pressure Based on Product Market Competition*, Advances in Mathematical Finance and Applications, 2020, Accepted Manuscript Available Online from 21 April 2020 (in Persian). Doi: 10.22034/amfa.2020.1867187.121

[4] Altman, E. I., *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy*, Journal of Finance, 1968, **23**(4), P.889-609. Doi:10.2307/2978933

[5] Anderson, R., *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press, 2007.

[6] Anwar, M. N., *Complexity Measurement for Dealing with Class Imbalance Problems in Classification Modelling*, Thesis for Doctor of Philosophy, Massey University, Institute of Fundamental Sciences, 2012.

[7] Arieshanti, I., Purwananto, Y., Ramadhani, A., Nuha, M. U., Ulinnuha, N., *Comparative Study of Bankruptcy Prediction Models*, TELKOMNIKA (Telecommunication Computing Electronics and Control), 2013, **11**(3), P.591-596. Doi: 10.12928/TELKOMNIKA.v11i3.1095

[8] Azayite, F. Z., Achchab, S., *Hybrid Discriminant Neural Networks for Bankruptcy Prediction and Risk Scoring*, Procedia Computer Science, 2016, P.*83*, P.670–674. Doi:10.1016/j.procs.2016.04.149

[9] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J., *Benchmarking state-of-the-art classification algorithms for credit scoring*, Journal of the Operational Research Society, 2003, **54**(6*)*, P.627–635. Doi:10.1057/palgrave.jors.2601545

[10] Balcaen, S., Ooghe, H., 35 Years of Studies on Business Failure: An Overview of the Classic Statistical Methodologies and Their Related Problems. British Accounting Review, 2006, **38**(1), P.63-93. Doi:10.1016/j.bar.2005.09.001

[11] Beaver, W., *Financial Ratios as Predictor of Failure*, Journal of Accounting Research, 1996, **4**, P.71-111. Doi:10.2307/2490171

[12] Bishop, C. M., *Pattern Recognition and Machine Learning*, Springer, 2006.

[13] Boser, B., Guyon, I., Vapnik, V., *A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh, 1992.

[14] Botshkan, M., Salimi, M., Mottahedjoo, S., *Developing a Hybrid Approach for Financial Distress Prediction of Listed Companies in Tehran Stock Exchange*, Journal of Financial Research, 2018, *20*, P.173-192 (in Persian).

[15] Boyle, T., *Dealing with Imbalanced Data: A Guide to Effectively Handling Imbalanced Datasets in Python, 2018.*

[16] Breiman, L, *Random Forests.* Machine Learning, **45**(1), 2001, P.5-32. Doi:10.1023/A:1010933404324

[17] Brown, I., Mues, C., *An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets*, Expert Systems with Applications, 2012, **39**(3), P.3446-3453. Doi: org/10.1016/j.eswa.2011.09.033

[18] Buda, M, *A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. Royal Institute of Technology*, School of Computer Science and Communication, Sweden, 2017, Doi:10.1016/j.neunet.2018.07.011

[19] Campa, D., Camacho, M., *hhe Ipp att of S''' s pre-bankruptcy Financial Distress on Earnings Management Tools*. International Review of Financial Analysis, 2015, **42**, P.222-234. Doi:10.1016/j.irfa.2015.07.004

[20] Chaudhuri, A., De, K., *Fuzzy Support Vector Machine for Bankruptcy Prediction*. Applied Soft Computing, **11**, 2011, P. 2472–2486. Doi:10.1016/j.asoc.2010.10.003.

[21] Chawlaet, N.V., *Data Mining for Imbalanced Datasets: An Overview, 2005.*

[22] Chen, Y.S, *an empirical study of a hybrid imbalanced-class DT-RST classification procedure to elucidate therapeutic effects in uraemia patients*, 2016, **54**(6), P.983–1001. Doi:10.1007/s11517-016-1482-0

[23] Chuang, C.L., *Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction*, Information Sciences, 2013, **236**, P.174–185. Doi:10.1016/j.ins.2013.02.015

[24] Danenas, P., Garsva, G., *Selection of support vector machines based classifiers for credit risk domain*. Expert Syst. Appl, 2015, **42**(6), P.3194–3204. Doi:10.1016/j.eswa.2014.12.001

[25] D'aveni, R. A, *The aftermath of organizational decline: A Longitudinal Study of the Strategic and Managerial Characteristics of Declining Firms*. Academy of Management Journal, 1989, **32**(3), P.577-605. Doi:10.5465/256435

[26] De Andres, J., Landajo, M., Lorca, P., *Bankruptcy prediction models based on multinorm analysis: An alternative to accounting ratios*. Knowledge-Based Systems, 2012, **30**, P.67-77. Doi:10.1016/j.knosys.2011.11.005

[27] Desman, J., *Statistical Comparisons of Classifiers over Multiple Data Sets*, Journal of Machine Learning Research, 2006, *7*, P.1–30.

[28] Desai, V. S., Crook, J. N., Overstreet, G. A. Jr., *A comparison of neural networks and linear scoring models in the credit union environmen*t, European Journal of Operational Research, 1996, **95**(1), P.24–37. Doi:10.1016/0377-2217(95)00246-4

[29] Dilsha, M., Kiruthika, A., *Neural network approach for microfinance credit scoring*, J. Stat. Manag. Syst, 2015, **18**(1–2), P.121–138. Doi:10.1080/09720510.2014.961767

[30] Ding, Y., Song, X., Zen, Y., *Forecasting Financial Condition of Chinese Listed Companies Based On Support Vector Machine*, Expert Systems with Applications, 2008, P.3081–3089. Doi:10.1016/j.eswa.2007.06.037

[31] Fallahpoor, S., Eram, A., *Predicting companies' financial distress using ant colony algorithm*, Journal of Financial research, 2016, **18**(2), P.347-368 (in Persian).

[32] Fitzpartrick, P.J., *A comparison of ratios of successful industrial enterprises with those of failed firms, Certif*. Publ. Account, 1932, **10**, P.598–605, **11**, P.656–662; **12**, P.727–731.

[33] Friedman, M, *A Comparison of Alternative Tests of Significance for the Problem of Rankings*, Annals of Mathematical Statistics, 1940, **11**(1), P. 86–92.

[34] Gepp, A., Kumar, K., Bhattacharya, S., *Business Failure Prediction Using Decision Trees*, Journal of Forecasting, 2010, **29**(6), P.536-555 Doi:10.1002/for.1153.

[35] Ghasemi, S., Sarlak, A., *Investigating the Impact of the Financial Crisis on Conservative Accounting and Transparency of Banking Information*, Advances in Mathematical Finance and Applications, **3**(3), 2018, P.53-68 Doi: 10.22034/AMFA.2018.544949

[36] Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G. (2008). On the Class Imbalance Problem. Fourth International Conference on Natural Computation (pp. 192-201). Jinan, China: IEEE. Doi: 10.1109/ICNC.2008.871

[37] He, H., Garcia, E. A., *Learning from Imbalanced Data*, IEEE Transactions on Knowledge and Data Engineering, **21**(9), 2009, P.1263-1284. Doi: 10.1109/TKDE.2008.239

[38] Henley, W. E., Hand, D. J., *Construction of a k-nearest Neighbour Credit Scoring System*, IMA Journal of Management Mathematics, 1997, **8**(4), P.305–321. Doi: 10.1093/imaman/8.4.305

[39] Hsu, C.W., Chang, C.C., Lin, C.J., *A Practical Guide to Support Vector Classification. Technical Report*, Department of Computer Science and Information Engineering, National Taiwan University, 2004.

[40] Huang, X.B., Liu, X.L., Ren, Y.Q., *Enterprise credit risk evaluation based on neural network algorithm*, Cogn. Syst. Res. 2018, P.52 317–324. Doi:10.1016/j.cogsys.2018.07.023

[41] Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., Wu, S., *Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study*, Decision support systems, 2004, **37**(4), P.543-558. Doi:10.1016/S0167-9236 (03)00086-1

[42] Iturriaga, F. J. L., Sanz, I. P., *Bankruptcy Visualization and Prediction Using Neural Networks: A Study of US Commercial Banks*, Expert Systems with Applications, 2015, **42**(6), P.2857-2869. Doi:10.1016/j.eswa.2014.11.025

[43] Jo, H., Han, I., Lee, H., *Bankruptcy Prediction Using Case-Based Reasoning, Neural Networks, and Discriminant Analysis*, Expert Systems with Applications, 1997, **13**, P.97–108. Doi: 10.1016/S0957-4174 (97)00011-0

[44] Kasabov., *Evolving Connectionist Systems for Adaptive Learning and Knowledge Discovery: Trends and Directions*. Knowledge. -Based Syst. 2015, **80**, P.24–33. Doi:10.1016/j.knosys.2014.12.032

[45] Khashman, A., *Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes*, Expert Systems with Applications, 2010, **37**, P.6233–6239. Doi:10.1016/j.eswa.2010.02.101

[46] Kim, M. J., Han, I., *The Discovery of Experts' Decision Rules from Qualitative Bankruptcy Data Using Genetic Algorithms*, Expert Systems with Applications, 2003, **25**(4), P.637-646. Doi: 10.1016/S0957-4174(03)00102-7

[47] Kim, M. J., Kang, D. K., Kim, H.B, *Geometric Mean Based Boosting Algorithm with Over-Sampling to Resolve Data Imbalance Problem for Bankruptcy Prediction*, Expert Systems with Applications, 2015, **42**(3), P.1074-1082. Doi:10.1016/j.eswa.2014.08.025

[48] Kim, S.Y., P*rediction of Hotel Bankruptcy Using Support Vector Machine, Artificial Neural Network, Logistic Regression, and Multivariate Discriminant Analysis*, The Service Industries Journal, 2011, **31** (3), P.441-468. Doi:10.1080/02642060802712848

[49] Kim, T., Ahn, H, *A., Hybrid Under-Sampling Approach for Better Bankruptcy Prediction*, Journal of Intelligence and Information Systems, 2015, **21**(2), P.173-190.

[50] Kotsiantis, S., Kanellopoulos, D., Pintelas, P., *Handling Imbalanced Datasets*: *A Review*. GESTS International Transactions on Computer Science and Engineering, 30, 2006.

[51] Kumar, P. R., Ravi, V., *Bankruptcy Prediction in Banks and Firms via Statistical and Intelligent Techniques: A Review*, European Journal of Operational Research, 2007, **180**, P.1–28. Doi:10.1016/j.ejor.2006.08.043

[52] Lane, P. C., Clarke, D., Hender, P., *On Developing Robust Models for Favorability Analysis: Model Choice, Feature Sets and Imbalanced Data*, Decision Support Systems, 2012, **53**(4), P.712-718. Doi:10.1016/j.dss.2012.05.028

[53] Lee, T., Chiu, C., Chou, Y., Lu, C, M*ining the Customer Credit Using Classification and Regression Tree and Multivariate Adaptive Regression Splines*, Computational Statistics and Data Analysis, 2006, **50**, P.1113–1130. Doi:10.1016/j.csda.2004.11.006

[54] Lessmann, S., Baesens, B., Mues, C., Pietsch, S., *Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings*, IEEE Transactions on Software Engineering, 2008, **34**(4), P.485–496**.** Doi: 10.1109/TSE.2008.35

[55] Li, H., Sun, J., *Ranking-order Case-Based Reasoning for Financial Distress Prediction*, Knowledge-based Systems, 2008, **21**, P.868–878. Doi:10.1016/j.knosys.2008.03.047

[56] Li, H., Sun, J., *Forecasting Business Failure: The Use of Nearest- Neighbor Support Vectors and Correcting Imbalanced Samples - Evidence from the Chinese Hotel Industry*, Tourism Management, 2012, **33**, P.622–634. Doi:10.1016/j.tourman.2011.07.004

[57] Li, X., Wang, F., Chen, X., *Support vector machine ensemble based on choquet integral for financial distress prediction,* Int. J. Pattern Recognit. Artif. Intell, 2015, **29**(4), P.1–16. Doi:10.1142/S0218001415500160

[58] Liao, J.-J., Shih, C.-H., Chen, T.-F., Hsu, M. F., An Ensemble-Based Model for Two-Class Imbalanced Financial Problem, Economic Modeling, 2014. Doi:10.1016/j.econmod.2013.11.013

[59] Lin, S.W., Ying, K.C., Chen, S.C., Lee Z.J., *Particle swarm optimization for parameter determination and feature selection of support vector machines*, Expert Systems with Applications, 2008, **35**, P.1817-1824. Doi:10.1016/j.eswa.2007.08.088

[60] Lin, W. Y., Hu, Y. H., Tsai, C. F., *Machine Learning in Financial Crisis Prediction: A Survey*, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2012, **42**, P.421–436. Doi**:** 10.1109/TSMCC.2011.2170420

[61] Lopez, V., Fernandez, A., Garcia, S., Palade, V., Herrera, F., *An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics*., Information Sciences, 2013, **250**, P.113-14. Doi:10.1016/j.ins.2013.07.007

[62] McKee, T. E., Greenstein, M., *Predicting Bankruptcy Using Recursive Partitioning and a Realistically Proportioned Data Set*, Journal of Forecasting, 2000, **19**(3), P.219-230. Doi:10.1002/(SICI) 1099-131X(200004)19

[63] Messier, Jr., W., Hansen, J., *Inducing Rules for Expert System Development: An Example Using Default and Bankruptcy Data*, Management Science, 1988, **34**(12), P.1403–1415. Doi:10.1287/mnsc.34.12.1403

[64] Min, J., Lee, Y., *Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters*, Expert Systems with Applications, 2005, **28**, P.603-614. Doi:10.1016/j.eswa.2004.12.008

[65] Min, S.-H., Lee, J., Han, I., *Hybrid Genetic Algorithms and Support Vector Machines for Bankruptcy Prediction*, Expert Systems with Applications, 2006, **31**, P.652–660. Doi:10.1016/j.eswa.2005.09.070

[66] N. Sung, T. K., Chang, N., Lee, G., *Dynamics of Modeling in Data Mining: Interpretive Approach to Bankruptcy Prediction*. Journal of Management Information Systems, 1999, 16, P.63–85. Doi:10.1080/07421222.1999.11518234

[67] Namazi, M., Kazemnezhad, M., Nematelahi, M., *Comparing Different Feature Selection Methods in Financial Distress Prediction of the Firms Listed in Tehran Stock Exchange*, Journal of Financial Engineering and Securities Management, 2016, **29**(7), P.193-212 (in Persian).

[68] Nemenyi, P. B, *Distribution-free Multiple Comparisons*, Ph.D. Thesis. Princeton University, 1963.

[69] Odom, M., Sharda, R., A neural networks model for bankruptcy prediction, in: Proceedings of the IEEE International Conference on Neural Network, 1990, **2**, P.163-168. Doi: 10.1109/IJCNN.1990.137710

[70] Ohlson, J.A., *Financial ratios and probabilistic prediction of bankruptcy*, J. Account. Res. 1980, **18**(1), P.109–131. Doi: 10.2307/2490395

[71] Ooghe, H., Joos, P., *Failure Prediction, Explanation of Misclassifications and Incorporation of Other Relevant Variables: Result of Empirical Research in Belgium*, Working paper, Department of Corporate Finance, Ghent University (Belgium), 1990.

[72] Pal, R., Kupka, K., Aneja, A.P., *Business health characterization: A hybrid regression and support vector machine analysis*, Expert Syst. Appl, 2016, **49**, P.48–59. Doi:10.1016/j.eswa.2015.11.027

[73] Piri, S., Delen, D., Liu, T., *A Synthetic Informative Minority Over-Sampling (SIMO) Algorithm Leveraging Support Vector Machine to Enhance Learning from Imbalanced Datasets*. Decision Support Systems, 2018, **106**, P.15-29. Doi:10.1016/j.dss.2017.11.006

[74] Rezaei, F., Tolaminejad, B., *The Financial Applications of the Colony Ant Algorithm*, Accounting and Auditing Studies, 2012, **3**(1), P.48-59 (in Persian).

[75] Rezaei, N., Javaheri, M., *The Predictability of Neural Network and nnn ccccc ggrhhm rrom Copp an'''' Financial Crisis*, Advances in Mathematical Finance and Applications, 2020, **5**(2), P.183-196 (in Persian). Doi: 10.22034/AMFA.2019.1863963.1195

[76] Rosner, R. L., *Earnings Manipulation in Failing Firms*, Contemporary Accounting Research, *2003*, **20**(2), P.361-408. Doi:10.1506/8EVN-9KRB-3AE4-EE81

[77] Sartre, F., Mazzucchelli, A., Gregorio, A. D., *Bankruptcy Forecasting Using Case-Based Reasoning: The Creeperie Approach*, Expert Systems with Applications, 2016, P.64, 400–411. Doi:10.1016/j.eswa.2016.07.033

[78] Saruei, S., *The Study of Performance of Springerit, Zimsky and Ahlson Models in Predicting Bankruptcy of Listed Companies in Tehran Stock Exchange*, M. A. thesis, Arak Islamic Azad University, Arak, Iran, 2010, (in Persian).

[79] Setayesh, M., Kazemnezhad, M., Hallaj, M., *The Usefulness of Random Forest Classifier and Relief Features Selection in Financial Distress Prediction: Empirical Evidence of Companies Listed on Tehran Stock Exchange*, Journal of Financial Accounting Research, 2016, **28**(8), P.1-24 (in Persian).

[80] Sun, J., Lang, J., Fujita, H., Li, H., *Imbalanced Enterprise Credit Evaluation with DTE-SBD: Decision Tree Ensemble Based on SMOTE*. Information Sciences, 2018, **425**, P.76–91. Doi:10.1016/j.ins.2017.10.017

[81] Tay, F. E., Cao, L., *Application of Support Vector Machines in Financial Time Series Forecasting*, Omega, **29**(4), 2001, P.309-317. Doi:10.1016/S0305- 0483(01)00026-3

[82] Thabtah F., *Machine Learning in Autistic Spectrum Disorder Behavioral Research*: A *Review and Ways Forward Informatics for Health and Social Care*, 2018, b, **43**(2), P.1-20. Doi:10.1080/17538157.2017.1399132

[83] Thabtah, F., Kamalov, F., Rajab, K., *A New Computational Intelligence Approach to Detect Autistic Features for Autism Screening*, International Journal of Medical Infromatics, 2018, *117*, P.112-124. Doi:10.1016/j.ijmedinf.2018.06.009

[84] Tian, S., Yu, Y., Zhou, M., *Data Sample Selection Issues for Bankruptcy Prediction*, Risk, Hazards and Crisis in Public Policy, 2015, **6**(1), P.91-116. Doi:10.1002/rhc3.12071

[85] Vapnik, V, *Statistical learning theory*, Wiley, New York, 1998. Doi: 10.1109/72.788640

[86] Wald, A., *On Statistical Problem Arising in the Classification of an Individual into One of Two Groups*, Annals of Mathematical Statistics, 1994, **15**(2), P.145-162.

[87] Wang, M., Chen, H., Li, H., Cai, Z., Zhao, X., Tong, C., Li, J., Xu, X., *Grey Wolf Optimization Evolving Kernel Extreme Learning Machine: Application to Bankruptcy Prediction*, Engineering Applications of Artifcial Intelligence, 2017, **63**, P.54 – 68. Doi:10.1016/j.engappai.2017.05.003

[88] Weiss, G. M., Provost, F. J., *Learning when training data are costly: The effect of class distribution on tree induction*, Journal of Artificial Intelligence Research, 2003, **19**, P.315–354. Doi:10.1613/jair.1199

[89] Weiss, G.M., *Mining with Rarity: A Unifying Framework*, SIGKDD Explor, 2004, **6**(1), P.1–7. Doi:10.1145/1007730.1007734

[90] Wilson, R. L., Sharda, R., *Bankruptcy Prediction Using Neural Networks*, Decision Support Systems, 1994, **11**(5), P.545- 557. Doi:10.1016/0167-9236 (94)90024-8

[91] Xia, Y., Liu, C., Li, Y.Y., et al, *a boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring*, Expert Syst Appl, 2017, **78**, P.225–241. Doi:10.1016/j.eswa.2017.02.017

[92] Yeh, I. C., Lien., C. H., *The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients*, Expert Systems with Applications, 2009, **36**, P.2473–2480. Doi:10.1016/j.eswa.2007.12.020

[93] Zhou, L., Lai, K. K., Yen, J., *Bankruptcy Prediction Using SVM Models with a New Approach to Combine Features Selection and Parameter Optimization*., International Journal of Systems Science, 2014, **45**(3), P.241-253. Doi:10.1080/00207721.2012.720293

[94] Zmijewski, M. E., *Methodological Issues Related to the Estimation of Financial Distress Prediction Models*, Journal of Accounting Research, 1984, **22**, P.59–82. Doi:10.2307/2490859