Int. J. Nonlinear Anal. Appl. 15 (2024) 1, 199–214
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.29040.4049

Detecting financial fraud using machine learning techniques

Jafar Nahri Aghdam Ghalejoogh[a], Nader Rezaei[a,], Yaghoub Aghdam Mazrae[b], Rasoul Abdi[a]

[a]Department of Accounting, Bonab Branch, Islamic Azad University, Bonab, Iran

[b]Department of Accounting, Sofian Branch, Islamic Azad University, Sofian, Iran

(Communicated by Seyyed Mohammad Reza Hashemi)

Abstract

Financial fraud detection is a challenging problem due to four primary reasons: the constantly changing fraudulent behavior, the lack of a mechanism to track fraud data, the specific limitations of available detection techniques (such as machine learning algorithms), and the highly dispersed financial fraud dataset. Thus, it can be declared that teaching algorithms are complex. The current study used machine learning techniques, including support vector machine regression and boosted regression tree, to detect financial fraud in the Iranian stock market. The findings indicated that the boosted regression tree machine model has the lowest RMSE. Furthermore, concerned with the sensitivity value of the models, the boosted regression tree model has the highest sensitivity in the sense that they had correctly detected the absence of financial fraud Tehran Stock Exchange market the Tehran Stock Exchange market. The boosted regression tree has the highest kappa coefficient indicating the appropriate performance of this model compared to other models used in the research.

Keywords: Support vector machine regression, Boosted regression tree, Financial fraud
2020 MSC: 91G15, 62G08

1 Introduction

Accounting is a service activity that provides helpful information about economic units to internal and external decision-makers. Thus, accounting should identify, analyze, and appropriately record the activities of an economic unit in order to report effectively to stakeholders. At present, according to the increasing demand of managers to obtain correct financial data in order to make managerial decisions in the long term and the need to attract domestic and foreign investors to provide capital and compete on this issue, the information provided about the financial status and performance of a company, is essential for users of financial statements. Because as it was mentioned above, financial statements are necessary to evaluate the company's performance and financial health and monitor management; it is the basis for decision-making and capital allocation. In addition, it provides valuable financial information for efficient resource allocation and conscious economic decision-making.

Therefore, financial reports should be presented correctly, accurately, and well-timed to provide helpful information about organizations' financial and non-financial performance to internal and external users. This will lead to the correct and honest presentation of the company's financial achievements and play an essential role in the financial system's stability.

Email addresses: jafarnahri@gmail.com (Jafar Nahri Aghdam Ghalejoogh), naderrezaeimiyandoab@gmail.com (Nader Rezaei), aghdam.acc@gmail.com (Yaghoub Aghdam Mazrae), abdi_rasool@yahoo.com (Rasoul Abdi)

Received: April 2022 *Accepted:* July 2022

Correct financial information is the most fundamental component in making correct managerial and investment decisions. Any intentional or accidental error in the financial information will bring about non-optimal decisions in shareholding and corporate affairs and will seriously endanger the interests of the companies' stakeholders. Thus, financial reports should be presented correctly, accurately and well-timed to actual and potential beneficiaries. Nevertheless, it is worth mentioning that fraud in financial reports has been increasing in recent years. Fraud in financial reporting is defined as the intentional distortion of the results of financial statements to present a false image of the company, like exaggerating assets, minimizing costs, and understating stolen assets. Deception is done through documentation and manipulation or changing accounting records or supporting documents for the preparation of financial statements and incorrect presentation or intentional omission of events, transactions, or other vital information in financial statements.

Higson [10] analyzed the results of 13 interviews with senior auditors and certified public accountants on whether their clients reported possible fraud to external users. Although some companies reported cases, some were cautious in this regard.

Chen et al. [7] used distributed data mining to detect fraud. Their research showed the possibility of detecting credit card fraud for risk management.

Feroz [9] used a sample of 42 incorrect reports and 90 actual financial reports. The test was done based on logistic regression and neural networks. The experimental results demonstrated that the neural network method is better than the logistic regression method.

Abbott et al. [1] checked and measured auditing independence and activity in reducing the risk of fraud. Using logistic regression analysis, they found that companies with dormant audit committee boarder of directors with at least two annual sessions are less inclined to false or delusive reporting.

Bell and Carcello [3] used the logistic regression method to check fraud detection. They found that prevarication of the manager to the auditor, weak internal control system, aggressive behavior of the manager, the undue emphasis of the manager to achieve the predicted profit, and considerable problems in transaction auditing are among the factors which are different between fraudulent and non- fraudulent managers.

Beasley et al. [4] investigated fraudulent and non- fraudulent companies from 1980 to 1990. Their study was conducted in the three technology, treatment, health, and financial services departments. Fraud was different in different industries. In the technology companies, the fraud was mainly in income, while assets fraud and misallocation of the budgets in the financial services department were more common.

Their research showed that weak control was more common in fraud and compared non-fraudulent companies in the three industries. Furthermore, the results supported the previous research in fewer independent audit committee members in fraudulent companies. In addition, Bisley et al. understood that fewer audit committee sessions were held in fraudulent companies' health and technology departments and that fraudulent companies had weaker internal control. However, the experience of significant fraud worldwide shows that accounting and auditing encounter severe challenges in creating financial transparency and presenting reliable information. Some managers resort to fraud and manipulation of financial reports to achieve their goals and attract investors. This shows that all auditing and financial experts and the related fields of study should be equipped with more education knowledge than before, the knowledge which can make them more familiar with fraud and manipulation methods in financial reports and provide the possibility to discover and predict financial fraud. In this regard, machine learning techniques in detecting financial fraud can be essential.

2 Material and method

The theoretical literature was collected through desk research (including foreign and domestic books and journals and scientific treatises) and searching through scientific databases. It attempted to present the concepts and fundamental premise as exhaustive and abridged as possible.

The required statistical data were provided through the accepted company's financial statements, the published financial data on the Codal site, the Rahavard Novin software, the Central Bank, and the set forth. Thus, the research instruments were documents in the current study. In general, the present research will be accomplished in two stages. In the first stage, the library resources were used to prepare the research's theoretical literature; in the second stage, the authenticated and confirmed statistical sources were used to collect the desired data.

In this study, Excel software was used to prepare the data: extracting the data of the variables from the mentioned resources, they were entered in the worksheets of the software, and the necessary calculations were performed to access

the new variables.

Finally, the calculated variables were transferred to MATLAB software to implement the machine learning techniques.

2.1 Statistical sample and sampling method

According to the spatial scope of the study, the statistical population includes all the listed companies in Tehran Stock Exchange whose stock will be exchanged on the floor or the first market from 2009 to 2020. However, the statistical sample population had the following criteria, according to the spatial and temporal scope of the research:

1. They were listed in Tehran Stock Exchange at least since the beginning of the fiscal year 2009.
2. The sample companies were not among the investment banking companies.
3. The sample companies were not stopped during the years 2008 to 2017 in order to regard the stock price ordinary.
4. The companies which 20 of March (The last day of a solar year) is the end of their fiscal year.
5. The sample companies whose fiscal year was not changed from 2009 to 2021.
6. The critical research data were provided for the Stock Exchange until the end of the fiscal year of 2021.

The machine learning techniques approach was used to detect financial fraud to achieve the research objectives. Machine learning is a branch of computer science that has become a hot topic in recent years to find the standards and generalize them to the future. Machine learning aims to give the previous data acquired power. In other words, the data in a particular algorithm learn to adapt themselves under different circumstances and develop themselves as a complete identity.

One of the advantages of machine learning is implementation without heavy programming. Many recent attempts have been made to use this knowledge in predicting financial-accounting variables. The machine learning techniques do not have a fixed classification since diverse, new, and combined models are constantly presented. They can be classified differently. However, the placing techniques in different categories depends on the different perspectives through which we see learning. These techniques use inductive learning and a set of data to make an Approximation of Productivity. Machine learning can be classified into three general categories:

1. Supervised learning
2. Unsupervised learning
3. Semi-supervised learning

A brief explanation will be provided for each of them in the following:

1. Supervised learning is a type of machine learning in which the input and output are specified, and there is no supervisor to provide the data for the learner. Accordingly, the system attempts to acquire a function of the input to the output.
2. In unsupervised learning, contrary to supervised learning, there are no specified data in advance, and the purpose is not the relationship between the input and output. However, classification is essential, and the learner should search for a specific structure in the data.
3. Semi-supervised learning is a type of learning that uses classified (labelled) and unclassified (unlabeled) data simultaneously to improve learning accuracy.

Since the present study intends to predict and detect financial fraud and provides specified data for the techniques, supervised learning was used. Supervised learning includes diverse techniques, the most important of which are:

1. Bayesian linear regression
2. Logit regression
3. Boosted regression tree
4. Neural network regression,
5. Support vector regression,

The data for the listed companies in the stock and OTC market from 2008 to 2021 will be used to implement the regression mentioned above. In general, according to the experimental and theoretical studies, the following model

will be used to investigate the factors affecting fraud in the financial reporting of companies in each machine learning technique:

$$\begin{aligned} F_{it} =& f(OCFTNI_{it}OCFTSAL_{it}, NITSAL_it, NITTA_{it}, NITEQ_{it}, TDTEQ_{it} \\ & TLTTA_{it}, CATCL_{it}, CASHTTA_{it}, INVTTA_{it}, ARTTA_{it}, CATTA_{it}CCC_{it} \\ & , INVTSAL_{it}, ARTSAL_{it}, COGTSAL_{it}, SALTTA_{it}, APTSAL_{it}, COGTINV_{it} \\ & , Board_{it}, ESTQ_{it}, NESTQ_{it}, TAX_{it}, UNEM_{it}, GS_{it}, INF_{it}, \log(EXC_{it}) \\ & , \log(GDP_{it}), \log(GDPP_{it}), MTGDP_{it}, e_{it}. \end{aligned}$$

In which: FF_{it}: is a virtual variable. This variable is assigned a value of zero for the years when the company's financial reports are accepted and is assigned the value of one, otherwise (conditional, rejected, and no comment);

$OCFTNI_{it}$: Ratio of operating cash flow to net income for i-th Company for the year t;

$OCFTSAL_{it}$: Operating cash flow to sales ratio for i-th Company for the year t;

$NITSAL_{it}$: Net income to sales ratio for i-th Company for the year t;

$NITTA_{it}$: Return on assets for i-th Company for the year t;

$NITEQ_{it}$: Return on equity for i-th Company for the year t;

$TDTEQ_{it}$: Total debt-to-equity for i-th Company for the year t;

$TLTTA_{it}$: Total debt to total assets for i-th Company for the year t;

$CATCL_{it}$: Current assets to current liabilities for i-th Company for the year t;

$CASHTTA_{it}$: Cash to total assets for i-th Company for the year t;

$INVTTA_{it}$: Inventory to total assets for i-th Company for the year t;

$ARTTA_{it}$: accounts receivable to total assets for i-th Company for the year t;

$CATTA_{it}$: Current assets to total assets for i-th Company for the year t;

CCC_{it}: Cash flow cycle for i-th Company for the year t;

$INVTSAL_{it}$: Cash balance to sales for i-th Company for the year t;

$ARTSAL_{it}$: Accounts receivable to sales for i-th Company for the year t;

$SOGTSAL_{it}$: Cost of goods sold to sales for i-th Company for the year t;

$SALTTA_{it}$: Asset turnover ratio for i-th company for the year t; (net sales divided by average total assets)

$APTSAL_{it}$: Accounts payable to sales for i-th company for the year t;

$COGTINV_{it}$: Inventory turnover for i-th Company for the year t; (cost of goods sold divided by average inventory)

$Boord_{it}$: The number of board members for i-th Company for the year t;

$ESTQ_{it}$: The number of board members required for i-th Company for the year t;

$NESTQ_{it}$: The number of dormant member of directors for i-th company for the year t;

TAX_{it}: The income tax rate in the whole economy for the year t;

$UNEM_{it}$: Unemployment rate for the year t;

GS_{it}: Ratio of government expenditure to Gross domestic product (GDP) for the year t;

OIL_{it}: Oil revenues for the year t;

INF_{it}: Inflation rate for the year t;

$log(EXC_{it})$: Natural logarithm of the exchange rate in the informal market for the year t;

$log(GDP_{it})$: Logarithm of GDP for the year t;

$log(GDPP_{it})$: Logarithm of GDP per capita for the year t;

$MTGDP_{it}$: Ratio of liquidity to GDP for the year t;

e_{it}: The residual of the regression model for i-th company in the year t.

2.2 Data analysis - machine learning methods

2.2.1 Boosted regression tree

The boosted regression tree model combines the results of many weak classifiers to build a robust classifier [2]. In the boosted regression tree method, decision trees are created using the classification and regression tree method (CART), one of the classification one can find in [12]. The boosted regression tree model is a non-parametric method that uses the classification and regression tree model to solve linear regression problems. The three ways to analyze the classification and regression tree model are to create full trees first, choose the best trees, and do the cross-validation process with pruning steps.

The boosted regression tree method has the capabilities of two algorithms, including regression trees, the models which describe the response to predictors through optimal and boosted binary (two-mode) separation. It is also an adaptive method to combine many simple models to obtain the appropriate performance [8]. The better performance of the algorithm depends on the correct adjustment of the options related to the boosted trees and the stopper parameters of the branching trees. The boosting operations should be performed to improve the predictive power of the regression tree. During this operation, the results of several models are used. The boosting operation is a step-forward process in which the repetition of models is fitted to a part of the training dataset at every step. Accordingly, two main parameters are proposed for the model.

1. The division rate that determines the percentage of training data in each iteration that the user determines;
2. The reduction rate, which expresses the contribution of each tree in the modelling process and the number of nodes in each tree.

The previous studies revealed that a reduced rate of 0.1 or less tends to be a suitable model; for more minor data (n=500), this rate can be set to 0.005, and for more extensive data (n=5000), it can be set to 0.5. Among the essential advantages of this method, the following can be mentioned: it can analyze large volumes of data at high speed, it is less sensitive to overlap than other classification models, it does not require the assumption of data distribution, and it can determine the most effective and essential factors in classification [5].

2.2.2 Artificial neural network regression

An artificial neural network is a neuron-like structure in the human brain. An artificial neural network is an information processing system that imitates the biological nerves and can receive and combine multiple inputs to make predictions. An artificial neural network is a type of artificial intelligence device in which a mathematical method is used to create the ability of a computer to conclude the fast computing power of a computer. This should be done through a learning process (i.e. machine learning) so that it can have the ability to infer (i.e. someone tells it which conditions will lead to which result). If you tell it the good examples, it will answer you correctly. An artificial neural network can predict the possible outcome for previously unlearned examples.

The units that are processed by an artificial neural network are neurons. They have two objectives:

1. Passage, in which incomplete data points for node inputs do not significantly influence the network.
2. Adaptive learning refers to adjusting the connection weight between nodes.

The main structure of an artificial neural network encompasses an input layer, a hidden layer, and an output layer. The output value of each processing element is transferred to another processing unit and becomes the input value of that unit [11]. The input of the neural network will be the specificity vectors, and the network's output will represent different classes. Multi-Layer Feed-Forward Networks are one of the most critical and common neural networks in real-world applications. The most common type of artificial neural network is created from a set of essential neurons that form an input layer, one or more hidden layers, and an output layer. The input data is propagated through the network in a forward and layer-wise direction.

This type of neural network is a pre-feeder or multi-layer perceptron (MLP). The number of neurons in the input layer is equal to the number of elements of the input vector, and the number of neurons in the output layer is equal to the number of elements of the output vector. However, precise and realistic analysis to find the number of neurons in the middle layer is very complex. It can be declared that the number of neurons in the hidden layer is a function of the number of input vector elements and also the maximum number of regions of the input space that are linearly separated from each other. Thus, the number of hidden layer neurons is generally obtained experimentally.

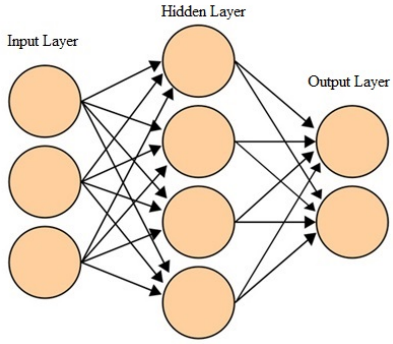

Figure 1: An example of an artificial neural network with a hidden layer

2.2.3 Logit regression

Logistic regression is a statistical method that allows researchers to create predictive models. This technique is used to understand the effect of several independent variables on a bivariate dependent variable. This method attempts to predict the dependent variable based on the independent variables.

Logistic regression predicts the rank of each sample company by assigning weights to independent variables. This rank determines the probability of membership in a given group (zero and one). The probability of success or failure (or any other binary quality ranking) in this model is calculated by the following cumulative distribution function:

$$p_i = E(Y = 1|X_i) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-a-\sum_{i=1}^{k} b_i x_i}}.$$

In the above relation, p_i is the probability that the dependent variable will be one, x_i is the independent variables and $b_i$is the coefficients of independent variables [13].

2.2.4 Evaluating the effectiveness of machine learning techniques

Confusion matrix

The confusion matrix is one of the most crucial and widely used tools to evaluate the performance of developed models in classification problems. Besides, this matrix is used in binary and multi-class classification problems.

Table 1: Confusion matrix

Original class	Expected class	
	Class 0	Class 1
Class 0	TP	TN
Class 1	FP	FN

The elements of the confusion matrix are defined as follows:

TP: correct prediction percentage of class 0

TN: correct prediction percentage of class 1

FP: false prediction percentage of class 0 as class 1

FN: false prediction percentage of class 1 as class 0

According to the elements of the confusion matrix, the criteria of accuracy, sensitivity, and specificity are used to check and evaluate the performance of machine learning techniques.

$$\texttt{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Specificity} = \frac{TN}{FP + TN}.$$

ROC curve

One of the suitable methods to evaluate the results of a classifier and its ability to identify the target class is to use the receiver operating characteristic curve or ROC curve to determine the method's sensitivity. Sensitivity is defined as the relationship between the amount of correctly classified cells and the incorrect ones.

The greater the deviation from the baseline for a particular class in the ROC curve, the higher the efficiency of the mentioned classifier in identifying that class. In addition to examining the Trend Curve of the target class, the area under the curve is also calculated. This area indicates the possibility that a randomly selected cell is classified correctly; the higher it is, the more reliable the mentioned method will be [6].

3 Results and discussion

The results and performance of the two famous machine learning models, including support vector machine (SVM) regression and boosted regression tree (BRT), were investigated to detect financial fraud in the Iranian stock market. In these models, the fraud variable (presence of fraud = 1 and absence of fraud = 0) was considered as the response variable (dependent), and 31 variables were considered as predictor variables (independent). The R statistical software was utilized to run these models. The obtained results unraveled differences in the influential predictor (independent) variables using each model.

The data used were the data of 125 companies active in the Tehran stock market during 12 years (from 2008 to 2019) and a total of 1500 observations. Among these observations, 845 cases (56.3%) had financial fraud, and 655 cases (43.7%) did not have financial fraud. Also, among the data, two observations related to the COGTINV variable and one observation related to the CCC variable were missing, replaced by the average of each corresponding variable. Among these 1500 observations, using set.seed (123), 1050 observations were considered train data, and the remaining 450 observations were considered test data randomly.

3.1 Boosted Regression Tree (BRT)

The BRT model was implemented in the present study with a learning rate of 0.005 and a bag fraction of 0.5 using R software and the dismo package. Firstly, the predictive deviance plot was drawn versus the number of trees in Figure 2-4. Afterwards, the holdout deviance plot was drawn versus the number of trees to determine the optimal number of trees used in this model (Figure 3-4). After running the model and according to Figure 3-4, 1900 trees were considered in this model.

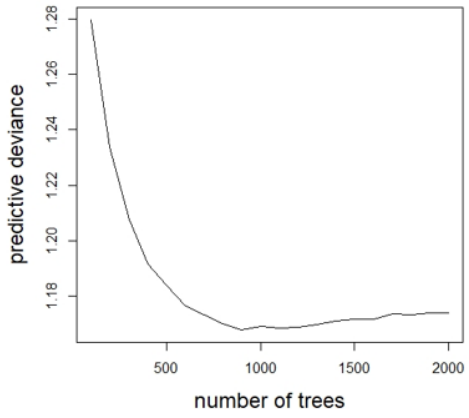

Figure 2: Predictive deviance plot versus the number of trees for BRT model

The functions fitted to each of the 31 predictor variables were initially drawn (Figures 4-4 to 4-6), and then the fitted values of these variables were calculated along with their weighted averages. The scatter plot of these fitted values is drawn versus the actual values of the variables in Figures 4-7 to 4-11.

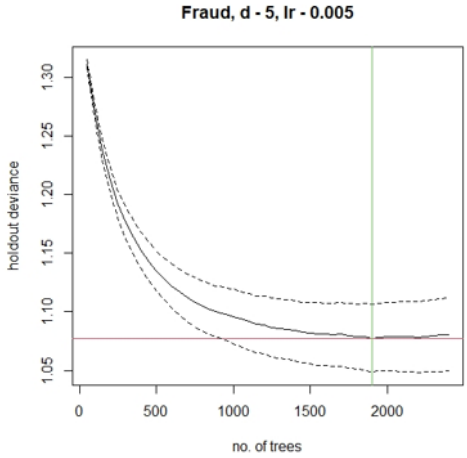

Figure 3: Holdout deviance plot versus the number of trees for the BRT model

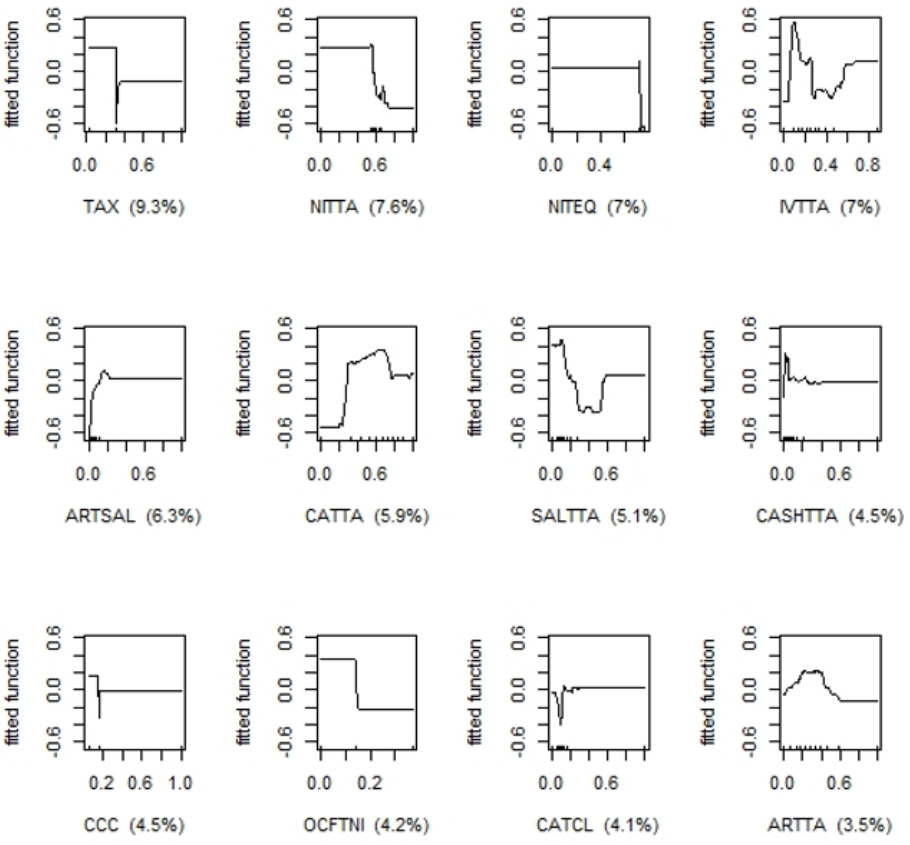

Figure 4: Functions fitted to predictor variables plot (first part)

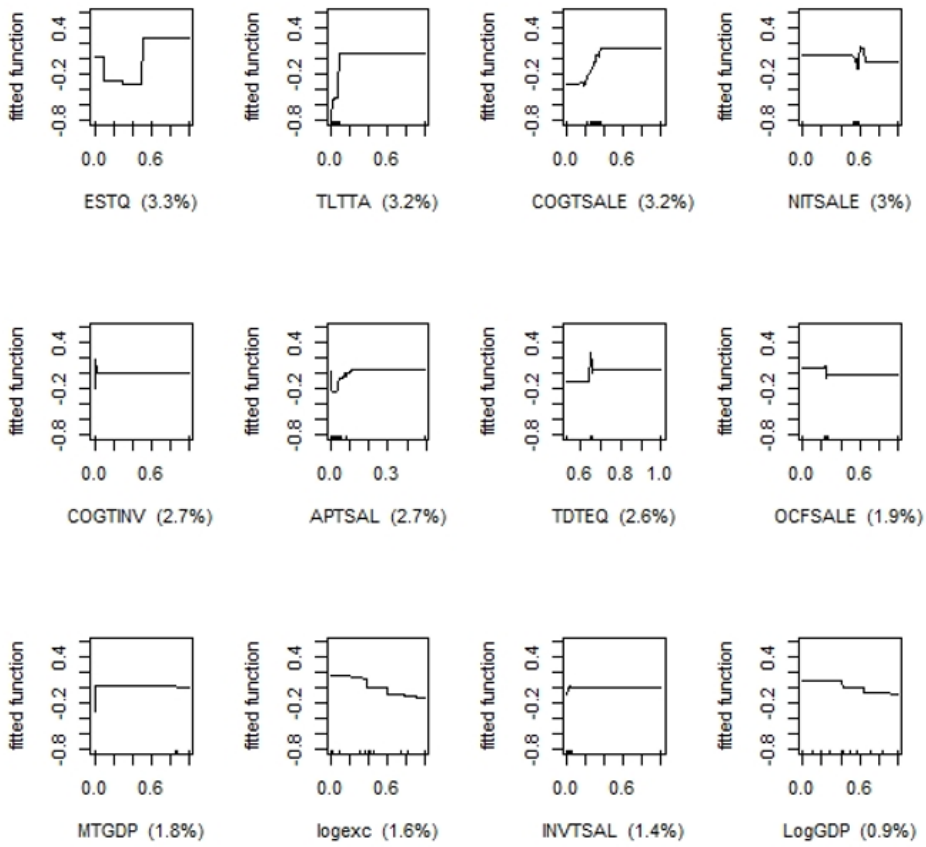

Figure 5: Functions fitted to predictor variables plot (second part)

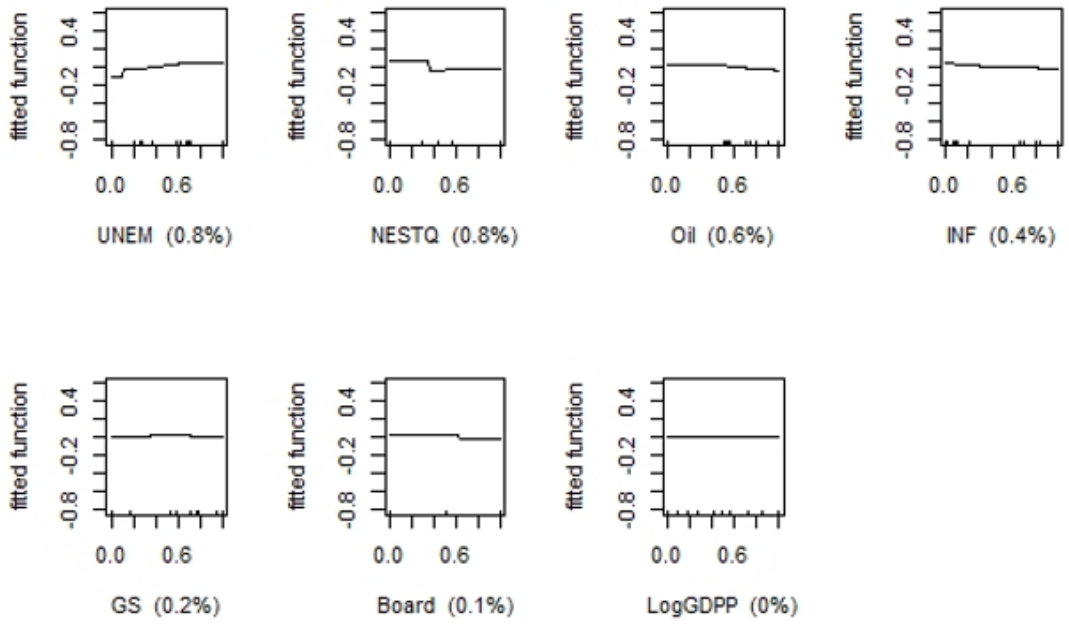

Figure 6: Functions fitted to predictor variables plot (third part)

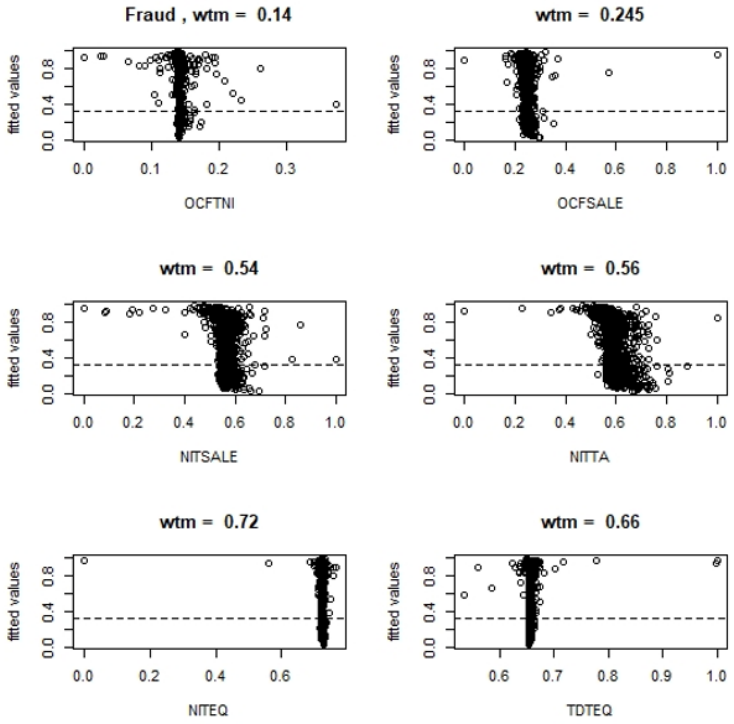

Figure 7: Scatter plot of the values fitted with the BRT model versus the actual values and the weighted average of the predictor variables (first part)

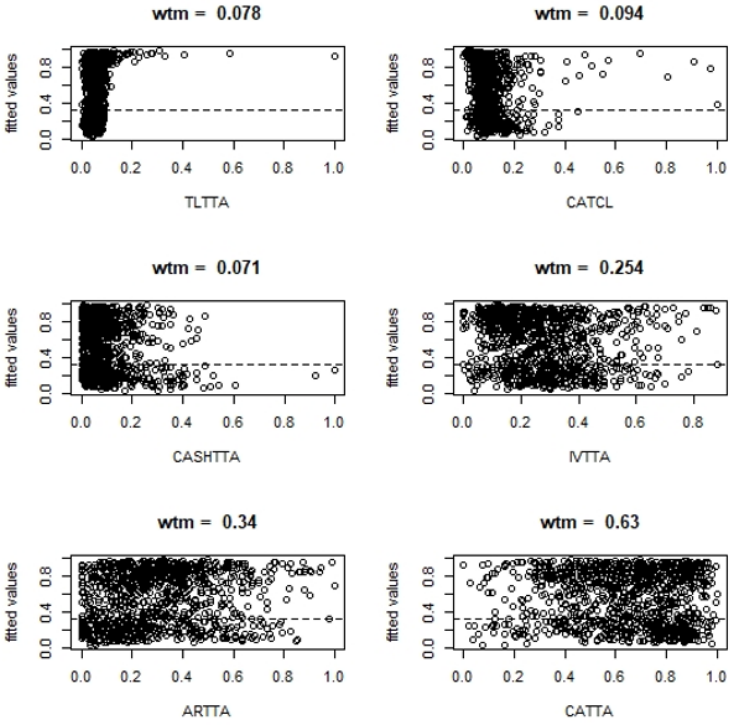

Figure 8: Scatter plot of the values fitted with the BRT model versus the actual values and the weighted average of the predictor variables (second part)

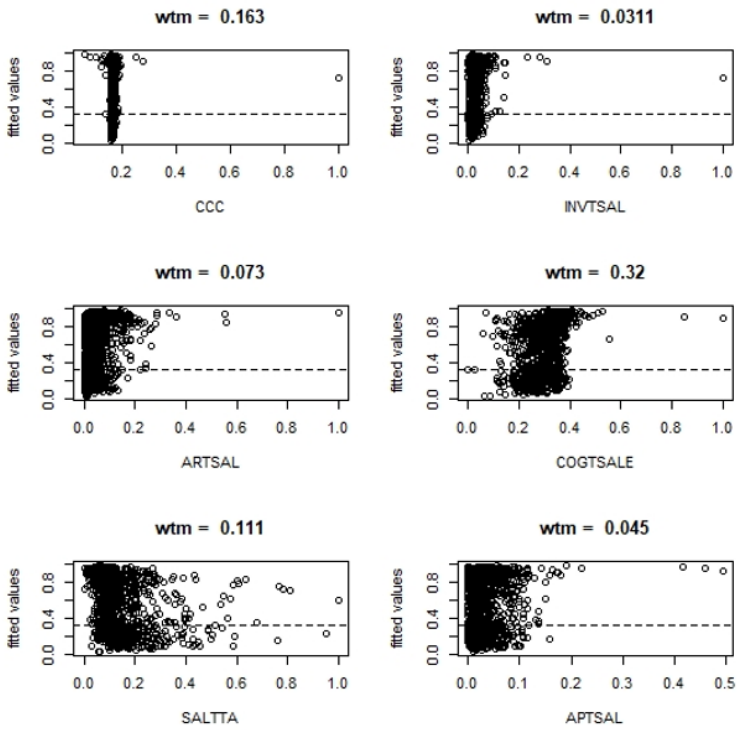

Figure 9: Scatter plot of the values fitted with the BRT model versus the actual values and the weighted average of the predictor variables (third part))

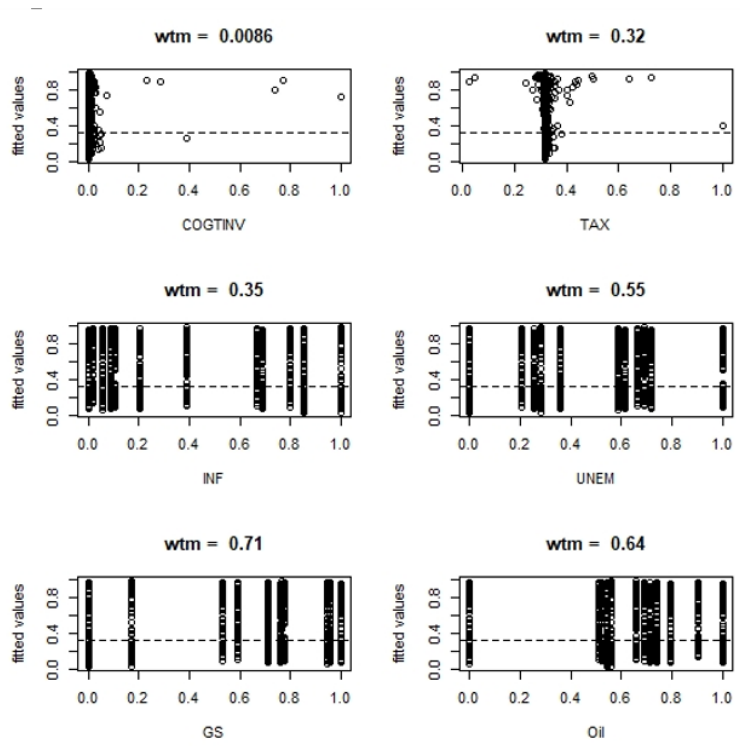

Figure 10: Scatter plot of the values fitted with the BRT model versus the actual values and the weighted average of the predictor variables (fourth part)

Firstly, the model was run with all 31 independent variables. The next step should determine which independent variables affected predicting financial fraud and which did not. To this aim, according to the changes in the pre-

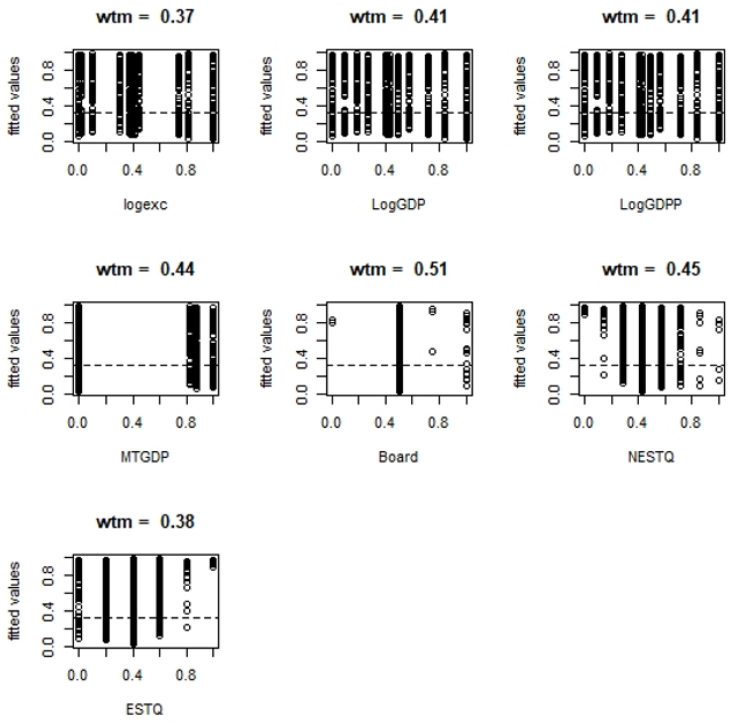

Figure 11: Scatter plot of the values fitted with the BRT model versus the actual values and the weighted average of the predictor variables (fifth part)

dictive deviance (Figure 12-4), 15 independent variables logGDPP, Board, GS, INF, LogGDP, Oil, NESTQ, UNEM, INVTSAL, logexc, TDTEQ, NITSALE, OCFSALE, ESTQ and APTSAL were removed, respectively.

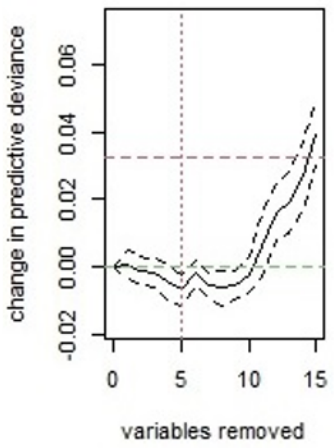

Figure 12: Changes in the predictive deviance versus independent variables removed in the BRT model

Then, the final model was run with the remaining 16 independent variables. To determine the importance of the remaining independent variables, the relative importance of these variables is demonstrated in Table 2 and drawn in Figure 12 for more intuition. According to these results, the three variables TAX, NITTA, and IVTTA, respectively, had the most influence on detecting financial fraud. Furthermore, the influence of COGTSALE, MTGDP, ARTTA, CCC, and COGTINV variables is almost half and, in some cases, even less than half of the influence of the three variables.

Afterwards, the model obtained from the train data was run on the test to check the goodness of fit for the BRT model. The root means square error (RMSE) value was calculated as 0.55104. Besides, the confusion matrix of the BRT model for the test data is reported in Table 3. Accordingly, the accuracy measure of this model was equal to

Table 2: Order of importance and relative importance of 16 remaining independent variables in the BRT model

Order	Variable Name	Relative Importance	Order	Variable Name	Relative Importance
1	TAX	9.663745	9	OCFTNI	5.801407
2	NITTA	8.581277	10	TLTTA	5.422749
3	IVTTA	8.365905	11	CATCL	5.394182
4	NITEQ	7.858480	12	COGTSALE	4.952860
5	ARTSAL	6.533000	13	MTGDP	4.910434
6	CATTA	6.319404	14	ARTTA	4.776843
7	SALTTA	6.287439	15	CCC	4.742899
8	CASHTTA	6.112588	16	COGTINV	4.276786

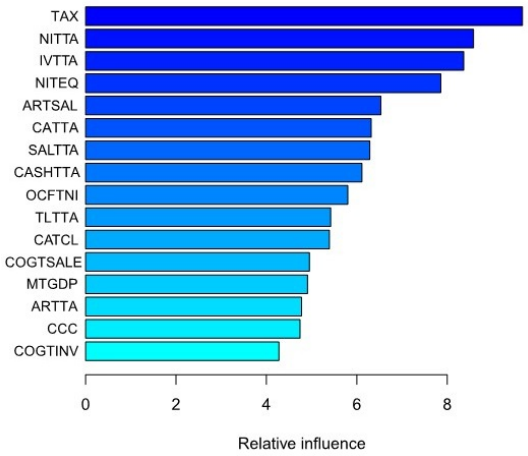

Figure 13: The relative importance of independent variables in predicting financial fraud in the BRT model

0.69111; the sensitivity and specificity values of this model were equal to 0.5736 and 0.7826, respectively.

Table 3: Confusion matrix of BRT model for test data

		Model prediction	
		No Fraud	Fraud
Reality	No Fraud	112(0,5685)	54(0,2134)
	Fraud	85(0,4315)	199(0,7866)

In addition, the receiver operating characteristic curve (ROC) is also drawn for this model in Figure 13. According to this plot, the area under the model's curve (AUC) value was obtained at 0.7563, which shows the acceptable accuracy of this model for data modelling.

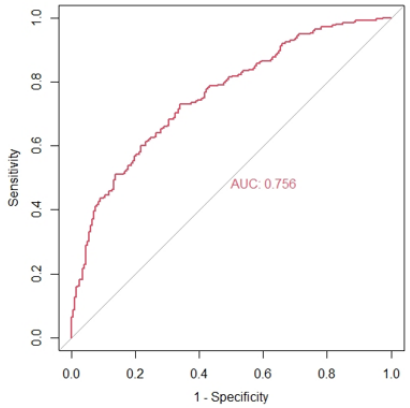

Figure 14: ROC curve based on test data for BRT model

A Kappa coefficient of 0.362 was obtained, which shows an almost average correlation between the results obtained

from this BRT model and the actual values of the test data. On the other hand, McNemar's non-parametric test rejected the (null) hypothesis of the independence of the values predicted by the model and the actual values (P-value=0.01755). Besides, the hypothesis, which states the measure of accuracy is smaller than the rate of ignorance (0.5622), was tested and rejected (P-value=1.394×10-8). Therefore, it is appropriate to use the BRT model for data modelling.

3.2 Support vector machines (SVM) model

In this study, two SVM models were run with radial and linear kernel functions and C-classification, respectively, using R software and e1071 package. Firstly, the SVM model was run with a linear kernel function, and the number of support vectors for this model was obtained as 764. To check the goodness of fit for this SVM model, the model obtained from the train data was run on the test data. The root means square error (RMSE) value was obtained as 0.58689. Also, the confusion matrix of this SVM model for test data is reported in Table 4. Accordingly, the accuracy measure of this model was equal to 0.65556; the sensitivity and specificity values of this model were equal to 0.5279 and 0.7549, respectively.

Table 4: Confusion matrix of SVM model with linear kernel function for test data

		Model prediction	
		No Fraud	Fraud
Reality	No Fraud	104(0,5279)	62(0,2451)
	Fraud	93(0,4721)	191(0,7549)

In addition, the ROC curve is also drawn for this model in Figure 15. Upon this plot, the AUC value of this model was obtained as 0.6414, which is less than 0.7. However, for the SVM model, the AUC measure is a weaker indicator than other indicators of the model's goodness of fit. It does not necessarily indicate that this model is inappropriate for data modelling. The Kappa coefficient obtained as 0.288 was obtained, which indicates a relatively weak correlation between the results obtained from this SVM model and the actual values of the test data.

On the other hand, McNemar's non-parametric test rejected the null hypothesis of the independence of the values predicted by the model and the actual values (P-value=0.01597). Furthermore, the hypothesis that the accuracy measure is smaller than the ignorance rate (0.5622) was tested and rejected (P-value=3.39×10-5). Therefore, the SVM model with a linear kernel is suitable for data modelling.

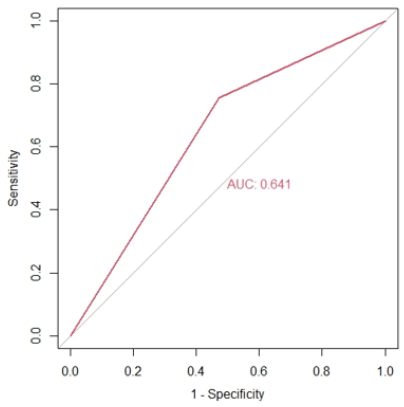

Figure 15: ROC curve based on test data for SVM model with linear kernel function

In the following, the SVM model was run with the radial kernel function, and the number of support vectors for this model was obtained as 817. To check the goodness of fit for this SVM model, the model obtained from the train data was run on the test data. The root means square error (RMSE) value was obtained as 0.54366. Besides, the confusion matrix of this SVM model for test data is reported in Table 5. Based on this, the accuracy measure of this model was equal to 0.704444; the sensitivity and specificity values of this model were equal to 0.5482 and 0.8261, respectively.

Furthermore, the ROC curve is also drawn for this model in Figure 16. Based on this plot, the AUC value of this model was obtained as 0.6872, which is less than 0.7. A Kappa coefficient equal to 0.384 was obtained, which indicates the almost average correlation between the results obtained from this SVM model and the actual values of the test

Table 5: Confusion matrix of SVM model with radial kernel function for test data

		Model prediction	
		No Fraud	Fraud
Reality	No Fraud	77(0,3909)	37(0,1462)
	Fraud	120(0,6091)	216(0,8538)

data. On the other hand, McNemar's non-parametric test rejected the null hypothesis of the independence of the values predicted by the model and the actual values (P-value=0.000136). Besides, the hypothesis that the accuracy measure is smaller than the ignorance rate (0.5622) was tested and rejected (P-value=3.878×10-10). Therefore, the SVM model with a linear kernel is suitable for data modelling.

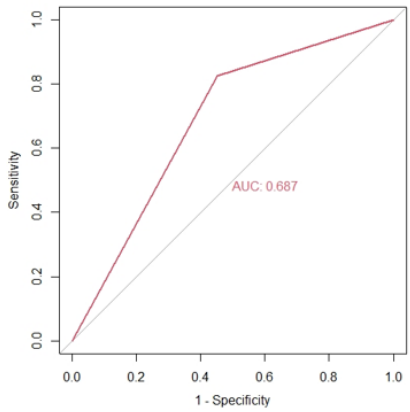

Figure 16: ROC curve based on test data for SVM model with radial kernel function

4 Conclusion

The results of implementing the logistic regression method in the fourth section showed that the root means square error (RMSE) was equal to 0.5696. Besides, using the confusion matrix, the accuracy measure of this model was equal to 0.67556; the sensitivity and specificity measures of this model were equal to 0.5635 and 0.7628, respectively. The area under curve (AUC) value of this model was obtained as 0.7166, which shows the acceptable accuracy of this model for data modelling. The Kappa coefficient results were obtained as 0.331, indicating an almost average correlation between the logistic regression model results and the test data's actual values. On the other hand, McNemar's non-parametric test also showed that the logistic regression model is appropriate for data modelling.

Logistic regression showed that among 31 independent variables, the following six variables had the most impact and had a significant effect on the financial fraud of the companies present in the Tehran Stock Exchange market: accounts receivable for sale, current assets to total assets, inventory to total assets, return on assets, operating cash flow to sales, accounts receivable to total assets, and the ratio of government expenditure to gross domestic product (GDP).

Also, increasing the error level to 0.10, the variables of net income to sales accounts payable to sales, return on equity, total debt to total assets, cash balance to sales, and cost of goods sold to sales had a significant influence on the financial fraud of companies present in the Tehran stock market. Other independent variables of the model did not significantly affect financial fraud. In addition, the variables of accounts receivable for sale, current assets to total assets, accounts payable to sales, total debt to total assets, cash balance to sales, and positive inventory turnover had a positive influence; in the sense that a large amount of these variables increases the possibility of financial fraud in the company. The influence of the variables of inventory to total assets, operating cash flow to sales, accounts receivable to total assets, the ratio of government expenditure to gross domestic product (GDP), net income to sales, and return on equity were negative; that is, the small amount of these variables reduces the possibility of financial fraud in the company.

The results of implementing the support vector machines method with linear kernel in the fourth section showed the root mean square error (RMSE) value as 0.58689. In addition, based on the confusion matrix, the accuracy measure of this model was equal to 0.65556; the sensitivity and specificity measures were equal to 0.5279 and 0.7549,

respectively. The area under the curve was found to be 0.6414 based on the ROC curve of this model; this shows the acceptable accuracy of this model for data modelling. (In the method of support vector machines, this index is a weaker index than other model goodness of fit indices and does not necessarily indicate that this model is unsuitable for data modelling.) The Kappa coefficient results were equal to 0.288, which indicates a relatively weak correlation between the results of the support vector machine model with a linear kernel and the actual values of the test data.

On the other hand, McNemar's non-parametric test also showed that using a support vector machine model with a linear kernel is suitable for data modelling. Implementing the support vector machine method with the radial kernel in the fourth section showed the root mean square error (RMSE) value as 0.54366. Also, based on the confusion matrix, the accuracy measure of this model was equal to 0.70444; the sensitivity and specificity measures of this model were equal to 0.5482 and 0.8261, respectively.

The area under curve value based on the ROC curve of this model was obtained as 0.6872, which indicates the acceptable accuracy of this model for data modelling. The Kappa coefficient was equal to 0.384, which indicates the average correlation between the results of the support vector machine model with radial kernel and the actual values of the test data. On the other hand, McNemar's non-parametric test also showed that using a support vector machine model with a radial kernel is suitable for data modelling.

References

[1] J.L. Abbot, Y. Park and S. Parker, *The effects of audit committee activity and independence on corporate fraud*, Manag. Finance **26** (2000), no. 11, 55–67.

[2] S.M. Abeare, *Comparisons of boosted regression tree, GLM and GAM performance in the standardization of yellowfin tuna catch-rate data from the Gulf of Mexico Lonline fishery*, MSc Thesis, Department of Oceanography and Coastal Sciences, Pretoria, 2009.

[3] T. Bell and J. Carcello, *A decision aid for assessing the likelihood of fraudulent financial reporting*, Audit.: J. Practice Theory **9** (2000), no. 1, 169–178.

[4] M. Beasley, J. Carcello, D. Hermanson and P. Lapides, *Fraudulent financial reporting consideration of industry traits and corporate governance mechanisms*, Account. Horizons **14** (2000), 113–136.

[5] M. Broghani, S. Pourhahashemi, M. Zarei and K. Aliabadi, *Spatial modeling of the sensitivity of dust centers to its emission in east of Iran using BRT boosted regression tree model*, Arid Regions Geog. Stud. **9** (2018), no. 35, 14–28.

[6] G. Camps-Valls, D. Tuia, L. Gomez-Chova, S. Jimenez and J. Malo, *Remote Sensing Image Processing*, Morgan & Claypool Publishers, 2012.

[7] P.K. Chan, W. Fan, A.L. Prodromidis and S.J. Stolfo, *Distributed data mining in credit card fraud detection*, IEEE Intel. Syst. Appl. **14** (1999), no. 6, 67–74.

[8] J. Elith, J.R. Leathwick and T. Hastie, *A working guide to boosted regression trees*, J. Animal Ecology **77** (2008), no. 4, 802–813.

[9] E. Feroz, K. Park and V. Pastens, *The financial and market effects of the SECs accounting and auditing enforcements releases*, J. Account. Res. **29** (2000), 42–107.

[10] A. Higson, *Why is management reticent to report fraud?, An exploratory study*, 22nd Ann. Cong. Eur. Account. Assoc., Bordeaux, 1999.

[11] H. Kamrani and B. Abedini, *Formulation of financial statement fraud detection model using artificial neural network and support vector machine approaches in companies listed in Tehran Stock Exchange*, J. Manag. Account. Audit. Knowledge **11** (2022), no. 41, 285–314.

[12] A. Kornejady and H.R. Pourghasemi, *Landslide susceptibility assessment using data mining models, a case study: Chehalis-Chai Basin*, Watershed Engin. Mang. **11** (2019), no. 1, 28–42.

[13] E. Tashdidi, S. Sepasi, H. Etemadi and A. Azar, *New approach to predicting and detecting financial statement fraud, using the bee colony*, J. Account. Knowledge **10** (2018), no 3, 139–167.