Adv. Math. Fin. App., 2024, **9**(3), P.1006-1023

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

Original Research

Comparing The Performance of Machine Learning Techniques in Detecting Financial Frauds

Jafar Nahari Aghdam Qala Jough[a], Nader Rezaei[a,*], Yaqub Aghdam Mazrae[b], Rasul Abdi[a]

[a]*Department of Accounting, Bonab branch, Islamic Azad University, Bonab, Iran.*
[b] *Department of Accounting, Sufian Branch, Islamic Azad University, Sufian, Iran*

ARTICLE INFO

Article history:
Received 2022-10-23
Accepted 2023-02-23

Keywords:
Bayesian linear Regression
Neural Network
Logistic Regression
Financial Fraud

ABSTRACT

Abstract

Detecting financial fraud is an important process in the activities of companies. In the last decade, much attention has been paid to fraud detection techniques. Financial fraud is a problem with far-reaching implications for shareholders. Today, financial fraud in companies has become a big problem. Companies and regulatory agencies must continuously develop their mechanisms to detect fraud. Machine learning and data mining techniques are currently commonly used to solve this problem. However, these techniques still need to be improved in terms of computational cost, memory cost, and dealing with big data that is becoming a feature of current financial transactions. In this research, machine learning techniques including logistic regression, neural network, and Bayesian linear regression were used to detect financial frauds in the Iranian stock market. According to the obtained results, the support vector machine model with radial kernel has the lowest RMSE and the highest accuracy criterion, and the support vector machine model with linear kernel and Bayesian linear regression has the highest RMSE and the lowest accuracy criterion for modeling the financial fraud of companies in they were Tehran stock market. Also, the models of artificial neural network model, Bayesian linear regression and support vector machine model with linear kernel respectively had the lowest characteristic values and did not perform relatively well in detecting the existence of financial fraud in the companies present in the Tehran stock market.

1 Introduction

Nowadays, due to the wide variety of audiences, customers, markets, variety and complexity of services and business environments, access to appropriate information is necessary for making correct decisions. Therefore, it is necessary and vital for organizations to use appropriate solutions for classification and generation of information from a large amount of data. Sometimes, financial reporting may not provide correct information to people, which can be caused by accounting mistakes, and sometimes there is a possibility of fraud. If the financial statements contain distortions or omissions of important events in order to deceive users, it is fraudulent. Fraudulent financial statements through the manipulation of its

* *Corresponding author.*
E-mail address: naderrezaeimiyandoab@gmail.com

constituent elements by overstating assets, sales and profits or understating liabilities, costs and losses are incurred. Despite this issue, in recent years, the issue of fraud in financial statements has attracted a lot of attention and fraud in financial reports is increasing. Fraud of financial statements is known as accounting fraud, management fraud, or distorted financial reporting. This happens when financial statements (reporting) contain false information or ignore financial facts (measures, disclosures or evidence) to deceive users [1-3]. Fraudulent financial statements have had many negative effects on the world economy and have led to significant losses for individuals and companies. The evidence shows that there is currently fraud in financial statements and these frauds are very common and costly for the business world [4-5]. Different forms of fraud in financial fields according to the report of the Association of Official Fraud Examiners are: financial corruption, misappropriation of assets and fraud in financial reporting. On the other hand, financial statement fraud has a significant impact on corporate taxes compared to other frauds [6]. Therefore, fraud has devastating consequences for the future of the company, managers, employees, auditors, investors and the whole society. Will have. Big frauds cause the collapse of economic units or companies and cause significant losses for investors, also impose significant legal costs and can cause the loss of trust in They become capital markets. Finally, the cost of fraud in the form of an increase in the price of goods or providing services will affect all members of society [7], and on the other hand, the price of the company's shares will decrease, and finally, when the accounting fraud is revealed, the impact It leaves a significant negative impact on the company's reputation, brand name and credibility [8]. Due to the importance of the issue, it is also mentioned in the international accounting and auditing standards. Therefore, investigating the root causes of fraud is very important. Therefore, in a business environment, there is a fundamental need for effective methods to prevent and detect fraud in financial statements. Managers, employees and even auditors, inspectors and all persons who are in the process of preparing financial reports may commit fraud. Although the amount and variety of management fraud is less compared to employee fraud, but management fraud causes more losses to the company. Because managers have more authority and their opportunity to cheat is more than those who work under them, and with high amounts, they significantly affect the global economy and the stock market [9]. On the other hand, the external auditor provides an independent opinion regarding the fair presentation of management financial statements. Users of financial statements rely on auditors, and auditors are expected to detect fraud in financial statements. Therefore, the fraud of financial statements not only leads to significant losses to investors, lenders and other stakeholders and companies, but also can damage the credibility of the auditing profession, the credibility of the audit report, and the trust of customers [10]. Meanwhile, the audit process is not designed to detect the fraud of financial statements; because the auditors only rely on the company's management questionnaires, while the management is trying to maintain the existing fraudulent method to prevent the detection of fraud. Preventing fraud and detecting it are related issues, but they are not the same concept. Fraud prevention includes policies, approaches, training, and communications that prevent fraud from occurring, while fraud detection emphasizes activities and methods that detect or detect fraud in real time and with time sensitivity. Fraud is about to happen [11]. Unfortunately, the audit profession is not equipped with the tools and methods to detect fraud as a result of the limitations in the audit processes and the secret nature of financial statement fraud. Therefore, it is very necessary to equip auditors and financial supervisors with various and effective techniques to identify financial frauds. Fraud risk indicators, known as red flags, are used to predict fraudulent financial reporting and make informed economic decisions. Based on this, in most of the studies conducted, only corporate level variables have been emphasized in detecting fraud, this is in the situation that the business environment can also be

effective in the occurrence of financial frauds. The business environment is a set of factors that have an effect on the management of companies but are outside the control of companies [12]. The costs of any activity are subject to two categories of factors: a) the costs that are necessary in different production processes; b) The costs that are imposed on the owners of the companies due to the inappropriate economic environment. Sometimes, such costs are so high that managers and investors face serious challenges. In general, the variables related to the business environment are the main macroeconomic factors and these factors can significantly affect the performance of companies [13]. [3] showed that financial reporting of companies is largely affected by changing economic conditions. According to these topics, it is necessary to consider the variables of the business environment in the detection of financial frauds of companies. Due to the necessity of these issues, in the research, the efficiency of various methods of detecting financial fraud in companies is investigated and compared with emphasis on the fundamental factors of the company level, the market and the business environment in order to find the optimal model to predict the risk of fraud. Financial statements of companies should be identified. In this regard, machine learning techniques are used in this study. [5] |Used an expert system to prevent financial abuse and fraud. The results of this study showed that it is possible to examine the situation of the changing stages of management fraud risks in a very specific way. [6] Used Benford's law in accounting. The hypothesis of this research was that managers tend to round up the amount of profit. This researcher used the expected frequencies of the second digit of Benford's law in order to evaluate the increase of the second zero appearing in the net profit of companies. CARSA's results for New Zealand companies showed a higher frequency than expected according to Benford's law. Karsa stated that managers round up profit numbers. [8]Examined the usefulness of danger signs in a study. His research method was experimental and practical by using the control list. His research revealed that the use of the control list does not have a major effect on fraud risk assessments and there is no difference in the evaluation of users and those who do not use the control list. [10] Used a neural network to develop a model to detect management fraud. They also compared the results of neural network with linear regression and logistic regression. The results of this model showed that possible financial statement frauds can be discovered by analyzing customer documents. Also, the results indicated that the quality of the neural network model is better than the quality of statistical models. [12] Used a qualitative model to predict management fraud based on a set of data prepared by an international accounting institute. Probit and logit methods were used in this model. In a research, [1] examined the effect of the number of committed and non-committed members in fraudulent and non-fraudulent companies. The results of his research showed that in companies with more non-executive members of the board of directors, due to continuous monitoring of the effectiveness of management decisions, the possibility of fraud is reduced compared to companies with fewer non-executive members of the board of directors. Also, the results of Beasley's research showed that the composition of the board of directors affects the reduction of financial fraud more than the existence of the audit committee. [3] Used the artificial neural network method to detect fraud in financial statements, and the results of this study showed that this method has the ability to detect fraud in financial statements. [5] Used artificial neural network (ANN) to design a management fraud detection model. They compiled a model with eight variables with a high probability of discovery. The results of this research showed that the neural network has provided better results compared to logistic regression and standard statistical models. [7] Investigated the relationship between the holders of confidential information and fraud. By using Logit Abshari model in the direction of detecting fraud, they found that the owners of confidential information had reduced their shares in such companies to a large amount. [9] Examined whether certain types of financial reporting fraud lead

to high-probability lawsuits against independent auditors. They presented a new classification of types of fraud documents, which includes 12 general categories. They found that auditors are more concerned when they discover financial reporting frauds of the usual kind or when they discover frauds caused by fictitious transactions. [11] Used a neural network to distinguish between "real" and "manipulated" data. Using these data, they designed six neural networks and selected the best among the six neural networks. The designed neural network was able to distinguish real data from manipulated data among 800 financial data series in 68% of cases. Machine learning with the aim of finding patterns and generalizing them to the future has become one of the hot topics for predictions in recent years. Machine learning tries to give past data the power of acquisition. In other words, the data in a particular algorithm learns how to adapt itself in different situations and improve itself as a more complete identity. One of the advantages of machine learning is their implementation without using heavy programming. In recent years, many efforts have been made to use this field of knowledge in predicting financial and accounting variables. However, machine learning method has not been used in predicting and detecting financial frauds in Iran.

2 Materials and methods

This article is classified based on the purpose of applied research. The purpose of applied research is to develop applied knowledge in a specific field. Applied research is research whose findings can be applied to social issues. Also, the present research is a descriptive correlational research in terms of method and nature. Correlation research includes all research in which the relationship between different variables is tried to be discovered and determined using the correlation coefficient. Therefore, the correlation coefficient is an accurate index that states how much the changes of the variables are dependent on another variable. In terms of the implementation process, it is of the type of quantitative research, and in terms of the implementation logic, it is part of the research with an inductive-inductive approach, and in terms of the time dimension, it is of the longitudinal-retrospective research type, and it is based on the historical information of the companies.

2.1 Statistical Research Sample and Sampling Method

The statistical population of the current research is all the companies accepted in the capital market (both stock and over-the-counter) during the period of 2018-2019. The required accounting information was extracted from the financial statements of the listed companies. The reason for this choice is the greater attention of investors and financial analysts to the capital market, the availability and also the transparency of the companies' accounting information. The requirements of the stock exchange for the timely release of accounting information have created a more suitable information environment for research. The meaning of the time domain is the period of time that the researcher uses the information of the studied companies during that period in his research. The time domain of the current research is the 12-year period of 2018-2019. Because, in this period, all the data required for research and the information of the studied companies are accessible. In relation to the literature of the subject and its theoretical foundations, the necessary information is collected through library studies (including foreign and domestic books and magazines and scientific treatises) as well as searching in reliable scientific databases and tries to Concepts and theoretical foundations should be as comprehensive and summarized as possible. The statistical information required for the research is also prepared from the financial statements of accepted companies, financial information published on the Kodal website, Rahevard

Navin software, as well as the website of Central Bank, etc. Therefore, the collection tool used in this research was the use of documents. In general, in this research, data collection will be done in two stages. In the first stage, library sources are used to compile the theoretical foundations of the research, and in the second stage, official and approved statistical sources are used to collect the desired data.In this study, Excel software will be used to prepare the information, in such a way that after extracting the information related to the investigated variables from the mentioned sources, this information will be entered in the worksheets created in the environment of this software. And then, the necessary calculations are done to obtain the investigated variables. Finally, for the purpose of statistical analysis, the calculated variables were transferred to the R software environment and using this software and the desired packages, machine learning techniques such as Bayesian linear regression, logit regression, regression Boosted tree, neural network regression and support vector regression are implemented.To implement each of the methods of Bayesian linear regression, logit regression, augmented tree regression, neural network regression, and support vector regression, the data of companies admitted to the stock exchange and over-the-counter during the period of 2009-2019 will be used. In general, following experimental and theoretical studies, the following model will be used to investigate the factors affecting fraud in financial reporting of companies in each machine learning technique:

$$FF_{it} \cong f\ (OCFTNI_{it}, OCFTSAL_{it}, NITSAL_{it}, NITTA_{it}, NITEQ_{it}, TDTEQ_{it}, TLTTA_{it}, CATCL_{it}, CASHTTA_{it}, INVTTA_{it}, ARTTA_{it}, CATTA_{it}, CCC_{it}, INVTSAL_{it}, ARTSAL_{it}, COGTSAL_{it}, SALTTA_{it}, APTSAL_{it}, COGTINV_{it}, Board_{it}, ESTQ_{it}, NESTQ_{it}, TAX_{it}, UNEM_{it}, GS_{it}, OIL_{it}, INF_{it}, log + EXC_{it}, , log + GDP_{it}, , log + GDPP_{it}, , MTGDP_{it}, e_{it})$$

where

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 a value of one for otherwise (conditional, rejected and no comment);
OCFTNI_{it}	: Ratio of operating cash flow to net income for i-th company for year t
$OCFTSAL_{it}$	: operating cash flow to sales for company i for year t;
$NITSAL_{it}$	: net sales revenue for company i for year t;
$NITTA_{it}$	: return on assets for company i for year t;
$NITEQ_{it}$	: return on equity for i-th company for year t;
$TDTEQ_{it}$	: sum of debt to equity for i-th company for year t;
$TLTTA_{it}$	: total debt to total assets for company i for year t;
$CATCL_{it}$	: current assets to current liabilities for i-th company for year t;
$CASHTTA_{it}$	: cash to total assets for company i for year t;
$INVTTA_{it}$	: balance to total assets for company i for year t;
$ARTTA_{it}$	: accounts receivable to total assets for company i for year t;
$CATTA_{it}$	: current assets to total assets for company i for year t;
CCC_{it}	: cash flow cycle for i-th company for year t;
$INVISAL_{it}$	: cash balance for sale for i-th company for year t;
$ARTSAL_{it}$	: Accounts receivable for sale for company i for year t;
$COGTSAL_{it}$	: cost of goods sold for company i for year t;
$SALTTA_{it}$	: asset turnover ratio for i-th company for year t; (net sales over average total assets)
$APTSAL_{it}$	: accounts payable for sales for company i for year t;

$COGTINV_{it}$	: inventory turnover for i-th company for year t; (cost of goods sold over average inventory)
$Board_{it}$	: the number of board members for i-th company for year t;
$ESTQ_{it}$	: the number of board members required for company i for year t;
$NESTQ_{it}$	: the number of non-obligatory board members for i-th company for year t;
TAX_{it}	: the income tax rate in the entire economy for year t;
$UNEM_{it}$	: unemployment rate for year t;
GS_{it}	: ratio of government spending to total gross domestic product for year t;
OIL_{it}	: oil revenues for year t;
INF_{it}	: inflation rate for year t;
$\log(EXC_{it})$	: natural logarithm of the exchange rate in the informal market for year t;
$\log(GDP_{it})$	: logarithm of GDP for year t;
$\log(GDPP_{it})$	: logarithm of GDP per capita for year t;
$MIGDP_{it}$	: ratio of liquidity to GDP for year t;
e_t	: is the residual of the regression model for the i-th company in year t.

3 Data Analysis Methods

3.1 Support Vector Regression

Support vector machine as a new neural network algorithm is presented by [3]. Usually, a large number of traditional neural network algorithms follow the principle of experimental risk minimization, while the support vector machine method is based on the principle of structural risk minimization. The principle of experimental risk minimization is to minimize the classification error or incorrect grouping or deviation from the correct solution of training data, and the principle of structural risk minimization is to minimize the upper limit of the generalization error. In addition, the support vector machine solution can become the overall (absolute) optimum, while the neural network models tend to the local optimum solution. Therefore, meta-learning rarely happens in the support vector machine method [4]. In other words, the support vector machine is actually a binary (two-mode) classifier that separates and groups two classes or groups using a linear border [7]. The main idea of the support vector machine method is that, assuming that the clusters are linearly separable, it obtains the superplanes with the maximum margin that separate the clusters. According to a theory in statistical learning theory, if the training data is correctly classified, among the linear separators, the one that maximizes the margin of the training data will minimize the generalization error. In the meantime, the closest training data to Separating plane cloud is called support vector. These vectors (points) are used to specify the boundary between the layers [9]. Support vector machine is one of the most powerful and accurate machine learning algorithms, which is used to classify and separate groups. This algorithm combines statistical methods and machine learning. Therefore, its theoretical basis is based on statistical learning theory [10]. With the specified research data, the support vector machine model divides the data into distinct groups. These models have the general properties of data classification with maximum generalization ability, reaching the optimal point of data separation, automatic determination of the optimal structure for the classifier, and the possibility of modeling non-linear data using analysis and analysis are the main components [11].

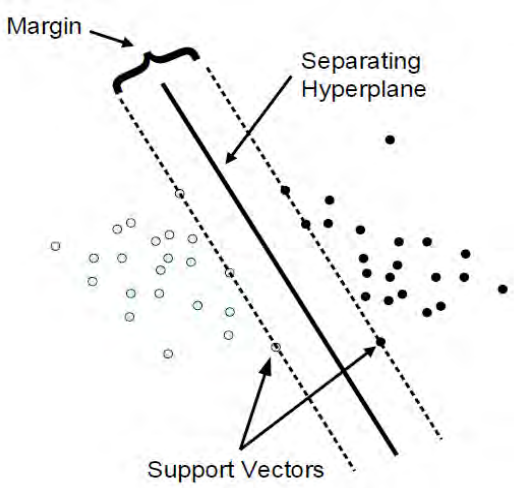

Fig. 1: Operation of support vector machine

Assume that there are n data points in the space, $\{(\bar{x}_1, c_1), (\bar{x}_2, c_2), \ldots, (\bar{x}_n, c_n)\}$ and $c_i \in \{+1, -1\}$ denote the classification symbol for the data point $\bar{X}_1$. These data are defined as training data to identify the optimal separation plane as follows:

$$\bar{w}.\bar{x} - \delta = 0$$

The symbol $\bar{w}$ indicates the separation margin and α is a constant. There are several solutions for $\bar{w}$, but the $\bar{w}$ is the optimal value with the maximum margin. The following equation is the solution to the optimization problem:

$$\text{minimize } \frac{1}{2}\|\bar{w}\|^2$$
$$\text{subject to } c_i(\bar{w}.\bar{x} - \alpha) \propto 1, 1 \propto i \propto n$$

After learning the network, $\bar{w}$ with the maximum margin, it is possible to create a classification $\hat{c}$ using the following equation on the test data that have not yet been classified [13].

$$\hat{c} = \begin{cases} -1, & \text{if } \bar{w}.\bar{x} - \alpha \propto -1 \\ +1, & \text{if } \bar{w}.\bar{x} - \alpha \propto 1 \end{cases}$$

3.2 Bayesian linear regression

In statistics, Bayesian linear regression is an approach to linear regression in which statistical analysis is performed in the framework of Bayesian inference. When the regression model has errors that are normally distributed and if a particular form of prior distribution is assumed for the parameters, explicit results for the posterior probability distributions of the model parameters can be calculated. In the Bayesian perspective, linear regression is formulated using probability distributions instead of point estimates. Bayesian linear regression is a type of conditional modeling in which the mean of one variable is described by a linear combination of other variables with the aim of obtaining the posterior probability of the regression coefficients (as well as other parameters describing the regression distribution).

In linear regression, the goal is to linearly estimate the dependent variable y from the independent variable x. In the standard linear regression, the conditional mean variable Yi is the condition of vector xi for i=1,...,n as follows.

$$y_i = x_i^T \beta + \varepsilon_i$$

Where β and x_i are $m \times 1$ vectors and ε_i are random and independent variables with a normal distribution with zero mean and variance γ^2 ($\varepsilon_i \square \text{Normal}(o, \gamma^2)$). In classical linear regression, parameters are estimated using ordinary least squares method as follows.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

In the classical method, $\hat{\beta}$ i.e. estimation of parameters is done only by using data. In linear Bayesian regression, it is assumed that the parameters, β i.e. a random variable itself, have a prior distribution $\pi(\beta)$. According to Bayes theorem, the prior distribution β is

$$\pi(\beta \mid D) = \frac{\pi(D \mid \beta) \partial \pi(\beta)}{\pi(D)}$$

The estimate $\hat{\beta}$ is obtained using the median or mean of the posterior distribution $\hat{\beta} = (\sigma_n X^T y)^T X$

4 Results and Discussion

The results and performance of three famous machine learning models, including logistic regression, neural network (ANN), Bayesian linear regression, were investigated to detect financial frauds 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). R statistical software was used to run these models. The obtained results revealed the existence of some differences in the predictor (independent) variables using each of these models.The data used was related to 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, which were replaced by the average of each of the corresponding variables. Among these 1500 observations, randomly and using set.seed(123), 1050 observations were considered as train data and the remaining 450 observations were considered as test data.

4.1 Artificial Neural Network (ANN) Model

In this study, an artificial neural network model with two hidden layers of 5 layers (5 by 5), 5000 training rounds (IPAC), Adam's optimization function and sigmoid action function was implemented using R software and ANN2 package. The value of the validation loss of this model was 0.62031. The graph of loss versus training period is shown in Fig. 2.

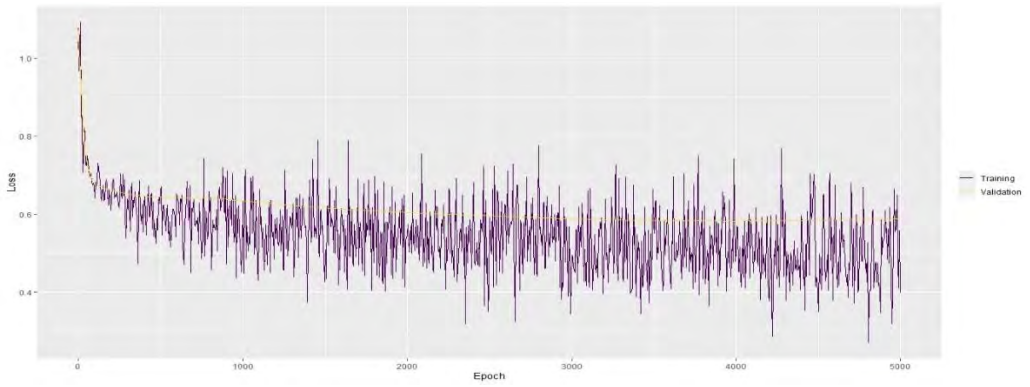

Fig. 2: Plot of loss versus training round for ANN model applied to train data

To check the goodness of fit of the ANN model, the model obtained from the train data was run on the test data. The root mean square error (RMSE) value was obtained as 0.56960. Also, the disturbance matrix of the ANN model for the test data is reported in Table 3-4. Based on this, the accuracy criterion of this model was equal to 0.67556, the sensitivity and specificity of this model were equal to 0.5888 and 0.7431, respectively.

Table 1: Perturbation matrix of ANN model for test data

		prediction	
		No Fraud	Fraud
reality	No Fraud	116((0.6888)	65((0.6509)
	Fraud	81((0.4112)	188((1.7431)

In addition, the ROC curve is also drawn for this model in Figure 3. Based on this graph, the AUC value of this model was equal to 0.7355, which indicates the acceptable accuracy of this model for data modeling.

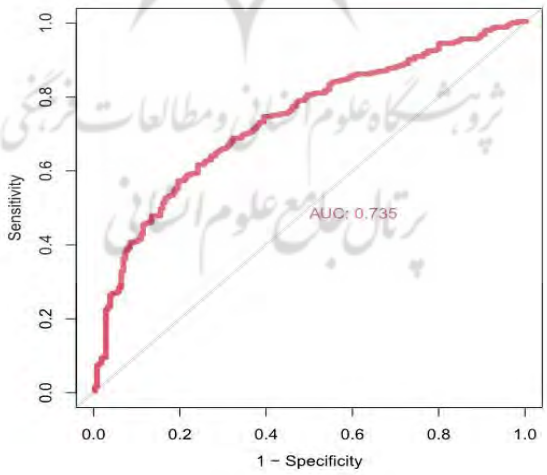

Fig. 3: ROC curve based on test data for ANN model

Kappa coefficient equal to 0.335 was obtained, which indicates the almost average relationship between the results obtained from the ANN model and the actual values of the test data. On the other hand, McNemar's non-parametric test rejected the hypothesis (zero) of the independence of the values predicted by the model and the actual values (P-value=0.2145). This result shows that the model did

not work very well in detecting financial fraud and the results obtained from the model were independent from the real data. Also, the hypothesis test that the accuracy criterion is smaller than the ignorance rate (0.5622) was tested and rejected (P-value=5.746x10-7). Therefore, based on this accuracy test, it is appropriate to use the ANN model for data modeling.

4.2 Bayesian Linear Regression Model

In this study, a Bayesian linear regression model was implemented using R software and mlbench, rstan, rstanarm, bayestestR, insight, broom and bayesplot packages. The generated Bayes chain diagram for each of the 31 predictor variables of the model is drawn in Figure 4. Also, the sample density diagram produced by Markov chain Monte Carlo (MCMC) for each of the predictor variables is drawn in Figure 5 and Figure 6.

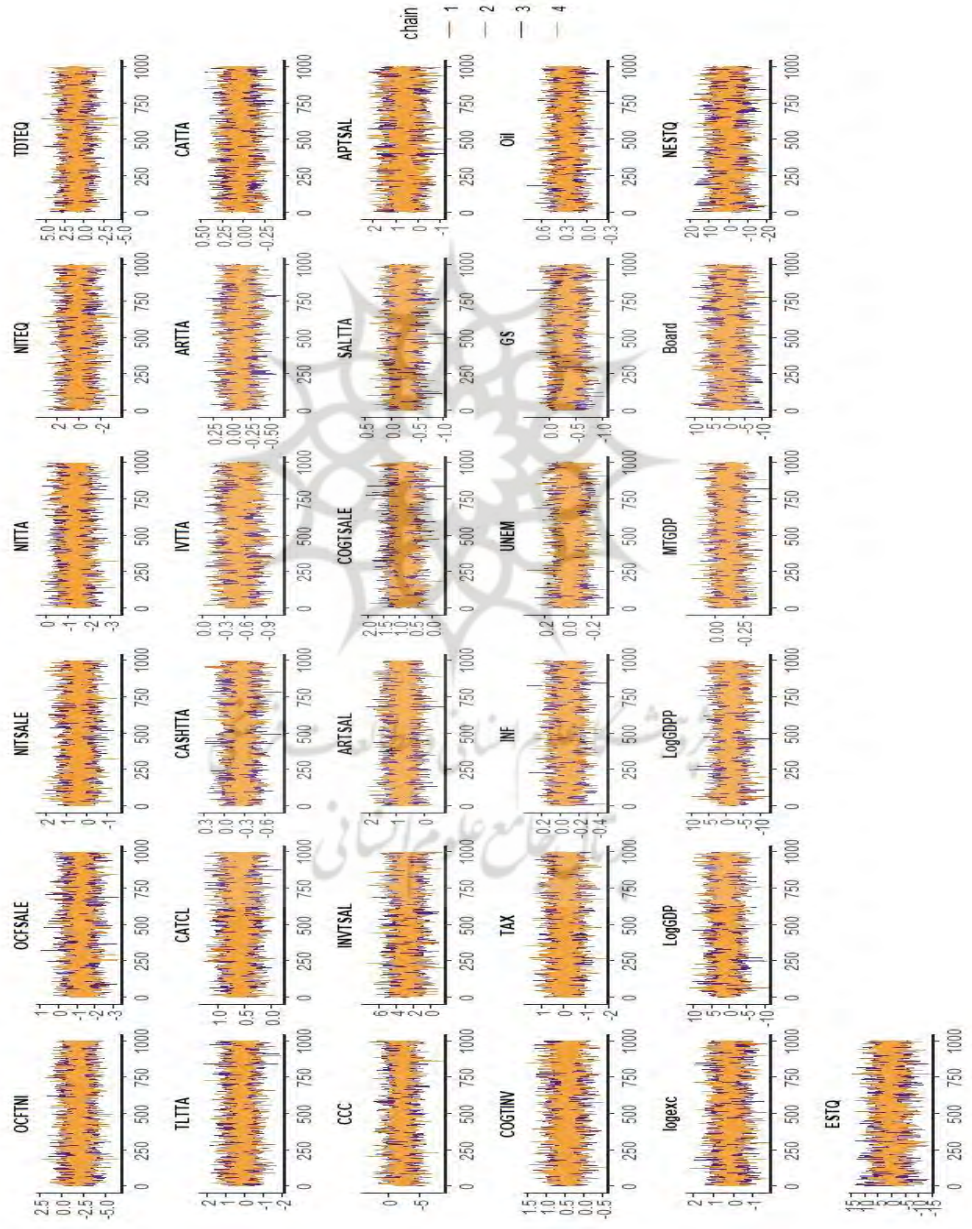

Fig. 4: Diagram of generated Bayesian chains for predictor variables of the model

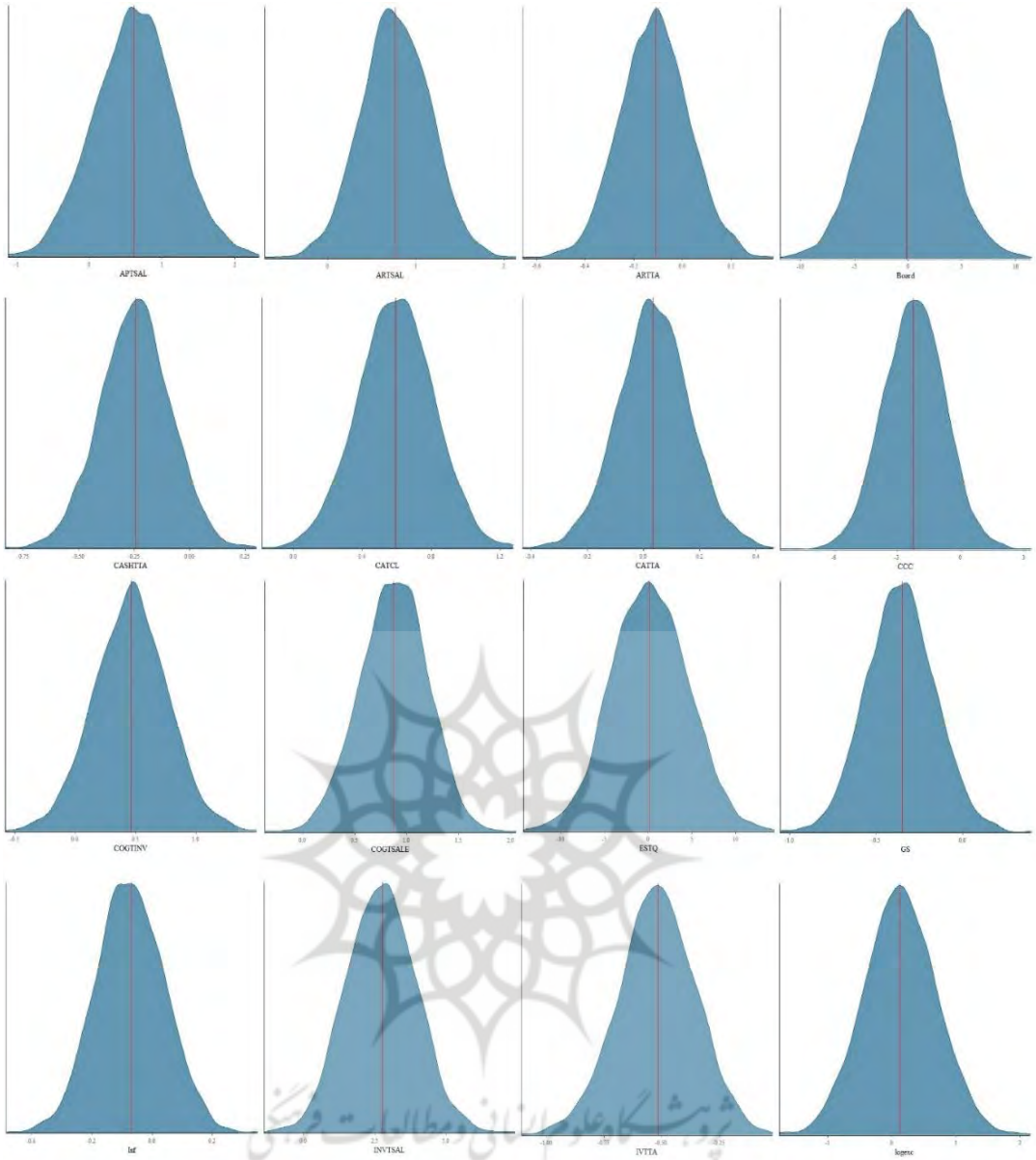

Fig. 5: Sample density diagram produced by MCMC for predictor variables (first part)

The results of running the Bayesian linear regression model are reported in Table 2. According to the results reported in Table 2, respectively, 10 variables IVTTA, CATCL, NITTA, COGTSALE, INVTSAL, OCFSALE, Oil, ARTSAL, GS and OCFTNI have the most impact and at the error level of 0.05 they have a significant impact on the financial fraud of the companies present in They had the Tehran stock market. Also, by increasing the error level to 0.10, the variables CASHTTA, CCC, COGTINV and MTGDP have had a significant effect on the financial fraud of companies present in the Tehran Stock Exchange. Other independent variables of the model did not have a significant effect on financial fraud. In addition, the influence of CATCL, COGTSALE, INVTSAL, Oil, ARTSAL and COGTINV variables was positive; In the sense that the large value of these variables indicates the presence of financial fraud in the company. The influence of IVTTA, NITTA, OCFSALE, GS,

OCFTNI, CASHTTA, CCC and MTGDP variables was negative; In the sense that the small value of these variables indicates the absence of financial fraud in the company. In Table 2, the median and mean values of the regression coefficients along with the probability of direction (PD), which is equivalent to the p-value of the classical approach in the Bayesian approach, and the Bayesian confidence interval with high posterior density (HPD) of 95% for these coefficients are reported. To check the goodness of fit of the Bayesian linear regression model, the model obtained from the train data was run on the test data. The root mean square error (RMSE) value was obtained as 0.59442. Also, the disturbance matrix of the Bayesian linear regression model for the test data is reported in Table 3. Based on this, the accuracy criterion of this model was equal to 0.64667, the sensitivity and specificity of this model were equal to 0.5076 and 0.7549, respectively.

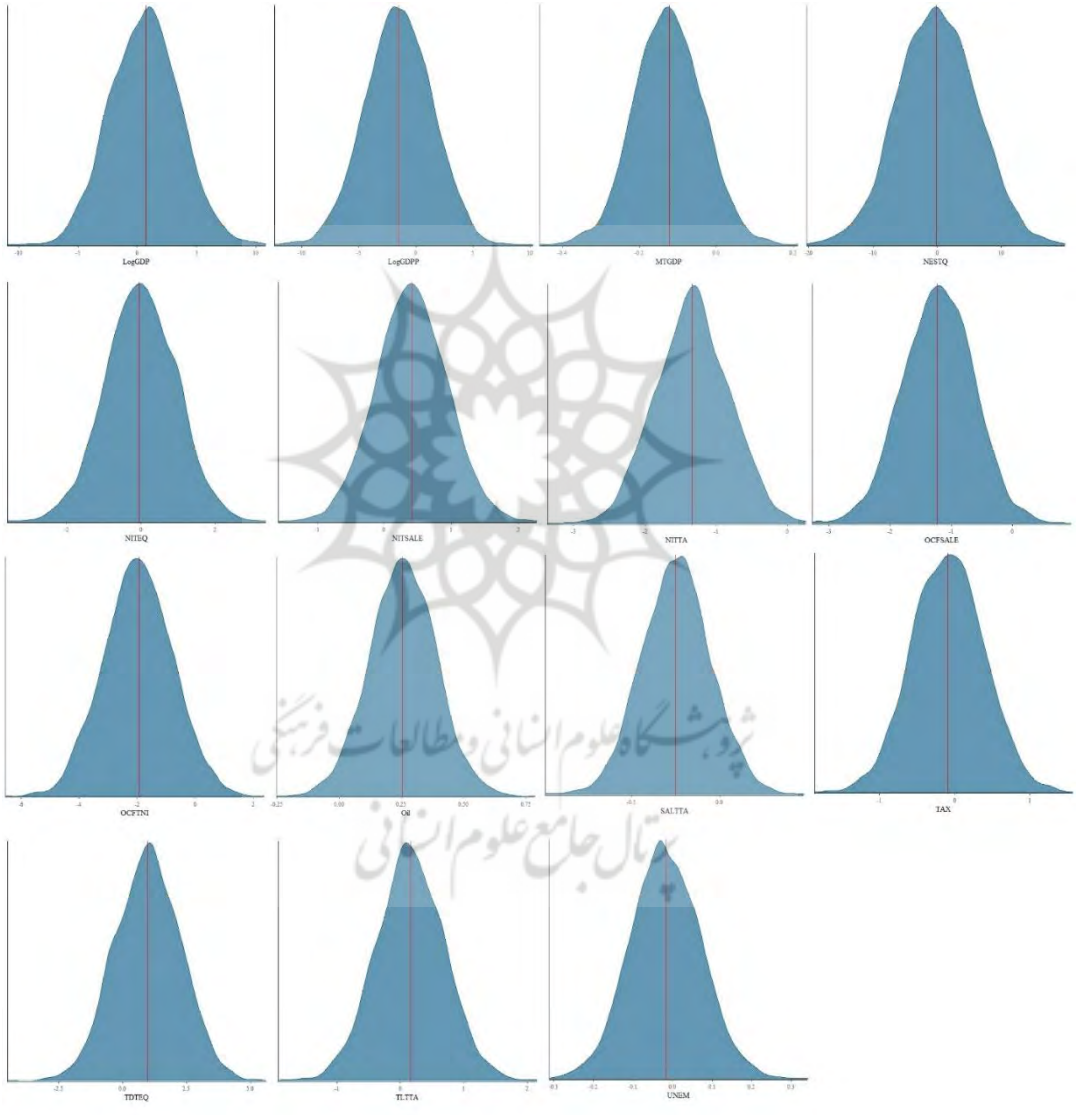

Fig. 6: Sample density diagram produced by MCMC for predictor variables (second part)

Table 2: The results of running the Bayesian linear regression model to detect financial fraud for Tehran Stock Exchange market data

Variable	median	mean	PD	HPD
Fixed coefficient	1.678	1.716	0.28875	[-4.23،،7.60]
IVTTA	-0.519	-0.519	0.00000	[-0.83,,-0.18]
CATCL	0.596	0.593	0.00225	[0.17,, 1.01]
NITTA	.1.329	-1.331	0.00350	[-2.28,,-0.32]
COGTSALE	0.883	0.875	0.00350	[1.49 ,0.25]
INVTSAL	2.737	2.731	0.00925	[4.93 ,0.27]
OCFSALE	-1.205	-1.213	0.01750	[-2.32,,-0.11]
Oil	0.256	0.256	0.02775	[-0.01,,0.51]
ARTSAL	0.771	0.775	0.02250	[0.05,,1.53]
GS	-0.358	-0.358	0.03525	[0.73,,0.02]
OCFTNI	-1.966	-1.961	0.04950	[-4.40,,0.17]
CASHTTA	-0.244	-0.245	0.05675	[-0.55 ,0.06]
CCC	-2.174	-2.194	0.05850	[-5.01,,0.61]
COGTINV	0.467	0.462	0.06450	[-0.14,,1.03]
MTGDP	-0.125	-0.124	0.07800	[-0.29,,0.05]
TAX	-0.085	-1.961	0.42600	[-1.00,,0.84]
NITEQ	0.006	0.002	0.49800	[-1.89,,1.85]
CATTA	0.036	0.037	0.37900	[-0.23 ,0.28]
SALTTA	-0.244	-0.246	0.10072	[-0.63,,0.12]
TLTTA	0.153	0.159	0.37875	[-0.90,,1.19]
ARTTA	-0.114	-0.115	0.20200	[-0.38,,0.17]
APTSAL	0.634	0.629	0.12775	[-0.45,,1.68]
ESTQ	0.110	0.161	0.48775	[-8.18,,,8..80]
NITSALE	0.406	0.408	0.21450	[-0.57,,1.48]
TDTEQ	0.991	0.981	0.22650	[-1.41,,3.51]
logexc	0.130	0.132	0.40550	[-0.92,,1.24]
UNEM	-0.021	-0.017	0.42100	[-0.18 ,0.17]
NESTQ	-0.058	0.019	0.49450	[-11.39 ,12.36]
LogGDP	0.713	0.652	0.40975	[-5.14,,6.13]
INF	-0.073	-0.071	0.26525	[-0.29,,0.15]
Board	-0.168	-0.201	0.47975	[-7.39,,6.12]
logGDPP	-1.498	-1.499	0.30875	[7.24,,4.13]

Table 3: Disturbance matrix of Bayesian linear regression model for test data

		prediction	
		No Fraud	Fraud
reality	No Fraud	100((0.5076)	62((0.2451)
	Fraud	97((0.4924)	191((0.7549)

In addition, the ROC curve is also drawn for this model in Fig. 6. Based on this graph, the AUC value of this model was equal to 0.7070, which indicates the acceptable accuracy of this model for data modeling.

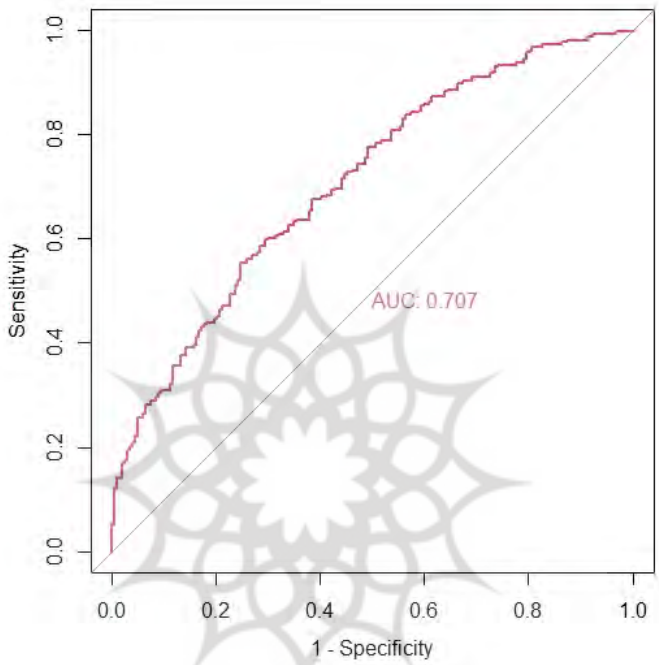

Fig. 7: ROC curve based on test data for ANN model

Kappa coefficient equal to 0.268 was obtained, which indicates a relatively weak relationship between the results obtained from this Bayesian linear regression model and the actual values of the test data. On the other hand, McNemar's non-parametric test rejected the hypothesis (zero) of the independence of the values predicted by the model and the actual values (P-value=0.000162). Also, the test of the hypothesis that the accuracy criterion is smaller than the ignorance rate (0.5622) was tested and rejected (P-value=0.00701). Therefore, using the Bayesian linear regression model is appropriate for data modeling.

4.3 Logistic Regression |Model

In this study, a logistic regression model was implemented using R software and the glm command. The results of running the Bayesian linear regression model are reported in Table 5-4. According to the Z-statistics reported in Table 5, the six variables ARTSAL, CATCL, IVTTA, NITTA, OCFSALE, ARTTA, and GS respectively have the most influence and at the error level of 0.05 they have had a significant effect on the financial fraud of the companies present in the Tehran stock market. . Also, by

increasing the error level to 0.10, the variables NITSALE, APTSAL, NITEQ, TLTTA, INVTSAL and COGTSALE have had a significant effect on the financial fraud of the companies present in the Tehran stock market. Other independent variables of the model did not have a significant effect on financial fraud. In addition, the influence of ARTSAL, CATCL, APTSAL, TLTTA, INVTSAL and COGTINV variables was positive; In the sense that the large amount of these variables increases the possibility of financial fraud in the company. The influence of IVTTA, OCFSALE, ARTTA, GS, NITSALE and NITEQ variables was negative; In the sense that the small amount of these variables reduces the possibility of financial fraud in the company.

Table 5: The results of implementing the logistic regression model to detect financial fraud for Tehran Stock Exchange market data

Variable	coefficient	Standard error	statistic Z	P -value
ARTSAL	11.9329	3.5076	3.402	0.00067
CATCL	5.2086	1.6022	3.251	0.001150
IVTTA	-2.6188	0.9272	-2.824	0.004739
OCFSALE	-11.9112	4.4325	-2.687	0.007200
ARTTA	-1.8962	0.7971	-2.379	0.017365
GS	-2.0586	0.9803	-2.100	0.035740
NITSALE	-8.3415	4.3546	-1.916	0.055420
APTSAL	6.2113	3.4204	1.816	0.069380
NITEQ	-0.9522	0.5453	-1.746	0.080780
TLTTA	7.3589	4.2330	1.738	0.082128
INVTSAL	14.5296	8.5872	1.692	0.090645
COGTSALE	3.2348	1.9635	1.647	0.099472

To check the goodness of fit of the logistic regression model, the model obtained from the train data was run on the test data. The root mean square error (RMSE) value was obtained as 0.56960. Also, the disturbance matrix of the logistic regression model for the test data is reported in Table 7-4. Based on this, the accuracy criterion of this model was equal to 0.67556, the value of sensitivity and specificity of this model was equal to 0.5635 and 0.7628, respectively.

Table 6: Logistic regression model disturbance matrix for test data

		prediction	
		No Fraud	Fraud
reality	No Fraud	111((0.5635)	60((0.2372)
	Fraud	86((0.4366)	193((0.7628)

In addition, the ROC curve is also drawn for this model in Figure 7. Based on this graph, the AUC value of this model was equal to 0.7166, which indicates the acceptable accuracy of this model for data modeling.

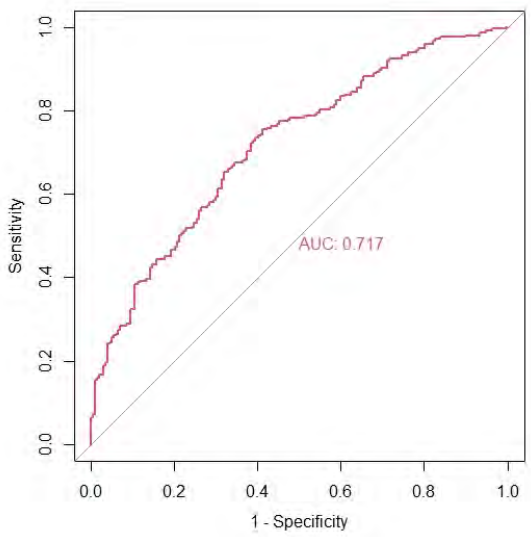

Fig. 8: ROC curve based on test data for logistic regression model

Kappa coefficient equal to 0.331 was obtained, which indicates an almost average relationship between the results obtained from the logistic regression model and the actual values of the test data. On the other hand, McNemar's non-parametric test rejected the hypothesis (zero) of the independence of the values predicted by the model and the actual values (P-value=0.03854). Also, the hypothesis test that the accuracy criterion is smaller than the ignorance rate (0.5622) was tested and rejected (P-value=5.746x10-7). Therefore, it is appropriate to use the logistic regression model to model the data.

5 Conclusion

The results of the implementation of the enhanced tree regression method in the fourth chapter showed that the average square root of the error was 0.55104. Also, based on the disturbance matrix, the accuracy criterion of this model was equal to 0.69111, the sensitivity and specificity of this model were equal to 0.5736 and 0.7826, respectively. The value of the area under the curve based on the ROC curve of this model was found to be 0.7563, which indicates the acceptable accuracy of this model for data modeling. The Kappa coefficient results were equal to 0.362, which indicates an average relationship between the results of the enhanced tree regression model and the actual values of the test data. On the other hand, McNemar's non-parametric test also showed that using the enhanced tree regression model is suitable for data modeling. The enhanced tree regression showed that out of the 31 independent variables selected, 16 variables are effective in detecting financial fraud based on changes in the prediction deviation, which are respectively the income tax rate in the entire economy (with a relative importance of 9.6637), return on assets (with relative importance of 8.5812), balance to total assets (with relative importance of 8.3659), return on equity (with relative importance of 7.8584), accounts payable for sale (with relative importance of 8.5330), current assets to total assets (with relative importance of 6.3194), asset turnover ratio (with relative importance of 6.2874), cash to total assets (with relative importance of 6.1128), operating cash flow ratio to net income (with relative importance of 5.8014), total debt to total assets (with relative importance of 5.4227), current assets to current liabilities (with relative importance of 5.3941), cost of goods sold (with relative importance of 4.9528), ratio of liquidity to gross domestic product (with relative importance of 4.9104), accounts receivable to total assets (with relative

importance of 4.7768), cash flow cycle (with relative importance of 4.7428) and inventory turnover for the company (with relative importance of 4.2767).The results obtained from the implementation of the artificial neural network model method in the fourth chapter showed that the average square root of the error was equal to 0.56960. Also, based on the disturbance matrix, the accuracy criterion of this model was equal to 0.67556, the sensitivity and specificity of this model were equal to 0.5888 and 0.7431, respectively. The value of the area under the curve based on the ROC curve of this model was found to be 0.7355, which indicates the acceptable accuracy of this model for data modeling. The Kappa coefficient results were equal to 0.335, which indicates an average relationship between the results of the artificial neural network model and the actual values of the test data. On the other hand, McNemar's non-parametric test also showed that the use of artificial neural network model is suitable for data modeling.The results of the implementation of Bayesian linear regression method in the fourth chapter showed that the mean square root of the error was equal to 0.55104. Also, based on the disturbance matrix, the accuracy criterion of this model was equal to 0.64667, the sensitivity and specificity of this model were equal to 0.5076 and 0.7549, respectively. The value of the area under the curve based on the ROC curve of this model was equal to 0.7070, which indicates the acceptable accuracy of this model for data modeling. The Kappa coefficient results were equal to 0.268, which indicates a relatively weak relationship between the results of the Bayesian linear regression model and the actual values of the test data. On the other hand, McNemar's non-parametric test also showed that the use of Bayesian linear regression model is appropriate for data modeling. Bayesian linear regression showed that among 31 independent variables, 10 variables are inventory to total assets, current assets to total assets, return on assets, cost of goods sold, cash balance to sales, operating cash flow. to sales, oil revenues, accounts receivable to sales, the ratio of government expenditures to the total gross domestic product and the ratio of operating cash flow to net income have the most impact and at the error level of 0.05 have a significant impact on the financial fraud of companies in the stock market They have had Tehran. Also, by increasing the error level to 0.10, the variables of cash to total assets, cash circulation cycle, inventory circulation and ratio of liquidity to GDP have had a significant effect on the financial fraud of companies present in the Tehran stock market. Other independent variables of the model did not have a significant effect on financial fraud. In addition, the influence of the variables of current assets on current liabilities, cost of goods sold, cash balance for sale, oil revenues, accounts receivable for sale and inventory turnover was positive; In the sense that the large value of these variables indicates the presence of financial fraud in the company. The influence of inventory variables on total assets, return on assets, operating cash flow to sales, the ratio of government expenditures to total gross domestic product, the ratio of operating cash flow to net income, cash to total assets, the cash flow cycle and the ratio of liquidity to production GDP was negative; In the sense that the small value of these variables indicates the absence of financial fraud in the company.

References

[1] Keshavarz, Ahmad, Ghasemian, Hassan, A Fast Algorithm Based on Support Vector Machine for Classifying Hyperspectral Images Using Spatial Correlation, *Iranian Electrical and Computer Engineering Journal*, 2014; 37-44.

[2] Karanjadi, Ayding, Pourqasmi, Hamidreza , Landslide susceptibility assessment using data mining models, a study Case*: Chahalchai Watershed*, 2018; 11(1) : 42

[3] Kim, K., Financial Time Series Forecasting Using Support Vector Machines, *Neurocomputing* 2003;55:307-319.

[4] Elith, J., J.R. Leathwick and T. Hastie, a working guide to boosted regression trees, *Journal of Animal Ecology*, 2008; 77(4): 802-813.
[5] Kamrani, Hossein, & Abedini, Bijan, Formulation of financial statement fraud detection model using artificial neural network and support vector machine methods in companies admitted to Tehran Bahadur Stock Exchange, *Knowledge of Accounting and Management Audit*, 1401;11(41):285-314.
[6] Abeare, S.M., Comparisons of boosted regression tree, GLM and game performance in the standardization of yellowfin tuna catch-rate data from the Gulf of Mexico online fishery. 2009
[7] Moradi, Mohsenshafiei Sardasht, Morteziabrahimpour, Maleeha, Predicting the financial distress of companies by support vector machine models and multiple audit analysis, *Stock Exchange Quarterly*, 2013; No. 18, 5(9): 113-136.
[8] Broghni Mehdi, Porfashmi Sima, Zarei Mehdi, Aliabadi Kazem, Spatial modeling of the sensitivity of dust centers to its emission in eastern Iran using BRT enhanced regression tree model, *Geographical studies of dry areas*. 1398; 9 (35): 14-28
[9] Kamrani, Hossein, & Abedini, Bijan, Formulation of financial statement fraud detection model using artificial neural network and support vector machine methods in companies admitted to Tehran Bahadur Stock Exchange, *Knowledge of Accounting and Management Audit*, 1401;11(41): 285-314.
[10] Essani, Elahe, Sepasi, Dr. Sahar, Etemadi, Dr. Hossein, Azar, Dr. Adel, presenting a new approach in predicting and detecting financial statement fraud using the bee algorithm, *Journal of Accounting Knowledge*, (2018; 10(3): 139-167. doi: 10.22103/jak.2019.13616.2927
[11] Che. Tsai, Hung. C, Automatically Annotating Images With Keywords: A Review Of Image Annotation Systems, *Recent Patents On Computer Science*, 2008; 55-68.
[12] Camps-Valls, G., Tuia, D., Gomez-Chova, L., Jimenez S. and Malo, J., Remote Sensing Image Processing, *Morgan & Claypool Publishers*, 2012;176
[13] Yara Alghofaili, Albatul Albattah & Murad A. Rassam, A Financial Fraud Detection Model Based on LSTM Deep Learning Technique, Journal of Applied Security Research, 2020; doi: 10.1080/19361610.2020.1815491