Journal of
Empirical Research In Accounting

Alzahra University, Faculty of Social Sciences and Economics
Empirical Research in Accounting, Winter 2022, V. 12, No. 45, pp.41-44

Financial Reporting Fraud Scheme Prediction via Machine Learning Approach – Multiclass Classification[1]

Tohid Kazemi[2], Parviz Piri[3]

Received: 2022/06/15
Accepted: 2022/11/20

Research Paper

INTRODUCTION

The financial statements presented by economic enterprises are an information source that contains useful information for decision-making with investors and creditors. Financial statements provide an overview of the economic activities of the enterprise. The accumulation of this information over a long time, courses information overload, and the analysis of this information takes time. On the other hand, fraudulent financial statements are an undesirable phenomenon that affects the market price, and the shareholder's value. It can harm investors' truth in the financial reporting system. On the other hand, fraudulent financial reporting disrupts the fair distribution of wealth by misleading investors, creditors, and the government. Following fraud in financial reporting, limited economic resources are directed toward unsuccessful economic enterprises. It causes to waste the economic resources. (Sajadi and Kazemi, 2015).

Discovering and extracting fraud patterns from financial statements with data mining techniques can aware of financial statements users. Data science originates from various sciences such as statistics, artificial intelligence, machine learning, pattern recognition, and database. Machine learning is a two-step process. In the first stage, the machine learning algorithm is applied to a set of data to identify useful patterns in the dataset. In the second stage, when the model is created, it is used for analysis (Kelher and Tierney, 2021). These patterns are presented in

1. DOI: 10.22051/JERA.2022.41290
2. Assistant Professor, Department of Accounting, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran. (Corresponding Author). (t.kazemi@alzahra.ac.ir).
3. Associate Professor, Department of Accounting, Faculty of and Economics and Management, Urmia University, Urmia, Iran. (p.piri@urmia.ac.ir).

different ways such as regression models, artificial neural networks, decision trees, support vector machines, and boosting algorithms.

Literature review in the research area shows previous research investigated machine learning models' performance in binary space. They attempted to predict occurring fraud in financial reporting, not fraud schemes that occurred in financial statements. The present research attempts to fill the research gap in the research area by developing machine learning models with a multi-classification approach.

MATERIALS AND METHODS

According to the literature review, the research hypothesis is determined as follows:

1. The Support Vector Machine performance is preferred to other machine learning models in financial reporting fraud scheme prediction via multiclass classification approach.
2. The Support Vector Machine performance is preferred to other machine learning models in financial reporting fraud scheme prediction via binary classification approach.

The present research is classified as applied research in terms of purpose. Library and document analysis methods have been used to collect data. The population is all the companies admitted to the Tehran Stock Exchange, which are examined in the period from 2009 to 2021. The statistical sample includes companies that a) were admitted to Tehran Stock Exchange in 2009. b) Fiscal year was adopted to Hijri Shamsi calendar and not changed c) information was available. d) Was not classified as financial intermediaries, investors and banks.

data The set contains financial ratios as independent variables and nominal variables which refers to fraud scheme as dependent variables. According to Forqandoost Haqiqi et al (2015), Khajavi et al (2018) and Rezaiee (2021) to recognize the fraud scheme, the auditor's reports were analyzed and the fraud scheme was inferences from the auditor’s qualified opinion. Logistic Regression, Decision Tree, Boosting Algorithms, and Support Vector Machine models were implemented with Python via multiclass and binary classification space. Performance metrics were calculated according to confusion matrixes. To compare models’ performance, Friedman's Two-Way Analysis of Variance by Ranks was performed.

RESULTS AND DISCUSSION

Logistic Regression, Decision Tree, Boosting Algorithm, and Support Vector Machine models were implemented via a multi-classification approach. Table 1 represented the results. According to the results, a significant difference between machine learning models' performance was approved. Support Vector Machin was preferred.

Table 1. Machine learning models performance via multi-classification approach

model	Accuracy	Recall		Precision		F1		Cohen's Kappa
		Macro	micro	Macro	micro	Macro	micro	
Logistic Regression	0.4955	0.2105	0.4955	0.1591	0.4955	0.1713	0.4955	0.0641
Decision Tree	0.4570	0.3009	0.4570	0.3098	0.4570	0.3029	0.4570	0.1997
GBoost	0.5252	0.2753	0.5252	0.244	0.5252	0.2675	0.5252	0.2071
Support Vector	0.5401	0.2812	0.5401	0.3170	0.5401	0.2788	0.5401	0.2284
Friedman's Two-Way Analysis of Variance								
Test Statistic	11.550					Sig	0.009	
Pairwise Comparisons								
		Test Statistic	Sig			Test Statistic	Sig	
Logistic Regression	Decision Tree	-.625	.333	Decision Tree	GBoost	-.125	.846	
Logistic Regression	GBoost	-.750	.245	Decision Tree	Support Vector	-1.500	.020	
Logistic Regression	Support Vector	-2.125	.001	GBoost	Support Vector	-1.375	.033	

Furthermore, for binary classification, Machine learning models were implemented to predict each fraud scheme exclusively. Table 2 represented the results.

Table 2. Machine learning models performance via binary classification approach

scheme	model	Accuracy	Recall	Precision	F1	AUC
Overstatement of assets, understatement of debt and expenses	Logistic Regression	0.555	0.943	0.532	0.680	0.781
	Decision Tree	0.611	0.231	0.960	0.37	0.611
	GBoost	0.627	0.819	0.595	0.688	0.708
	Support Vector Machine	0.610	0.799	0.582	0.671	0.711
	Friedman's Two-Way Analysis	Test Statistic		2.52	Sig	0.472
Overstatement assets, understatement expenses	Logistic Regression	0.527	1.00	0.515	0.679	0.784
	Decision Tree	0.500	1.00	0.500	0.667	0.683
	GBoost	0.589	0.965	0.551	0.701	0.744
	Support Vector Machine	0.723	0.715	0.730	0.720	0.775
	Friedman's Two-Way Analysis	Test Statistic		4.71	Sig	.194
understatement debt and expenses	Logistic Regression	0.500	1.00	0.500	0.666	0.645
	Decision Tree	0.592	0.629	0.590	0.606	0.620
	GBoost	0.540	0.921	0.524	0.666	0.688

scheme	model	Accuracy	Recall	Precision	F1	AUC
	Support Vector Machine	0.790	0.786	0.787	0.785	0.808
	Friedman's Two-Way Analysis	Test Statistic		5.69	Sig	.127
Overstatement assets and income	Logistic Regression	0.443	0.433	0.433	0.422	0.344
	Decision Tree	0.611	0.900	0.589	0.708	0.588
	GBoost	0.639	0.900	0.608	0.722	0.69
	Support Vector Machine	0.836	1.00	0.763	0.865	0.792
	Friedman's Two-Way Analysis	Test Statistic		14.02	Sig	.003

According to the results to predict fraud schemes via binary classification, a significant difference between machine learning models' performance was not approved except to predict the "Overstatement assets and income" scheme. Friedman's Two-Way Analysis of Variance pairwise comparisons implemented on models' performance to predict the "Overstatement assets and income" scheme. Support Vector Machin was preferred to Logistic Regression and Decision Tree model.

CONCLUSION

According to the results via a multi-classification approach, a significant difference between machine learning models' performance was approved. Support Vector Machin was preferred in multiclass problem space with the unbalanced data set. To predict fraud scheme via binary classification, a significant difference between machine learning models' performance was not approved except to predict the "Overstatement assets and income" scheme. Support Vector Machin was preferred to Logistic Regression and Decision Tree model. The present research attempts to fill the research gap in the research area by developing machine learning models with a multi-classification approach.

Keywords: Fraud Scheme, Fraudulent Financial Reporting, Machine Learning, Multi-Classification.

JEL Classification: M41, M42, G32.