ASSESSING AND FORECASTING THE RISK OF FINANCIAL STATEMENT FRAUD OF ENTERPRISES LISTED ON THE VIETNAMESE STOCK MARKET - A LOGISTIC REGRESSION MODEL APPROACH

PhD. Nguyen Thu Thuy* - Ngo Tran Thanh Ngan* - Nguyen Thu Hang*
Nguyen Thi Thanh Huyen* - Bui Trinh Minh Ngoc*

Abstract: ***This article aims to assess and forecast the risk of financial statement fraud of enterprises listed on the Vietnamese stock market from 2018 to 2023. Based on the application of the Logistic Regression model with machine learning technique together with the famous M-Score model of Beneish (first published in June 1999) on the data collected from 25 enterprises, including 279 observations. The results of the study show that the application of the logistic regression model is capable of detecting fraud risks in financial statements with relatively high reliability. Most indicators have the same impact on the likelihood of fraudulent financial reporting.***

• Keywords: *financial statements, fraud risk assessment and forecasting, logistic regression model, m-score model, vietnamese stock market.*

JEL codes: B26, C58, D53

Date of receipt: 10th Oct., 2024
Date of receipt revision: 26th Nov., 2024
Date of delivery revision: 12th Oct., 2024
Date of approval: 30th Nov., 2024
DOI: https://doi.org/10.71374/jfar.v25.i1.15

1. Introduction

Financial statements are a set of documents created by an enterprise to reflect the financial situation of an enterprise at a certain time or in a certain period. Financial statements provide information on the situation of assets, liabilities, equity, and business results of the enterprise. Financial statement fraud is the act of deliberately manipulating, falsifying, or omitting information in financial statements to deceive relevant parties, such as investors, creditors, and regulatory agencies. The main forms of fraud are false declaration of expenses (hiding part of the cost to increase profits), declaring debts (minimizing liabilities to improve the financial situation), increasing revenue shorts (recording unrealized revenue or increasing revenue higher than reality), and misvaluation of assets (valuing assets higher than their real value to increase total assets).

Financial statement fraud is a sore problem in Vietnam's stock market, as many listed companies deliberately falsify financial information to deceive investors and regulators. In recent years, the number of fraud detections has tended to increase, especially in high-risk industries such as real estate, banking, and finance. Some typical cases can be mentioned, such as FLC Group with false disclosure of financial statements, Tan Hoang Minh related to fraudulent bond issuance, or Louis Holdings with allegations of manipulating stock prices. These frauds not only cause great losses to investors but also reduce confidence in the stock market, negatively affecting capital flows and financial stability. Therefore, to detect fraudulent acts early and protect the interests of stakeholders, it is extremely necessary to build fraud risk forecasting models. This study focuses on the application of the logistic regression model to assess and forecast the risk of financial statement fraud of enterprises listed on the Vietnamese stock market.

Logistic regression is a statistical technique widely used in machine learning to predict the probability of a binary event (with or without fraud) thanks to its ability to determine the probability of fraud occurring based on a variety of input financial variables, such as debt-to-equity ratio, operating cash flow, net profit, or irregularities in revenue fluctuations. One of the key advantages of logistic regression is its high explainability, which makes it easy for managers,

* *Thuongmai University, email: nguyenthuthuy@tmu.edu.vn*

auditors, and investors to understand the impact of each factor on the likelihood of fraud. In other words, it helps us answer the question: "What are the chances of this event happening?" The authors based on the Financial Fraud Assessment Model was made public by Beneish in June 1999 and some other application, like Score (2016). A logistic regression model (see Alpaydin (2020)) will be applied on the financial statement data of 25 companies in the period 2018-2023 to assess and forecast the risk of financial statement fraud of enterprises in the Vietnamese stock market.

In addition to the introduction, the rest of the research paper is presented as follows: Section 2 is an overview of previous studies, section 3 presents research results, and section 4 contains conclusions.

2. Research Overview

In the context of Vietnam's growing stock market, ensuring the transparency and accuracy of corporate financial information is extremely important. However, financial statement fraud is still a worrying issue, affecting investor confidence and market stability. There have existed numerous studies on assessing and forecasting the risk of financial statement fraud of enterprises as listed belows.

Dong et al. (2018) assembled a distinctive dataset comprising 64 fraudulent firms and an equivalent sample of 64 non-fraudulent firms, along with social media data preceding each firm's alleged fraudulent violation as documented in Accounting and Auditing Enforcement Releases (AAERs). The proposed framework automatically extracted various indicators, including sentiment features, emotional characteristics, topic-related attributes, lexical elements, and social network metrics, which were subsequently input into machine learning classifiers for fraud detection. The performance of the algorithm was evaluated and compared against baseline methods that relied exclusively on financial ratios or language-based features.

With a similar topic, Maranzato et al. (2010) examined the detection of fraudulent activities targeting reputation systems in e-markets. The primary objective was to generate a ranked list of users (sellers) based on their likelihood of engaging in fraud. Initially, transaction-related characteristics indicative of fraudulent behavior were identified and extended to sellers. A basic ranking method was outlined, which classified sellers by tallying these fraud-related attributes. Subsequently, additional characteristics unsuitable for the counting approach were incorporated, and logistic regression was applied to both the original and enhanced datasets. Using real data from a prominent Brazilian e-market for training and evaluation, the enhanced approach with logistic regression demonstrated superior performance. The final ranked list identified 32.1% of the most probable fraudsters targeting the reputation system, achieving a 110% increase in identified cases while confirming zero false positives.

In addition, Aftabi et al. (2023) presented an innovative approach leveraging generative adversarial networks (GAN) and ensemble models to address the scarcity of non-fraudulent samples while effectively managing the high-dimensional feature space. A new dataset was developed by compiling annual financial statements from ten Iranian banks and extracting three categories of features as outlined in the study. Experimental results on this dataset revealed that the proposed method excelled in generating synthetic samples susceptible to fraud, demonstrating its efficacy.

Previously, Beneish (1999) used financial measures to analyze 363 samples collected and obtained from 49 violating enterprises. It was pointed out that companies with unusual revenue growth rates compared to the industry are highly likely to have manipulated profits. While recently, Mutemi & Bacao (2024) presented a text-based fraud detection framework designed to mitigate financial losses effectively. The framework consisted of four essential components: text preprocessing, representation, knowledge extraction through machine learning algorithms, and model evaluation. By incorporating data augmentation techniques, it improved the performance of classifiers in identifying fraudulent activities. The proposed approach, which employed a combination of FastText and Random Forest classifiers, attained a remarkable F1 score of 0.833 and an AUC score of 0.99 on an augmented dataset, outperforming traditional keyword-based models.

Thanks to that kind of idea, this research based on the M-score model and applying the decision tree model to assess and forecast risk of Financial Statement Fraud of Enterprises Listed on the Vietnamese Stock Market with data extracted from financial statements of enterprises listed on the Vietnamese stock market with 25 enterprises, including 279 observations, in the period from 2018 to 2023, which is an updated version for seeking new empirical results.

3. Research results

3.1. Research data and model

The data used in this analysis of 25 enterprises listed on the Vietnamese stock market was collected from https://cafef.vn in the period from 2018 to 2023, including 279 observations. The dataset of 8

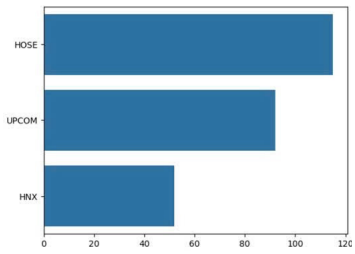
observation variables was built and processed based on the M-score financial fraud assessment model (see Beneish (1999)).

*M-score = – 4.84 + 0.92*DSRI + 0.528*GMI + 0.404*AQI + 0.892*SGI + 0.115*DP – 0.172*SGAI + 4.679*TATA – 0.327*LVGI.*

Where M-score is the dependent variable, predicting possibility of financial statement fraud, independent variables include DSRI (Days' Sales in Receivables Index) which is the average collection period change index, GMI (Gross Margin Index) is the index that measures the decline in a company's profit margin over time, AQI (Asset Quality Index) measures the increase in long-term assets other than fixed assets, SGI (Sales Growth Index) measures sales growth rate, DP (Depreciation) is the index of depreciation reduction of the company, SGAI (Sales, General and Administrative Expense) assesses the change in the ratio of selling and administrative expenses to net revenue, TATA (Accruals) is the accrual variable on total assets, and LVGI (Leverage Index) compares total corporate debt to total assets.

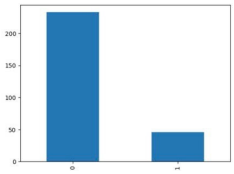
From the overview of the research data, it can be seen that in the research dataset, there is a difference between the number of enterprises collected on stock exchanges and the number of enterprises collected each year. The HOSE stock exchange has the largest proportion of enterprises included in the study and is followed by UPCOM and HNX, respectively. In Figure 1, the vertical axis offers stock exchanges, while horizontal one presents the quantities of businesses studied.

Figure 1. The number of businesses included in the study of each exchange



Source: Drawn by the authors

Figure 2. Number of enterprises at risk of financial statement fraud and not at risk of financial



Source: Drawn by the authors

In Figure 2, businesses at risk of financial statement fraud are labeled 1; businesses without risk of financial statement fraud are labeled 0. Additionally, the research data is divided into two parts in an 80/20 ratio, corresponding to 80% of the data for the model training process and 20% of the data for the model performance testing process.

3.2. Empirical results

A decision tree model is programmed in Python and trained on the training dataset; after completing the training, the effectiveness of the model is clarified. Table 1 presents the results on decision tree model accuracy executed on the test dataset.

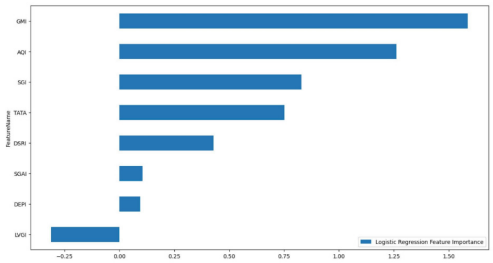
Table 1. Decision tree model accuracy results on the test dataset

```
Confusion matrix:
    0  1
0  45  0
1   5  6

Classification report:
              precision    recall  f1-score   support

           0       0.90      1.00      0.95        45
           1       1.00      0.55      0.71        11

    accuracy                           0.91        56
   macro avg       0.95      0.77      0.83        56
weighted avg       0.92      0.91      0.90        56

Logistic Regression accuracy on the test set:  0.9107
```

Source: Summarized by the authors

Out of a total of 45 cases of non-fraudulent financial statement predictions, there were 45 cases of accurate predictions, none of which were fraudulent. Similarly, out of a total of 11 cases of financial statement fraud predictions, there were 6 cases of accurate predictions and the remaining 5 cases of non-fraud. The model with high accuracy (> 0.9) means that the accuracy of the points found is high, and a high recall means a high true positive rate, which means that the rate of missing positive points is low, indicating that the model does not miss a significant number of cheats. In addition, a high F1 score and an overall accuracy of 91.07% also indicate that the model has good predictive ability. Detailed influence of each factors included in M-Score model on the likelihood of fraud is illustrated in Figure 3.

Figure 3. The impact of the indicators on the likelihood of fraud



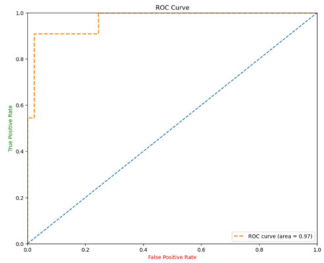
Source: Summarized and drawn by the authors

One can see that the indices "DSRI," "AQI," "GMI," "SGI," "TATA," "DEPI," and "SGAI" have a

positive impact on the likelihood of fraudulent financial reporting. Specifically, the index with the most impact on the logistics regression model is "DSRI," followed by "AQI," "GMI," "SGI," "TATA," "DEPI," and "SGA"; the impact level decreases, respectively. In contrast, the variable "LVGI" hurts the target in the logistic regression model, which means that when this indicator decreases, the likelihood of fraud increases.

The ROC line of the logistic regression model is very close to the upper left corner of the chart and tends to approach the y-line = 1, as in Figure 5, showing that the classification and forecasting ability of the model is appropriate.

Figure 4. ROC Curve Graph



Source: Drawn by the authors

In addition, Area Under the Curve (AUC) = 0.97 (close to 1) also indicates that this model has good performance.

4. Conclusion

This study has shown that the application of the Logistic regression model in assessing and forecasting the risk of financial statement fraud brings high efficiency thanks to the ability to accurately classify, clearly interpret, and have flexible applicability on the financial data of listed enterprises. However, to improve practice efficiency, it is necessary to have a close combination of mathematical models and traditional audit methods and apply them to automated financial monitoring systems to detect businesses with signs of fraud early. Auditing firms, regulatory agencies such as the State Securities Commission, and institutional investors can integrate this model into the risk analysis system to build a set of criteria for assessing financial transparency, thereby improving the quality of audits and risk management.

Although logistic regression has proven effective in research, there are still some limitations that need to be overcome to further optimize the model. One of the key improvements is the optimization of input variable selection, which can be done using the Lasso Regression or PCA method to reduce noise and eliminate variables that are not important. In addition, standardizing the data and using methods to handle data imbalances such as SMOTE (Synthetic Minority Over-sampling Technique) will help improve accuracy when the percentage of fraudulent businesses in the study sample is usually very low compared to non-fraudulent businesses. Another direction is to experiment with nonlinear regression models, such as multinomial logistic regression or in combination with Bayesian logistic regression algorithms, to enhance accuracy in cases where data has many nonlinear characteristics.

In addition to the logistic regression model, future research can be expanded by combining it with other machine learning algorithms in the supervised learning and unsupervised learning teams. In the supervised learning group, models such as Random Forest, XGBoost, or SVM (Support Vector Machine) can improve accuracy by leveraging offline learning capabilities and minimizing overfitting. In particular, ensemble models such as bagging and boosting can be applied to combine logistic regression with more powerful algorithms to optimize predictive performance. In the Unsupervised Learning group, methods such as K-Means Clustering or Autoencoders can help detect anomalies in financial data, thereby providing additional indicators for the Logistic Regression model.

In conclusion, this study not only confirms the effectiveness of logistic regression in detecting financial reporting fraud but also opens up many directions for improvement and combination with other machine learning models to improve accuracy and practical application. The implementation of the model in the automatic financial monitoring system will help Vietnam's stock market become more transparent, contributing to protecting the interests of investors and improving the stability of the economy.

References:

Aftabi, S. Z., Ahmadi, A., & Farzi, S. (2023), Fraud detection in financial statements using data mining and GAN models. Expert Systems with Applications, 227, Article 120144. https://doi.org/10.1016/J.ESWA.2023.120144.

Alpaydin, E. (2020). Introduction to machine learning (fourth edition). MIT Press.

Beneish, M. D. (1999), Incentives and penalties related to earnings overstatements that violate GAAP. The Accounting Review, 74(4), 425-457.

Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging financial social media data for corporate fraud detection. Journal of Management Information Systems, 35(2), 461–487. https://doi.org/10.1080/07421222.2018.1451954.

Maranzato, R., Pereira, A., Do Lago, A. P., & Neubert, M. (2010), Fraud detection in reputation systems in e-markets using logistic regression. In Proceedings of the ACM Symposium on Applied Computing (pp. 1454–1459). https://doi.org/10.1145/1774088.1774400.

Mutemi, A. & Bacao, F. (2024), "Balancing act: Tackling organized retail fraud on e-commerce platforms with imbalanced learning text models", International Journal of Information Management Data Insights, Vol. 4, Iss. 2, 100256, https://doi.org/10.1016/j.jjimei.2024.100256.

Score, K. M. B. (2016), "Detecting financial statement fraud by Malaysian public listed companies: The reliability of the Beneish M-Score model". Journal Pengurusan, 46, 23-32.

https://cafef.vn.