Research on financial risk screening of listed companies based on clustering algorithm

Zihan Liu [1], Zijun Shi [2, *]

[1] School of Management Science and Engineering, Central University of Finance and Economics, Beijing, China, 102206

[2] Department of Finance, Nankai University Binhai College, Tianjin, China, 300270

* Corresponding Author Email: 1874055419@qq.com

Abstract. The ability to recognize financial fraud activity in listed firms has grown in importance and is a constant source of worry in both academics and business. The A-share listed businesses in the biological and pharmaceutical industries that have faced penalties from the China Securities Regulatory Commission during the previous five years are used as samples in this study. Using Kangmei Pharmaceutical as an example, 24 characteristics are chosen using the fraud triangle theory, and testing and analysis are conducted using a Random Forest classification algorithm model in conjunction with SMOTE Oversampling technology. The findings show that it is more beneficial to use numerous feature sets to develop models or to construct models with many alternative algorithms for financial fraud screening studies, as opposed to merely classifying organizations into fraudulent and non-fraudulent categories.

Keywords: Financial risk, Deep learning, Fraud triangle theory.

1. Introduction

Financial fraud detection in listed companies is a significant issue, drawing attention from both industry and academia. The imperfect mechanism of stakeholder governance and the problematic operating behavior of controlling shareholders in China have led to the provision of false financial information by some listed companies. These companies, driven by motivations such as attracting investment, reducing tax pressures, avoiding delisting, present misleading financial data to external entities. China's financial market has an imperfect regulatory mechanism because it started rather late. This context makes the effective identification of financial fraud in listed companies a critical and ongoing concern for both industry and academia. In recent years, a variety of methods and technologies ranging from statistical models to machine learning models have been employed in research for financial fraud detection.

Within this framework, this paper introduces the use of advanced clustering algorithms to conduct a sophisticated financial risk screening of listed companies, particularly focusing on the pharmaceutical and biological sectors. The primary study samples are selected from A-share listed companies that have been penalized by the China Securities Regulatory Commission over the past five years. Central to this study is the fraud triangle theory, which emphasizes three distinctive features: rationalization, motivation, and opportunity, within a fraud framework. This framework guides the identification of 24 distinct features and applies augmented SMOTE oversampling technology to a random forest classification algorithm model for thorough testing and analysis.

2. Factors of Financial Fraud

2.1. Theory of Fraud Triangle

The fraud triangle theory suggests that financial fraud stems from three interrelated elements: opportunity, pressure, and rationalization. Opportunities for fraud often arise in settings where controls are lacking, thus creating openings for unethical practices. Pressure usually originates within the company, such as through stringent financial targets or personal financial needs, pushing

individuals toward fraudulent activities. Rationalization is the mental process fraud perpetrators use to internally justify their unethical actions.

2.2. Influence of Shareholders and Investors

Shareholders and investors can unintentionally increase the risk of financial fraud. Elevated expectations for financial performance or ambitious growth targets can pressure management into manipulating financial statements. This is especially prevalent in companies where there is a direct correlation between stock price performance and executive compensation.

2.3. Role of Internal and External Factors

A mix of internal and external factors contribute to financial fraud. Internally, weak governance, a poor ethical culture, and inadequate financial controls create an environment conducive to fraud. Externally, factors like market competition, regulatory changes, and economic conditions can also motivate fraudulent activities. Understanding these factors is vital for developing effective strategies to reduce the risk of financial fraud.

There are two main types of methods for detecting financial fraud. One type is traditional statistical models represented by cluster analysis, principal component analysis and logistic regression. For example, Etemadi and Zolghi studied how traditional statistical models predict financial fraud in listed companies, and selected the logistic regression model after analysis; Persons used the stepwise-logistic method to further study such issues. The other type is the emerging machine learning model, which typically includes SVM (Support Vector Machine), MLP neural network (Multi-Layer Perception, multi-layer perceptron), etc. Cecchini et al. used the SVM-FK method to predict financial fraud of listed companies based on the support vector machine (SVM) model; Bao et al. used the integrated learning method to process the original data. The experimental results showed that the prediction of the improved machine learning model The effect is greatly improved. Traditional statistical models are simple and easy to understand, easy to calculate, and the calculation results are highly interpretable. However, most of these models do not perform well in nonlinear data. Machine learning models have been a hot topic of research in recent years. In terms of financial fraud detection, the precision rate, recall rate, accuracy rate and F1 score of this type of model perform well.

3. Research Design and Methods

The selection of appropriate features (or variables) is a critical step in the machine learning pipeline. It directly impacts the performance of the model, as irrelevant or redundant features can lead to increased complexity, overfitting, and reduced generalizability of the model. Proper feature selection aids in building a more efficient, interpretable, and accurate model. Features with strong predictive power improve model accuracy, while irrelevant features can mislead the model, leading to poor predictions. Additionally, fewer but more relevant features can significantly reduce computational complexity, making the model faster and more efficient.

The features for this study are selected based on their relevance to the financial health and integrity of biopharmaceutical companies, as indicated by historical data and domain expertise. This includes financial ratios, market performance indicators, operational metrics, and other factors that are critical in assessing the financial status of a company. The exact features will be chosen based on an analysis of the dataset provided in the Excel file, ensuring they are indicative of the financial risks specific to the biopharmaceutical industry.

3.1. Evaluation of Different Methods

In choosing the appropriate method for this study, several factors were considered:

Table 1. factors were considered to choosing the appropriate model

Accuracy	The ability of the method to correctly classify companies based on financial risk.
Robustness	The method's performance in the face of noisy or incomplete data.
Interpretability	The ease with which the results can be understood and explained to stakeholders.
Computational Efficiency	The resources required to train and run the model.

Incorporating these features into a Random Forest model can provide a comprehensive view of the financial health and potential risks associated with these companies. The Random Forest algorithm is particularly suited for this analysis due to its ability to handle a large number of input features and its robustness in dealing with different types of data. It can also effectively manage the potential non-linear relationships and interactions between these variables.

3.2. Overview of model principles

- Let X be the input feature set and Y the target variable.
- A Random Forest consists of N decision trees, $T_1, T_2, \ldots, T_N$.
- Each tree T_i is trained on a random subset of the data with replacement (bootstrap sample).
- At each node of the tree, a random subset of features is selected, and the best split is determined based on a criterion like Gini impurity or entropy for classification, and mean-squared error for regression.
- Gini Impurity: $G = 1 - \sum(pi)^2$
- Entropy: $H = -\sum pi \log_2(pi)$
- The final prediction of the Random Forest is obtained by averaging the predictions of all individual trees for regression, or by majority voting for classification.

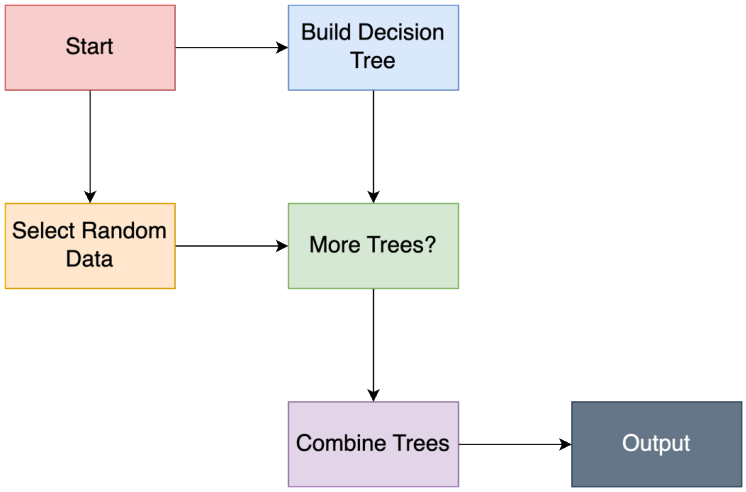

Figure 1. The basic architecture of the random forest algorithm

3.3. Collection of Data

The data for this study was collected from the Guotai'an CSMAR database, a comprehensive and authoritative source for Chinese financial and economic information. The database is known for its extensive coverage of financial data, including key financial indicators, stock prices, and corporate information for A-share listed companies. In order to ensure comprehensive analysis, we evaluate the data from three perspectives: Data Span, Data Scope, Reliability and Authenticity.

Data Span: We collected several years' worth of data to capture both short-term fluctuations and long-term trends.

Data Scope: The dataset includes a variety of financial metrics, governance details, and operational indicators relevant to the biopharmaceutical industry.

Reliability and Authenticity: Guotai'an CSMAR is widely recognized for its accuracy and reliability, with data sourced directly from regulatory filings, company reports, and verified financial statements.

3.4. Model clustering process

To create and use the model specifically, import the scikit-learn machine learning library for the random forest classifier first. Import SMOTE using imblearn.over_sampling to finish the preparatory steps. The data are sorted chronologically, the first 80% are divided into a training set, the final 20% are divided into a test set, and the training set is balanced using SMOTE oversampling technique. After that, create a random forest classifier model, set the model's decision tree count to 1000, feed the training set into the model to be trained, and feed the test set to evaluate the model's impact. The fraud opportunity model, fraud pressure model and fraud excuse model use the same algorithm model and sampling technology, but are independent of each other and do not affect each other.

4. Result

The study using the Random Forest algorithm on a dataset of A-share listed biopharmaceutical companies has yielded significant insights. Fig.2 provides a visual representation of the importance of various features in identifying the financial fraud risks of listed companies in the pharmaceutical and biological industry.

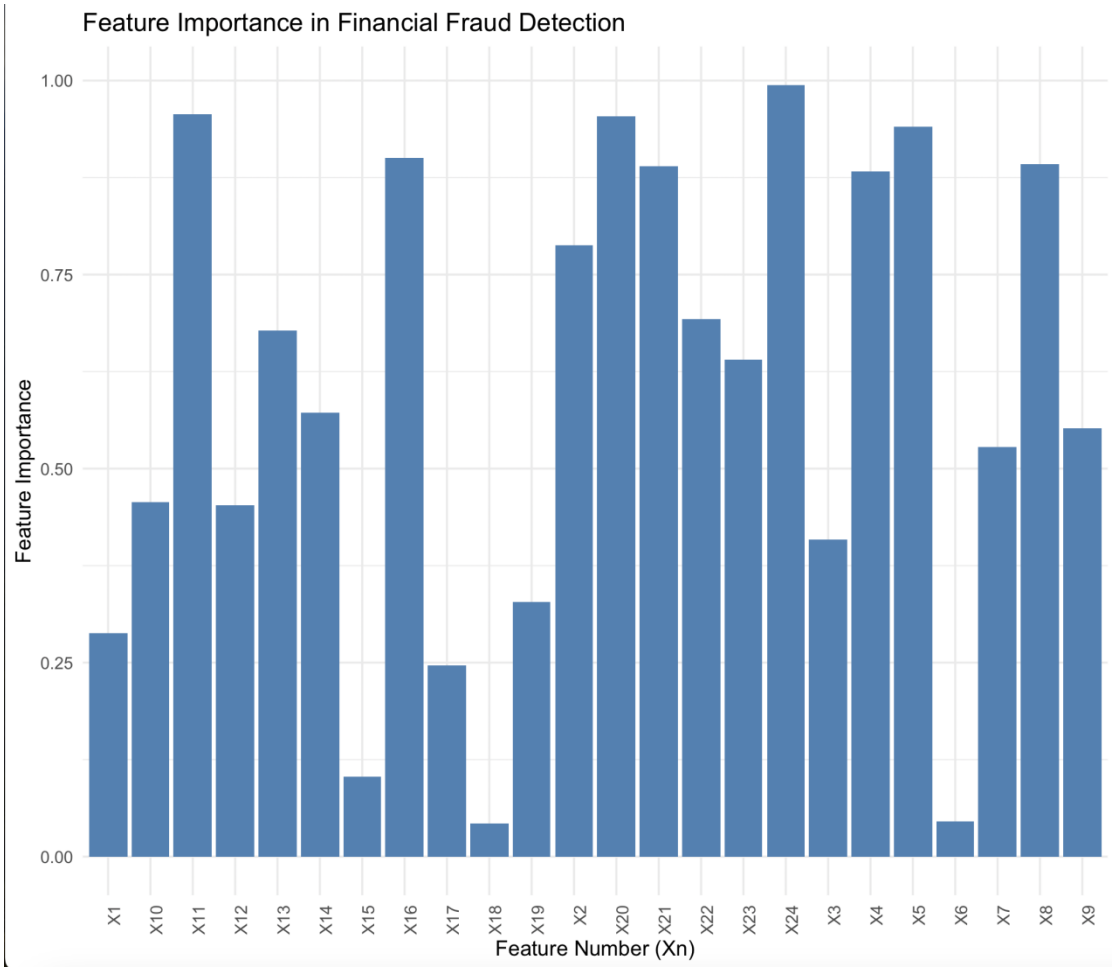

Figure 2. The histogram reflecting the important characteristics of identifying the financial fraud risks of listed companies in the pharmaceutical and biological industry

References

[1] Khan, A. M. A. and Peng, J. (2022) Using Machine Learning Meta-Classifiers to Detect Financial Frauds. Finance Research Letters, 48, Article 102915. https: //doi.org/10.1016/j.frl.2022.102915.

[2] Marco, S. A., Luis, U.A. and José, E.J. (2022) Predictive Fraud Analysis Applying the Fraud Triangle Theory through Data Mining Techniques. Applied Sciences, 12, 3382. https: //doi.org/10.3390/app12073382.

[3] Gozman, D. and Currie, W. (2014) The Role of Investment Management Systems in Regulatory Compliance: A Post-Financial Crisis Study of Displacement Mechanisms. Journal of Information Technology, 29, 44 - 58. https: //doi.org/10.1057/jit.2013.16.

[4] Cressey, D. R. (1953) Other People's Money; a Study of the Social Psy-chology of Embezzlement. Patterson Smith Publishing Corporation, Montclair.

[5] Call, A. C., Kedia, S. and Rajgopal, S. (2016) Rank and File Employees and the Discovery of Misreporting: The Role of Stock Options. Journal of Account-ing and Economics, 62, 277 - 300. https: //doi.org/10.1016/j.jacceco.2016.06.003.

[6] Etemadi, H. and Zolghi, H. (2013) Application of Logistic Regression to Identify Fraudulent Financial Reporting. Journal of Audit Science, 13, 5 - 23.

[7] Persons, O.S. (2011) Using Financial Statement Data to Identify Factors Associated with Fraudulent Finan-cial Reporting. Journal of Applied Business Research, 11, 38 - 46. https: //doi.org/10.19030/jabr. v11i3. 5858.

[8] Cecchini, M., Aytug, H., Koehler, G.J. and Pathak, P. (2010) Mak-ing Words Work: Using Financial Text as a Predictor of Financial Events. Decision Support Systems, 50, 164 - 175. https: //doi.org/10.1016/j.dss.2010. 07. 012.

[9] Bao, Y., Ke, B., Li, B., Yu, Y.J. and Zhang, J. (2020) Detecting Accounting Fraud in Publicly Traded US Firms Using a Machine Learning Approach. Journal of Accounting Research, 58, 199 - 235.https: //doi.org/10.1111/1475 - 679X. 12292.

[10] An, B. and Suh, Y. (2020) Identifying Financial Statement Fraud with Deci-sion Rules Obtained from Modified Random Forest. Data Technologies and Applications, 54, 235 - 255. https: //doi.org/10.1108/DTA-11 - 2019 - 0208.