Application of Machine Learning in Financial Fraud of Listed Companies: An Innovative Prediction Model

Zehao Wang[1](✉), Moqin Yang[2], Yizhan Du[3], and Hanqing Hu[4]

[1] School of Management, Huazhong University of Science and Technology, Wuhan 430074, People's Republic of China
m202177730@hust.edu.cn

[2] School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, People's Republic of China
202112190085@stu.zuel.edu.cn

[3] School of Law, Zhongnan University of Economics and Law, Wuhan 430073, People's Republic of China
duyizhan@stu.zuel.edu.cn

[4] School of Economics, Huazhong University of Science and Technology, Wuhan 430074, People's Republic of China
huhq@hust.edu.cn

Abstract. The over-reliance on financial statements published by listed companies as the main reference data can lead to great losses to capital market investors and hinder the orderly and healthy development of the capital market in the event of financial fraud by the company. In this context, the establishment of effective forecasting models to predict and analyze financial fraud has become the focus of research to avoid these economic traps. In this paper, we take the financial statement data of Shanghai and Shenzhen A-share listed companies in China during 2000–2020 as the observation sample, and establish a new universal and effective prediction model, which overcomes the unbalanced training of machine learning, and the innovative index system is finally externally verified with a prediction accuracy of 98.0% after three rounds of screening by psychological preference survey, feature engineering and model evaluation, leading all similar current financial fraud prediction models of listed companies.

Keywords: Financial Fraud · Predictive Models · Machine Learning · Feature Engineering

1 Introduction

For external investors, the most mainstream path to understand the information needed for investment is through the financial statements published by listed companies, and financial fraud will distort the information in the financial statements, and the analysis and conclusions built on them will be invalid, which will bring huge economic losses to external investors. As the number of listed companies continues to grow, financial fraud is a common occurrence.

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 957–965, 2023.
https://doi.org/10.2991/978-94-6463-198-2_100

Financial fraud is actually an irregularity organized by the company's management with the purpose of falsifying profits and deliberately misstating and omitting financial data in the financial statements, and systematically falsifying them. Reliable and accurate forecasting can effectively reduce the loss and impact caused by financial fraud, so developing an effective forecasting model is of strong theoretical and practical value. In the academic field, the research on the prediction model of financial fraud of listed companies has been active with the increase of the number of listed companies and the outbreak of economic crisis. In fact the conclusions of domestic and foreign researches on this subject are not consistent.

To build a model for predicting the existence of financial fraud in a company, the selection of relevant financial indicators is the basis for building the model. There are various methods of feature selection, and since the problem studied in this paper belongs to the scope of supervised learning, we focus on supervised feature selection. Supervised feature selection contains three types: filter, wrapper and embedded. Filtering methods can be independent of the learning algorithm in the feature selection phase, which first assesses the importance of features, ranks them, and filters them according to a set threshold, mainly divided into relevance measures [1, 2], mutual information measures [3–5], distance metric [6, 7] and consistency metric [8].Packing methods can iteratively use the learning performance of the classifier or regression model to evaluate the quality of the selected features, and such methods often use search algorithms, such as the combination of SVM and PSO for feature selection proposed by Li et al. in 2012[9]. In contrast, embedded methods can use the intrinsic structure of the learning algorithm to embedded feature selection into the underlying model, feature selection is performed simultaneously with the learning algorithm to select valuable features, and features are ranked using the parameters inside the classifier, and embedded algorithms are classified into pruning class algorithms [10, 11], algorithms based on tree structure models [12] and regularization algorithms [13]. In practical applications, features are often influenced by rationalization factors such as users' personal qualities and psychological preferences, however, almost most of the previous studies did not pay attention to this point in feature selection, while in this paper, the influence of rationalization factors is fully considered and feature selection is performed based on psychological preference surveys.

This paper intends to build an accurate and universal model to analyze and study the financial statement data of Chinese listed companies in Shanghai and Shenzhen A-shares during 2000–2020, covering the most important 90 industries, taking into account the influencing factors such as time and industry. This paper firstly tried different data filling methods and sample processing principles, then conducted feature screening using filter, wrapper and embedded methods, and determined the final evaluation index system based on their common performance in the three models of XGBoost, Random Forest and LightGBM, and finally based on SMOTETomek comprehensive sampling method Finally, the model was resampled based on the SMOTETomek integrated sampling method, and the balanced data set was used for model training, and the optimal model was determined through comparative screening to establish the financial fraud prediction model of listed companies. The established model is applicable to different industries and different categories of enterprises, and can correctly identify a small number of categories of fraudulent instances often more realistic value, early warning of enterprises

with the risk of financial fraud, which effectively reduces the economic losses caused by fraudulent behavior, not only to assist auditors and other professionals to make decisions, but also more relevant to the needs of general investors to use.

2 Materials and Methods

2.1 Data Sources

The data are obtained from the Chinese Guotaian CSMAR database for the period from January 1, 2000 to December 31, 2020. The data set has a total of 46,697 data, including 283 indicators of profit statements, cash flow statements and asset and liability statements of all listed companies in Shanghai and Shenzhen A-shares during 20 years. The financial violation penalties and violation announcements issued by the Securities Regulatory Commission, Shenzhen Stock Exchange, Shanghai Stock Exchange, listed companies, Ministry of Finance and other institutions during the period from January 1, 2000 to December 31, 2020 were collated and the list of companies where financial fraud occurred was obtained after screening. This study strictly follows the provisions of China Auditing Standard for Certified Public Accountants No. 1141 and other regulations to define financial fraud. Financial fraud is defined as "fictitious profit", "false listing of assets", "false record (misleading statement)", "delayed disclosure ", "material omission", "inaccurate disclosure (other)", and "appropriation of company assets".

2.2 Processing Method

The dataset was processed in Python and it was found that there were missing values in the dataset, and some of the features had a missing ratio of more than 90%, and these features would reduce the performance of training for the model. The correlation test between such features and whether the company was fraudulent was performed, and the results showed no correlation, so the features with more than 90% missing were removed. The ratio of normal samples to abnormal samples in the dataset was about 51:1, which was unbalanced and had a large gap between them, so the SMOTETomek integrated sampling method was applied to resample the original data, and finally a balanced dataset was obtained for model training.

3 Results and Discussion

3.1 Model Building

In this paper, five models, eXtreme Gradient Boosting, Logistic Regressiont, Light Gradient Boosting Machine, Adaptive Boosting and Random Forest, are selected and the evaluation metrics used are accuracy, detection rate, recall rate and F1 score. From Table 1, the XGBoost model has the best performance, with the four metrics of accuracy, accuracy, recall and F1 score up to 0.992 (Table 1).

In order to obtain more robust and reliable models, GridSearchCV with cross-validation is used to find the optimal parameters of the models and to make them achieve

Table 1. Performance of each model.

Model	Accuracy	Precision	Recall	F1 score
XGBoost	0.992	0.992	0.992	0.992
Logistic Regressiont	0.583	0.589	0.583	0.575
LightGBM	0.904	0.910	0.904	0.903
Adaboost	0.683	0.684	0.683	0.682
Random Forest	0.738	0.754	0.738	0.734

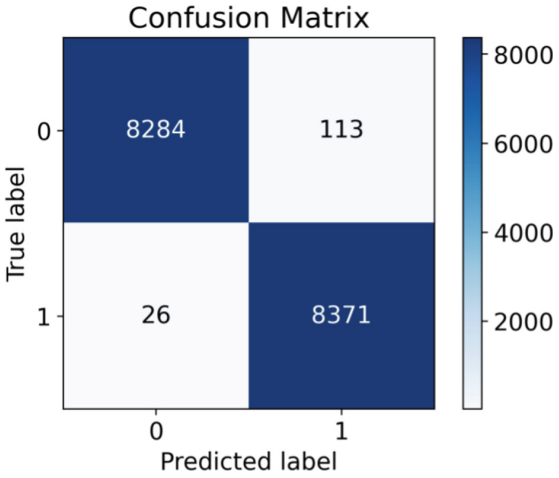

Fig. 1. Confusion matrix of XGBoost.

their optimal performance under their respective conditions (Fig. 1, 2, 3, 4 and 5). In this way, the generalization error of the models is evaluated and the approximate values of the generalization error of the models are obtained, and the performance of the five models is compared to visually find the model with the lowest error. Figures and Tables.

3.2 Model Evaluation

The ROC curve can be used to verify the overall predictive ability of the model. The number of normal and abnormal samples in the studied dataset is severely unbalanced, and considering issues such as the accuracy and recall of the model, the AUC value is introduced as an evaluation criterion for model training, which can take into account the accuracy and recall of the classification model in an integrated manner. Figure 2 shows the comparison of ROC curves of the five models (Fig. 6) and shows the comparison of AUC values of the five models (Fig. 7). An AUC value between 0.7 and 0.8 indicates a model with average prediction ability, and between 0.8 and 0.9 indicates a model with good prediction ability. The AUC values of XGBoost and LightGBM are 1.000 and

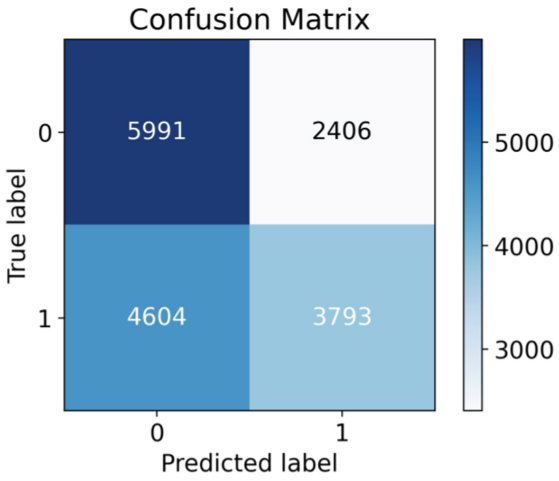

Fig. 2. Confusion matrix of Logistic Regression.

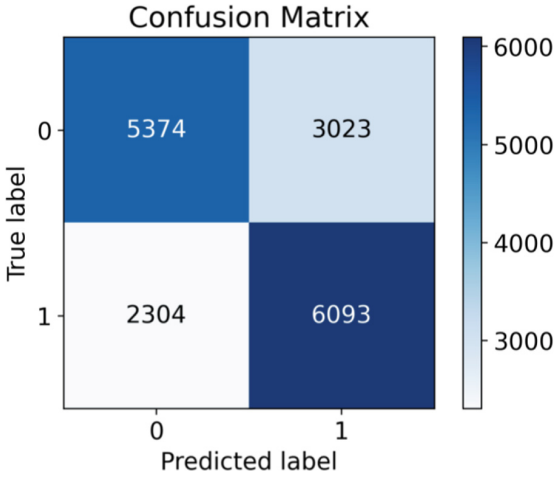

Fig. 3. Confusion matrix of LightGBM.

0.968, respectively, and their prediction ability is much higher than other models, so we select these two models to perform external validation.

When building the prediction model of financial fraud of listed companies. The features should be selected with full consideration of their significance in practice, and the psychological motives of financial fraudsters can be tapped by combining the needs and professional habits of auditors and other professionals. It is necessary to establish a scientific and reasonable feature screening system, so the indicators of the model must be updated with the development of the times, such as "registered capital", "number of employees" and other features, with the changes in the social environment and relevant policies and regulations, the actual reference significance The prediction results

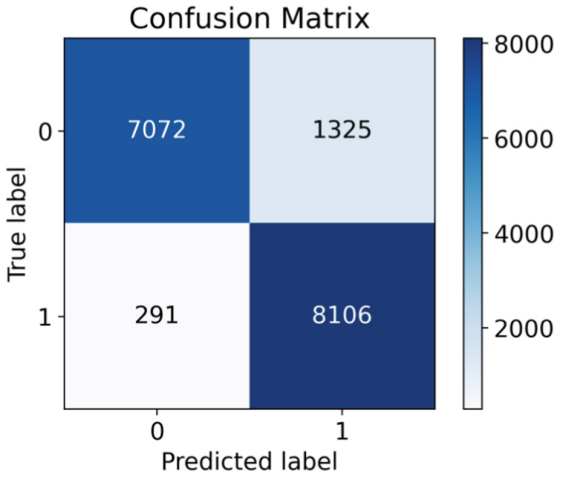

Fig. 4. Confusion matrix of Adaboost.

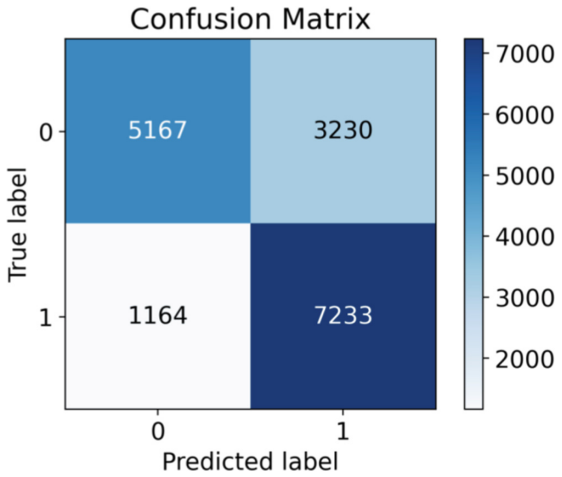

Fig. 5. Confusion matrix of Random Forest.

of models and studies containing such features are affected by the changes in the social environment and related policies and regulations, which are present in various countries and global capital markets. With the development of society and economy, more comprehensive and appropriate features that integrate data from different industries and different enterprises can improve the generalizability and accuracy of the model, thus making the financial fraud prediction model more practical.

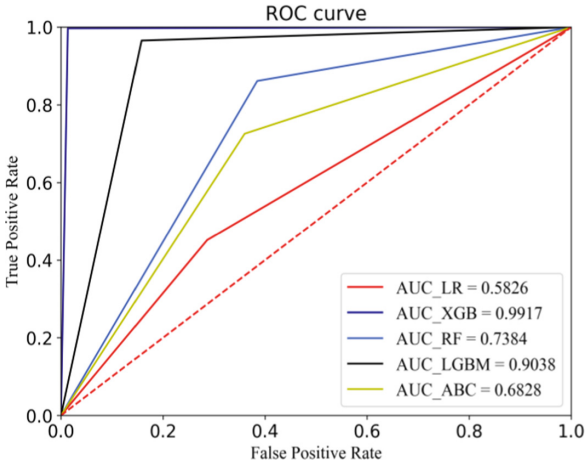

Fig. 6. Comparison of ROC curves for each model.

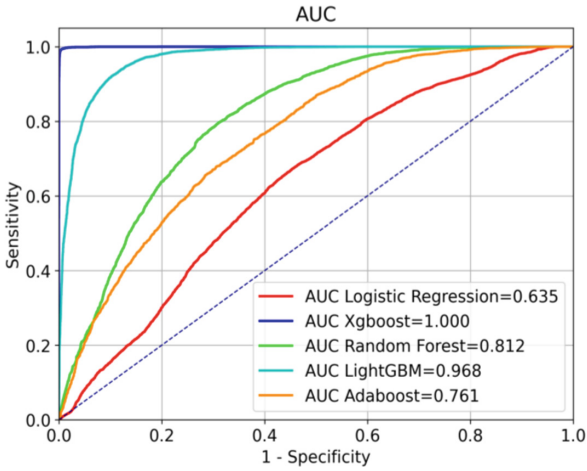

Fig. 7. Comparison of the AUC values of the models.

4 Conclusions

In order to cope with the features of financial fraud with hindsight, concealment and randomness, so as to establish a universal and accurate prediction model of financial fraud of listed companies, this paper uses the financial statements and data of listed companies from January 1, 2000 to December 31, 2020 in the Guotaian CSMAR database, solves the problem of data imbalance in model training, and establishes a universal and accurate prediction model of financial fraud of listed companies based on comprehensive consideration of major unexpected events, the innovative index system is established based on the influence of factors such as social environment and legal revision, and there are three rounds of model feature screening in this paper. The final screened new

metric system performed well for each model in both internal and external testing of the models, especially the XGBoost model with 98% accuracy in the results of external dataset validation. The feature screening system and model selection ideas of this study will improve the efficiency of subsequent research in this field and provide an important reference basis for capital market investors, analysts, auditors and relevant regulators in predicting and judging the financial fraud of Chinese listed companies.

Acknowledgment. This work was supported by the Fundamental Research Funds for the Central Universities 2022, and is also the phased achievement of the postgraduate innovation fund project of Huazhong University of Science and Technology "Application of Machine Learning in Financial Fraud of Listed Companies" Company: "Innovation Indicator System and Model Prediction" staged results (project number YCJJ202204013).

References

1. Blum, A. L., and P. Langley. "Selection of Relevant Features and Examples in Machine Learning." [In English]. Artificial Intelligence 97, no. 1–2 (Dec 1997): 245–71. https://doi.org/10.1016/S0004-3702(97)00063-5.
2. Hall, Mark. "Correlation-Based Feature Selection for Machine Learning." Department of Computer Science 19 (06/17 2000).
3. Shishkin, Alexander, Anastasya A. Bezzubtseva, Alexey Drutsa, Ilia Shishkov, Ekaterina Gladkikh, Gleb Gusev, and Pavel Serdyukov. "Efficient High-Order Interaction-Aware Feature Selection Based on Conditional Mutual Information." Paper presented at the NIPS, 2016.
4. Peng, H., F. Long, and C. Ding. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy." IEEE Trans Pattern Anal Mach Intell 27, no. 8 (Aug 2005): 1226–38. https://doi.org/10.1109/TPAMI.2005.159. https://www.ncbi.nlm.nih.gov/pubmed/16119262.
5. Yu, Lei, and Huan Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Vol. 2, 2003.
6. Lei, Xu, Yan Pingfan, and Chang Tong. "Best First Strategy for Feature Selection." Paper presented at the [1988 Proceedings] 9th International Conference on Pattern Recognition, 14 May-17 Nov. 1988 1988.
7. Robnik-Šikonja, Marko, and Igor Kononenko. "Theoretical and Empirical Analysis of Relieff and Rrelieff." Machine Learning 53, no. 1/2 (2003/10/01 2003): 23–69. https://doi.org/10.1023/a:1025667309714.
8. Dash, Manoranjan, and Huan Liu. "Consistency-Based Search in Feature Selection." Artificial Intelligence 151, no. 1–2 (2003/12/01/ 2003): 155–76. https://doi.org/10.1016/s0004-3702(03)00079-1. https://www.sciencedirect.com/science/article/pii/S0004370203000791.
9. Li, B., and M. Q. Meng. "Tumor Recognition in Wireless Capsule Endoscopy Images Using Textural Features and Svm-Based Feature Selection." IEEE Trans Inf Technol Biomed 16, no. 3 (May 2012): 323–9. https://doi.org/10.1109/TITB.2012.2185807. https://www.ncbi.nlm.nih.gov/pubmed/22287246.
10. Zhang, Junying, Shenling Liu, and Yue Wang. "Gene Association Study with Svm, Mlp and Cross-Validation for the Diagnosis of Diseases." Progress in Natural Science 18, no. 6 (2008/06/10/ 2008): 741–50. https://doi.org/10.1016/j.pnsc.2007.11.022. https://www.sciencedirect.com/science/article/pii/S1002007108001159.

11. Zhou, X., and D. P. Tuck. "Msvm-Rfe: Extensions of Svm-Rfe for Multiclass Gene Selection on DNA Microarray Data." Bioinformatics 23, no. 9 (May 1 2007): 1106–14. https://doi.org/10.1093/bioinformatics/btm036. https://www.ncbi.nlm.nih.gov/pubmed/17494773.
12. Mantas, Carlos J., Javier G. Castellano, Serafín Moral-García, and Joaquín Abellán. "A Comparison of Random Forest Based Algorithms: Random Credal Random Forest Versus Oblique Random Forest." Soft Computing 23, no. 21 (2019/11/01 2018): 10739–54. https://doi.org/10.1007/s00500-018-3628-5.
13. Kim, Youngsoon, Jie Hao, Tejaswini Mallavarapu, Joongyang Park, and Mingon Kang. "Hi-Lasso: High-Dimensional Lasso." IEEE Access 7 (2019): 44562-73.