




Article

A Financial Fraud Prediction Framework Based on Stacking Ensemble Learning

Shanshan Zhu ^{1,†}, Haotian Wu ^{2,†} , Eric W. T. Ngai ³, Jifan Ren ^{1,*} , Daojing He ^{4,*}, Tengyun Ma ⁴ and Yubin Li ¹ ¹ School of Economics and Management, Harbin Institute of Technology, Shenzhen 518000, China; 23b357002@stu.hit.edu.cn (S.Z.); liyubin@hit.edu.cn (Y.L.)² Faculty of Computer, Harbin Institute of Technology, Harbin 150001, China; 23b936025@stu.hit.edu.cn³ Department of Management and Marketing, The Hong Kong Polytechnic University, Hong Kong 00852, China; eric.ngai@polyu.edu.hk⁴ School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518000, China; tengyunma@stu.hit.edu.cn

* Correspondence: renjifan@hit.edu.cn (J.R.); hedaojing@hit.edu.cn (D.H.)

† These authors contributed equally to this work.

Abstract: With the rapid development of the capital market, financial fraud cases are becoming increasingly common. The evolving fraud strategies pose significant threats to financial regulation, market order, and the interests of ordinary investors. In order to combine the generalization performance of different machine learning methods and improve the effectiveness of financial fraud prediction, this paper proposes a novel financial fraud prediction framework based on stacking ensemble learning. This framework, based on data from listed companies, comprehensively considers financial ratio indicators and non-financial indicators. It uses the stacking ensemble technique to integrate numerous base models of machine learning algorithms for predicting financial fraud. Furthermore, the proposed framework has high versatility and is suitable for various tasks related to financial fraud prediction, addressing the problem of model selection difficulties in previous research due to different scenarios and data. We also conducted case studies on specific companies and industries, confirming the significant interpretability and practical applicability of the proposed framework. The results show that the recall rate and Area Under Curve (AUC) of our framework reached 0.8246 and 0.8146, respectively, surpassing mainstream machine learning models such as XGBoost and LightGBM in existing studies. This research study is of great significance for predicting the increasing number of financial fraud cases, providing a reliable tool for financial regulatory institutions and investors.

Keywords: financial statement; fraud prediction; machine learning; stacking model



Citation: Zhu, S.; Wu, H.; Ngai, E.W.T.; Ren, J.; He, D.; Ma, T.; Li, Y. A Financial Fraud Prediction Framework Based on Stacking Ensemble Learning. *Systems* **2024**, *12*, 588. <https://doi.org/10.3390/systems12120588>

Academic Editor: Vladimír Bureš

Received: 15 November 2024

Revised: 6 December 2024

Accepted: 19 December 2024

Published: 23 December 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the development of modern enterprises, financial frauds are becoming increasingly common around the world. According to the 2020 Global Occupational Fraud and Abuse of Power Investigation Report by the Association of Certified Fraud Examiners (ACFE) [1], three dominant forms of fraud prevail: corruption, asset misappropriation, and financial statement fraud. Financial statement fraud is typically the most detrimental form of fraud, causing severe negative consequences for companies and stakeholders. As a result, it has garnered significant attention and research. Financial statement fraud is generally defined as the intentional misrepresentation of financial information in a manner contrary to generally accepted accounting principles, leading to significant omissions or false statements. Within the realm of financial statement fraud, various distinct types emerge. According to the U.S. Securities and Exchange Commission (SEC) [2,3], improper recognition of revenue and fabricated revenue are the most prevalent. Meanwhile, the disclosures of the China Securities Regulatory Commission (CSRC) reveal that false assets and

fictitious profits account for 60% of fraud cases. Therefore, in order to maintain financial market stability and strengthen regulation and compliance, it is necessary to find efficient methods for identifying financial fraud. This has important reference value for regulators, management, auditors, and investors.

In theoretical research on financial fraud, the fraud triangle theory proposed by Cressey [4] and the fraud octagon theory proposed by Imoniana et al. [5] provide different theoretical frameworks to explain and understand the mechanisms behind fraud behavior. Cressey proposed the fraud triangle theory in 1953, which explains the occurrence of financial fraud through three key factors: pressure, opportunity, and rationalization. The theory emphasizes that fraud is likely to occur at the intersection of these three factors. As financial fraud has become more complex and diversified, Imoniana et al. proposed the fraud octagon theory in 2016, which extends and deepens Cressey's triangle model. Their model adds multiple dimensions, including pressure, opportunity, rationalization, arrogance, greed, lawbreaker, sycophant, and internal controls. Unlike the traditional fraud triangle theory, the fraud octagon theory attempts to incorporate more social, psychological, and organizational factors, offering a more multi-dimensional framework for understanding the occurrence of fraud.

Due to the diverse forms and covert methods of financial statement fraud, manual detection by regulators is challenging [6]. Some investigations have revealed that auditors can only unveil a small portion of such malfeasance [1,7]. The research team led by Meredith et al. [8] emphasized the necessity of collaboration between decision support systems and auditors. Thanks to the advancements in machine learning and artificial intelligence models in pattern recognition, automated fraud detection in financial statements has received increasing research attention [9–11]. Financial fraud detection refers to the process of identifying and discovering financial fraud that has already occurred. Past research has largely focused on financial fraud detection, which is a passive, retrospective approach. However, if we could identify early warning signals and anomalous patterns from a company's historical data that may indicate a heightened risk of future fraud, it could help concentrate limited auditing and compliance resources on the highest-risk areas, thereby improving the efficiency and effectiveness of fraud prevention efforts. This paper focuses on the task of financial fraud prediction, aiming to proactively forecast the potential fraud risks that enterprises may face in the future.

Machine learning methods can rapidly analyze the features in financial data and provide accurate predictions. When combined with statistical methods, they can quantify the effectiveness of financial features and provide recommendations for fraud prediction and financial auditing. On the other hand, although artificial intelligence methods utilizing deep learning also achieve good results and robustness, they often lack interpretability, making it difficult to quantify the role of individual features in prediction. Grinsztajn et al. [12], through their research, discovered that tree-based machine learning models can even outperform neural networks in processing tabular data. Therefore, machine learning methods are capable of handling the complex associations and influences between financial and non-financial data, and they can analyze the importance and statistical significance of specific features. This makes them well-suited for financial fraud prediction tasks. In the early stages of machine learning, financial fraud detection mainly relied on some classic supervised learning methods, such as Logistic Regression (LR), Decision Trees, and C4.5 [13–15]. Ravisankar et al. [16] evaluated the performance of various machine learning methods in detecting fraud by using data collected from 202 listed Chinese companies. These studies provided in-depth analysis from various perspectives, including fraud proportion, the impact of financial features, and comparisons of different models. As the field of machine learning developed, ensemble learning methods gradually gained attention, playing an important role in improving detection accuracy. For example, Hassanniakalager et al. [17] proposed an ensemble learning method based on LogitBoost, which utilized financial data extracted from the Compustat database, and achieved good results. Bao et al. [9] used the RUSBoost method [18], which constructs sub-training sets through random under-sampling

(RUS) and combines it with the Boost ensemble learning method, effectively improving the model's performance on imbalanced data, thereby enabling more accurate detection of financial fraud.

Currently, research on financial fraud primarily focuses on how to detect abnormalities in financial statements. In 1966, Beaver [19] first proposed the use of financial ratio indicators to identify the risk status of publicly listed companies. Since then, financial indicators have been widely considered as one of the factors for identifying corporate risk and have become a major component in financial fraud detection. Current financial fraud detection primarily relies on three data sources: basic financial data indicators, textual information from financial statements (often used in conjunction with financial data), and other company-specific characteristics (non-financial data). Financial data is the most direct reflection of a company's financial status, but companies may manipulate the data and text in financial statements to conceal fraudulent activities. Therefore, combining non-financial data (which is harder to manipulate) can further reveal the possibility of financial fraud. However, previous studies rarely focus on all three data sources simultaneously, typically concentrating on one or two sources. For example, some early studies [20–22] were based on fraud determinations and corresponding financial information from the Accounting and Auditing Enforcement Releases (AAERs) published by the U.S. Securities and Exchange Commission (SEC), using neural networks (NNs) for classification and detection. Cecchini et al. [23] studied 23 financial data indicators in the context of U.S. companies' financial data by using support vector machine (SVM) for fraud detection. Hassanniakalager et al. [17] used financial data extracted from the Compustat database and designed an ensemble learning method based on LR and LogitBoost, achieving good results. Some studies have also used non-financial data to assist in fraud detection [24,25]. For instance, Kim et al. [26] analyzed the role of certain non-financial indicators in their detection model, such as abnormal employee turnover, CEO compensation, and stock returns, achieving significant complementary results. Hajek and Henriques [27] combined financial data with linguistic attributes extracted from annual reports. Brown et al. [28] introduced annotated thematic content and linguistic features from financial statements, showing that both contribute to the identification of fraudulent activities.

Previous studies [3,9,10] have shown that machine learning-based methods can handle the correlation between financial and non-financial data, and analyze their importance and statistical characteristics. However, these studies often overlook the comprehensive construction of various stages in machine learning, including feature analysis, model selection, and parameter optimization. Therefore, they fail to establish a comprehensive workflow for detecting and analyzing specific fraud cases. Research has shown that feature selection plays an important role in the final result [27,29,30]. In addition, existing studies usually rely on manual selection methods, and there are relatively few approaches that combine model-based feature selection. Moreover, parameter optimization is also a crucial process in model construction [31], but some studies have overlooked this step.

Some studies [32–34] compared different models and achieved more accurate classification performance. However, they also pointed out the limitations of certain models, and the performance of the same model can vary significantly based on different accounting standards, regions, and financial indicators. The most effective model in different scenarios may differ, making it difficult to choose an appropriate model in practical applications [27,35]. By taking into account these existing challenges and aiming to enhance the practical value of fraud prediction methods and provide forward-looking references for stakeholders such as investors, this study establishes a general financial fraud prediction framework based on stacking ensemble learning.

We present exemplar studies from 2010 to 2024 in Table 1.

Table 1. Summary of prior research on fraud detection.

Authors	Data	Key Model	Key Finding
Cecchini et al. (2010) [23]	Financial data	SVM	A specific kernel function, which enhances the detection capability of SVM, was developed for the financial domain.
Dechow et al. (2011) [13]	Financial data/ non-financial indicators	LR	A detailed analysis was conducted on the differences between fraudulent companies and non-fraudulent companies across various data types.
Perols et al. (2011) [14]	Financial data/ non-financial indicators	LR/SVM	Logistic regression and SVM performed well. Only six predictors examined are consistently selected and used by different classification algorithms: auditor turnover, total discretionary accruals, Big 4 auditor, accounts receivable, meeting or beating analyst forecasts, and unexpected employee productivity.
Ravisankar et al. (2011) [16]	Financial data/ financial ratio indicators	GP/PNN	PPN outperformed all techniques without feature selection. GP and PNN were superior to other techniques when feature selection and accuracy were roughly equivalent.
Throckmorton et al. (2015) [36]	Conference call audio files/financial data	GLRT	It combines financial numbers, linguistic behavior, and non-verbal vocal cues. If each feature category provides independent, complementary information about financial fraud, then a combination of features from these categories may improve detection performance beyond what can be achieved by any single feature category alone.
Hajek et al. (2017) [27]	Financial data/ annual report text	BNN	They found that ensemble methods outperformed the remaining methods in terms of true-positive rate (fraudulent firms correctly classified as fraudulent). In contrast, Bayesian belief networks (BBNs) performed best on non-fraudulent firms (true-negative rate).
Yao et al. (2019) [37]	Financial data/ non-financial indicators	SVM	The experimental results showed that the SVM data mining technique had the highest accuracy across all conditions; after using stepwise regression, 13 significant variables were screened, and the classification accuracy of almost all data mining techniques was improved.
Craja et al. (2020) [30]	Financial data/ annual report text	Hierarchical attention network	The paper proposes an approach for detecting statement fraud through the combination of information from financial ratios and managerial comments within corporate annual reports. The model captures both the content and context of managerial comments, which serve as supplementary predictors along with financial ratios in the detection of fraudulent reporting.
Bao et al. (2020) [9]	Financial data	RUSBoost	The paper introduces a new performance evaluation metric commonly used in ranking problems that is more appropriate for the fraud prediction task.
Papik et al. (2021) [25]	Financial data/ financial ratio indicators	DT/RF	The article indicates that by using financial ratios and applying the random forest method, unintentional accounting errors can be detected with very high accuracy.
Hassanniakalager et al. (2022) [17]	Financial ratio indicators	LogitBoost	The model is superior to other models in predicting fraudulent behavior beyond the current accounting period. It relies on fewer predictive variables compared with what was used in previous ML research, thereby minimizing concerns related to multicollinearity and potential overfitting associated with machine learning methods.
Wang et al. (2023) [11]	Financial data/ annual report text	RCMA	A novel attention-based multimodal deep learning approach, called RCMA, was proposed. A new loss function called Focal and Consistency Loss (FCL) was designed.

Table 1. Cont.

Authors	Data	Key Model	Key Finding
Duan et al. (2024) [38]	Financial ratio indicators/ non-financial indicators/ annual report text	LDA/Balanced Random Forest	A pre-fraud risk index was proposed. This study redefined fraud detection as an ongoing endeavor rather than a retrospective event, thus enabling managers and stakeholders to reconsider their operation decisions and reshape their entire operation processes accordingly.
Khaksari et al. (2024) [39]	Financial data/ financial ratio indicators	Beneish model/ Spathis model	The coefficients of the Beneish and Spathis models were adjusted by using logistic regression, and the predictive performance of the adjusted model in detecting fraudulent financial reporting was studied. The adjusted version of the model demonstrated superior predictive performance.
Rahman et al. (2024) [40]	Financial ratio indicators	fraud triangle theory	The results show that leverage and liquidity ratios positively affect fraud detection, whereas return on net equity, audit size, and independent director percentage negatively affect fraud detection.
Bhattacharya et al. (2024) [41]	Financial data/ annual report text	BERT	Based on the BERT-Base model, two models were obtained through fine-tuning training: BERTfirst and BERTlast. BERTfirst was trained on the first 512 tokens of each MD&A text sample, while BERTlast was trained on the last 512 tokens of each MD&A text sample. Both models demonstrated good performance.

This study used data from listed companies in China from 2013 to 2021 as an example and employs four violation indicators from the CSMAR database, including “fabricated profits”, “misstated assets”, “false records (misleading statements)”, and “general accounting improprieties” as fraudulent samples for model training and prediction. In terms of feature selection, considering the three major data sources, a preliminary selection of 425 financial indicators and 22 non-financial indicators (such as the shareholding ratio of the largest shareholder) from all companies was made. Most mainstream machine learning models proposed in previous research [9,35,42] were trained and compared, and the model parameters were optimized based on cross-validation results. Subsequently, several of the most effective machine learning models were integrated by using stacking ensemble learning for further learning and prediction. The experimental results demonstrate that the recall rate and AUC of this framework reached 0.8246 and 0.8146, respectively, surpassing other existing machine learning models. The comprehensive results also indicate that this framework enhances fraud prediction effectiveness through processes such as feature selection, parameter optimization, and model integration. Finally, the framework was applied to various financial fraud-related tasks, demonstrating its good generality. Real case analyses were conducted to further transform the machine learning models into practical solutions with strong interpretability.

This study makes several key contributions as follows: (1) A comprehensive fraud prediction framework based on stacking ensemble learning is proposed; it considers feature analysis, parameter optimization, and model selection, providing a complete workflow for fraud prediction in accounting. (2) The proposed framework exhibits high generality and can be applied to various tasks related to financial fraud prediction. It outperforms single machine learning models in different application environments, addressing the challenge of model selection difficulty in previous research due to varying scenarios. (3) Unlike past research on financial fraud detection, this study’s forward-looking prediction approach aims to uncover early warning signals and anomalous patterns from a company’s historical data that may indicate heightened future fraud risks. This approach can help regulatory organizations concentrate their limited auditing and compliance resources on the highest-risk areas, thereby improving the efficiency of fraud prevention efforts.

The remaining sections of this paper are as follows: Section 2 systematically introduces the machine learning framework based on the stacking ensemble model proposed in this study. Section 3 presents the experimental results. Section 4 further investigates and discusses the framework by incorporating real-world scenarios. Finally, Section 5 provides a summary of the entire paper.

2. Materials and Methods

The basic structure of the proposed financial fraud prediction framework based on stacking learning is illustrated in Figure 1. It mainly consists of modules such as data collection and preprocessing, feature selection, model optimization and screening, model stacking ensemble, and performance evaluation. Each module will be described in detail in the following subsections. Considering the characteristics of the stacking ensemble model, it is necessary to train and optimize each base learner before the ensemble. By combining the predictions of multiple learners, the stacking ensemble model can capture more data features and patterns, resulting in more accurate prediction results.

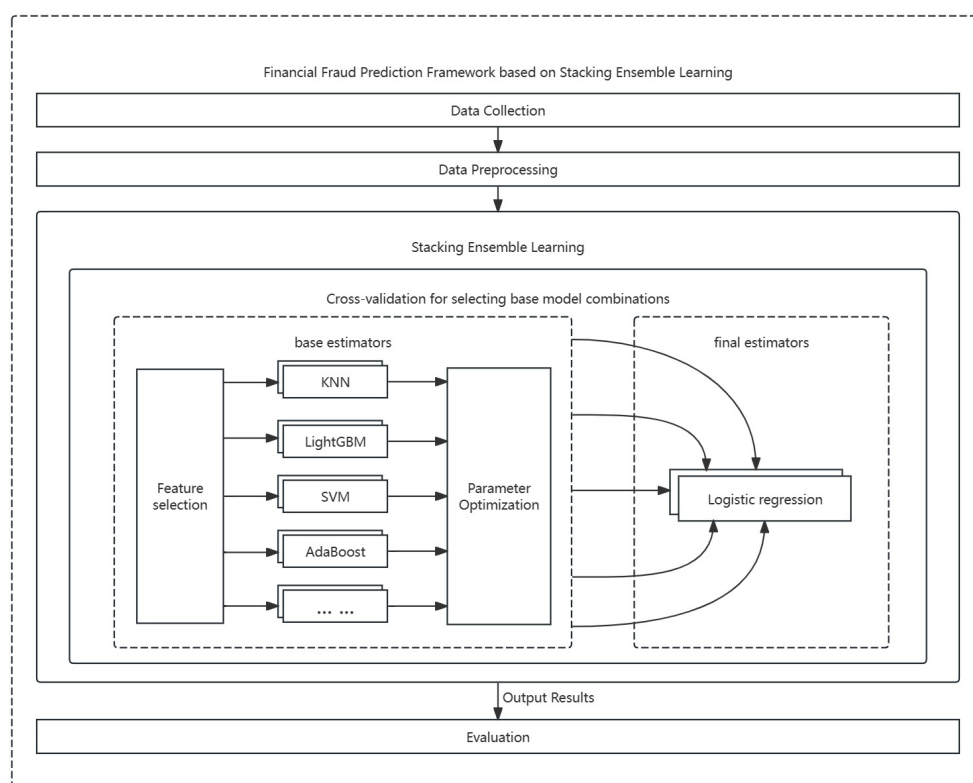


Figure 1. Schematic diagram of the proposed financial fraud prediction framework.

2.1. Data Collection and Processing

This paper used fraud information disclosed by Chinese securities regulatory agencies. Specifically, we obtained these data from the “Violation Information Summary Table” section of the CSMAR database, which contains a series of information on the violations of listed companies. The violation types of listed companies cover a wide range of aspects, including insider trading, general accounting mishandling, appropriation of company assets, material omissions, fraudulent listing, delayed disclosure, fictitious profits, false records (misleading statements), misstatement of assets, and so on. We primarily focused on financial fraud manifested through false statements regarding profits or assets, in addition to potential accounting treatment issues. Such false statements of assets and profits can be reflected in financial indicators, indicating that the enterprise has hidden operational risks. Therefore, listed companies with the following four types of violations were selected

as financial fraud companies: “fictitious profits”, “false presentation of assets”, “false entries (misleading statements)”, and “general accounting mishandling”. Compared with certain studies, our criteria can cover a wider range of fraud cases and effectively avoid the inclusion of many non-fraud cases when the scope is too broad.

In terms of financial data acquisition, the financial data of listed companies are sourced from the CSMAR database, which includes the annual reports of companies listed on the Shenzhen Stock Exchange and the Shanghai Stock Exchange. This study focused only on A-type annual reports. Data tables were obtained for 11 major categories of financial ratio indicators, including ratio structure, risk level, per-share indicators, disclosed financial indicators, solvency, operational capacity, cash flow analysis, development capacity, dividend distribution, profitability, and relative value indicators. Due to the special nature of financial businesses, the financial statements of financial sector companies need to consider additional factors and requirements. Therefore, this study excluded financial sector companies from the dataset. Typically, listed companies are penalized for financial fraud with a certain time lag, often occurring two years or even a decade after the actual occurrence of fraudulent activities. Therefore, this study traced back to the year in which the penalized companies actually engaged in financial fraud to determine whether the company had committed fraud in that year. Our analysis primarily centered around the period from 2013 to 2021 to ensure that as many accurate fraud cases as possible were included in our sample. The overall situation of financial fraud from 2013 to 2021 is shown in Table 2.

Table 2. Distribution of fraud cases.

Year	Fraud Number	Non-Fraud Number	Total Number
2013	163	2152	2315
2014	151	2274	2425
2015	207	2317	2524
2016	226	2608	2834
2017	333	3057	3390
2018	440	3032	3472
2019	415	3247	3662
2020	393	3700	4093
2021	358	4181	4539
Total	2686	26,568	29,254

As shown in Table 2, there exists a notable disparity between the number of fraudulent companies and non-fraudulent companies, with the former constituting approximately 9.1% of the total firms from 2013 to 2021. The serious sample imbalance tends to bias the model to classify the whole sample into the non-fraud category. To solve the sample imbalance problem, the random under-sampling method was used to balance the training set, and the majority class samples were randomly deleted, so that the fraud samples formed a definite proportion with non-fraud data. Then, we trained the model with the balanced training set and made prediction and evaluations on the test set.

We classified the variables into two groups: financial indicators and non-financial indicators. First, we used financial ratios from the financial statements as input variables because they possess objectivity and are easily obtainable. Unlike the absolute value of financial indicators, financial ratio indicators offer better comparability across enterprises of varying sizes. So, a total of 425 financial ratio indicators from the CSMAR database were selected as variables. Second, non-financial indicators provide a well-rounded reflection of internal and external environments, corporate governance, and other pertinent information about listed companies. Therefore, we also introduced 22 non-financial indicators (e.g., the readability score of the financial report, the proportion of shares held by the largest shareholder, historical violation records, internal control index, etc.) as input. We list the definition of non-financial indicators in Table 3.

Table 3. Non-financial indicator definitions.

Index	Variable Name	Variable Definition
1	Readability Score	(Average word count per clause + Proportion of adv. and conj. per sentence in the annual report)/2
2	Tone Score	(Positive words count—Negative words count)/(Positive words count + Negative words count)
3	ARF	Average ratio of sentiment-containing sub-text blocks to total sentiment words.
4	Largest Holder Rate	Proportion of shares held by the largest shareholder
5	Top Ten Holders Rate	Proportion of shares held by the top ten shareholders
6	Ownership	1 for state-owned enterprises, else 0
7	Board Size	Total number of board members
8	Proportion of Independent Directors	Proportion of independent directors to directors
9	Duality	1 if chairman and CEO are the same person, else 0
10	Auditor	1 if the auditor is from a Big Four accounting firm, else 0
11	Audit Committee	1 if the company has an audit committee, else 0
12	Receipt of Violation Notice in Current Year	1 if the company has received a violation notice in the current year, else 0
13	Receipt of Violation Notice in Previous Year	Set to 1 if the company has received a violation notice in the previous year, otherwise 0
14	Number of Regulatory Letters Received in Current Year	Quantity of regulatory functions received by the company in the current year
15	Number of Regulatory Letters Received in Previous Year	Quantity of regulatory functions received by the company in the previous year
16	Internal Control Index	Internal control index published by Shenzhen Dibo Company
17	Strategic Hierarchy Index	Strategic level index published by Shenzhen Dibo Company
18	Operational Hierarchy Index	Operational level index published by Shenzhen Dibo Company
19	Report Reliability Index	Report reliability index published by Shenzhen Dibo Company
20	Legal Compliance Index	Legal compliance index published by Shenzhen Dibo Company
21	Asset Security Index	Asset security index published by Shenzhen Dibo Company
22	Internal Control Disclosure Index	Internal control disclosure index published by Shenzhen Dibo Company

To ensure data quality, we excluded variables with missing values exceeding 25% from the initial set of 425 financial indicators, resulting in a refined collection of 407 indicators. For the remaining missing values, we used the median imputation technique to fill them in. Additionally, for all numerical data, apart from binary data with values of 0 and 1, standardization was performed. Finally, based on the comprehensive list of violation information, each sample was assigned a fraud class label. When the company is involved in fraud in the next year, the sample is labeled with 1; otherwise, the sample is labeled with 0.

2.2. Feature Selection

Feature selection plays a crucial role in machine learning, as it enhances model performance, reduces dimensionality, accelerates training speed, and improves the interpretability of data. This study compared three feature selection methods, namely, Spearman's coefficient, mutual information, and analysis of variance. Spearman's coefficient is employed to measure the monotonic relationship between two variables, making it suitable for handling nonlinear relationships. During feature selection, the Spearman's coefficient is calculated for each feature for the target variable. A higher coefficient indicates a strong association

between the feature and the target variable, making it an important feature to consider. Mutual information measures the correlation and dependency between two variables. For each feature, the mutual information with the target variable is computed. Features with high mutual information values indicate a strong correlation and information sharing with the target variable, thus serving as important features. Analysis of variance is used to compare the mean differences between different groups, particularly in the context of feature selection for classification problems. It calculates the variance of each feature among different categories or groups and determines the statistical significance through the analysis of the variance. Features with higher variance and significant differences are considered important features.

We compared the three feature selection methods and ranked the computed results, and the top ten features are shown in Table 4. It is evident that among all the features, “Earnings per Share”, “Earnings Surplus per Share”, and specific “index” features hold a relatively substantial influence on fraud prediction.

Table 4. Sorted feature selection results.

Spearman’s Coefficient	Mutual Information	Analysis of Variance
Legal Compliance Index	Largest Holder Rate	Legal Compliance Index
Earnings per Share TTM2	Earnings Surplus per Share	Internal Control Index
Earnings per Share2	Earnings Surplus per Share2	Report Reliability Index
Earnings per Share	Capital Surplus per Share2	Asset Security Index
Earnings per Share TTM	Capital Surplus per Share	Operational Level Index
Earnings per Share2	Internal Control Index	Strategic Level Index
Earnings per Share TTM2	Legal Compliance Index	History
Earnings per Share1	Operational Level Index	Earnings per Share TTM2 attributable to Parent Company
Earnings per Share TTM1	Report Reliability Index	Earnings per Share2 attributable to Parent Company
Earnings per Share4	Earnings per Share2 attributable to Parent Company	Earnings per Share2

Moreover, we intended to analyze the feature importance in different machine learning models for further model stacking ensemble. For each base estimator, we applied the three mentioned feature selection methods individually to conduct feature screening. With each method, we trained and tested the models by using the top 9.1% to 100% ranked features. Through cross-validation, we determined the optimal method and its corresponding feature selection ratio that yielded the highest AUC metric for the model. Table 5 shows the specific results.

Table 5. Feature selection.

Method	Feature Selection	Ratio
SVM	Analysis of variance	0.42
Random Forest	Analysis of variance	0.46
KNN	Analysis of variance	0.32
AdaBoost	Analysis of variance	0.56
XGBoost	Spearman’s Coefficient	0.96
LightGBM	Spearman’s Coefficient	0.96
ExtraTrees	Analysis of variance	0.44

For each model, we compared the results of the three feature selection methods to obtain the best choices. Then, based on these results, we further selected the base model for the framework.

2.3. Model Construction

Existing financial and non-financial characteristics can reflect a company's operating conditions through numbers, which is the most intuitive factor for determining financial fraud in companies. Machine learning methods are good at analyzing the data directly and can give appropriate explanations based on the statistical relationships of the data. Therefore, the fraud prediction framework in this paper adopts the supervised machine learning algorithm.

In building the fraud prediction model, we used traditional machine learning methods, i.e., logistic regression (LR), support vector machine (SVM), K-nearest neighbor (KNN), random forest, AdaBoost, XGBoost, LightGBM, and Extratrees. Moreover, a more advanced machine learning model ensemble method, stacking, was also introduced into this framework. The stacking model further improves model performance and generalization ability by feeding prediction results from multiple base estimators as new features into final estimators.

Logistic regression is the simplest and most popular method in previous studies [35,43,44]. It maps the input features to a fraud probability (between 0 and 1) by transforming a linear combination W on input features X with the logistic function (Sigmoid) as

$$P(y = 1|X) = \frac{1}{1 + e^{-(W^T X + b)}} \quad (1)$$

Here, y represents the target variable, which is the classification label. b is a constant that represents the bias term, used to adjust the model's output and improve prediction accuracy.

Support vector machine (SVM) is a one of the most widely used machine learning algorithms for classification and regression in the financial field [30,33,45]. SVM intends to find an optimal hyperplane (or hyperplane in higher dimensional space) denoted by

$$W^T X + b = 0 \quad (2)$$

which separates data points of different categories and maximizes the distance from the hyperplane to the nearest data point. Here, b is also the bias term. Its role is to adjust the position of the decision boundary, allowing the hyperplane to better separate data from different classes.

The K-nearest neighbor (KNN) algorithm can also be used for classification [33,37,46]. It determines the voting of an unknown data point by measuring the distance based on the features of the sample data point and assigns the point as the most common class among its k -nearest neighbors.

To reduce the bias and variance in machine learning, boosting ensemble learning was proposed in [47]. Subsequently, many improvement studies were conducted to further improve the ensemble performance.

Adaboost [48] assigns weights to each training sample and iteratively trains multiple weak classifiers, each of which is focused on the samples that were misclassified in the previous round, and ultimately combines these weak classifiers into one strong classifier. XGBoost [49] is an improved gradient-boosting tree algorithm that combines gradient boosting and regularization techniques to improve the performance and generalization ability of the model. Furthermore, LightGBM [50] aims to reduce memory consumption and increase training speed for machine learning tasks with large-scale datasets and high-dimensional features. These methods have been widely used as base and comparative models in recent years due to their generalization ability [27,30,43].

Moreover, some methods based on decision trees also showed strong classification capabilities. Random Forest, proposed by [51], consists of multiple decision trees, each of which is randomly generated. Then, the predictions of the individual trees are integrated by

voting (for fraud classification) to obtain a more robust and accurate model. This method was the first to apply decision trees to ensemble learning and achieved comparable results in previous studies [32,52]. ExtraTrees [53] employs more randomness in the construction of each tree to make it more robust to noisy data, achieves faster training speeds, and generally performs well in accuracy [34].

Overall, these algorithms exhibit different results in addressing various tasks. However, past research has often overlooked the potential of integrating them together. In the face of complex machine learning problems, a single model may not be able to fully capture the complexity and diversity of the data.

In recent years, ensemble methods have become an effective strategy for improving model performance and prediction accuracy, particularly when tackling complex machine learning problems. By combining the predictions of multiple base models, ensemble learning often achieves more accurate and robust overall models. Among these methods, stacking is a highly efficient and widely adopted technique. It integrates the predictions of diverse base estimators (weak learners) and utilizes a final estimator to aggregate these predictions, resulting in a more accurate final output.

Stacking is not only frequently employed as a blending technique by winning teams in competitions but is also regarded as a viable artificial intelligence solution in practical industrial applications. As a powerful ensemble method, stacking combines strong model performance, enhanced interpretability, and applicability to complex data. It is, therefore, considered one of the most practical and innovative approaches in the field of machine learning.

The primary objective of stacking is to feed the predictions of various base models into a meta-model, effectively leveraging the strengths of each. Once the meta-model is trained, it determines the optimal way to weight and combine these predictions into a final output. By aggregating the outputs of diverse base models, such as decision trees, support vector machines, and neural networks, stacking effectively captures the strengths of each while overcoming the limitations of individual models. This process reduces bias, enhances predictive capability, and improves the overall generalization ability of the model. Moreover, stacking mitigates the risk of overfitting since each base model may perform well in different aspects of the data, and the method leverages these strengths to enhance the generalization performance of the ensemble. With its flexible architecture, stacking is well suited for a wide range of machine learning tasks, including classification, regression, and time-series analysis.

In our framework, we combined the aforementioned base algorithms to construct a more powerful stacking model and compared it with these mainstream machine learning algorithms to demonstrate the superiority of the ensemble model.

2.4. Performance Evaluation

To comprehensively evaluate the performance of the proposed framework, we selected the following performance indicators: accuracy, recall, and AUC. This section introduces the calculation methods of each indicator.

First, for binary classification problems, the test set samples can be divided into four categories according to the combination of the true category and the predicted category: true positive (True Positive), false positive (False Positive), true negative (True Negative), and false negative (False Negative), denoted by TP, FP, TN, and FN, respectively; the confusion matrix is obtained according to the classification category.

Furthermore, according to the confusion matrix, commonly used performance indicators can be calculated: accuracy, recall, and AUC. Accuracy represents the proportion of correctly classified samples in the total samples, recall represents the proportion of true-positive samples in the actual true samples, and AUC is the area under the ROC curve. The larger the value of the above indicators, the better the prediction effect. The relevant calculation formulas are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

AUC (Area Under Curve) is a commonly used metric for evaluating the performance of binary classification models, representing the area under the ROC (Receiver Operating Characteristic) curve. The ROC curve depicts model performance with the false-positive rate (FPR) on the x-axis and the true-positive rate (TPR) on the y-axis. The FPR represents the proportion of negative samples incorrectly classified as positive, while the TPR represents the proportion of positive samples correctly classified. Essentially, AUC measures the ranking ability of the classifier, with a value closer to 1 indicating better classification performance. In simpler terms, AUC is the probability that when one sample is randomly selected from the positive class (1) and another from the negative class (0), the classifier will assign a higher score to the positive sample than the negative sample. The probability that the positive sample is ranked higher than the negative sample equals AUC. AUC does not require the setting of a threshold for predicted probabilities, making it an effective metric for assessing classifier performance, even when the samples are imbalanced.

The advantage of AUC lies in its focus on the ranking of classification results rather than specific probability values or threshold settings, which avoids the impact of threshold variations on model evaluation. Additionally, AUC is insensitive to the class distribution, making it particularly suitable for scenarios with imbalanced sample distributions. Therefore, in the context of financial fraud prediction, the importance of AUC is especially significant, as it better reflects the performance differences between models.

On the other hand, in the task of financial fraud prediction, using recall as a model evaluation metric, in addition to accuracy, has significant advantages. Missed detection (failure to identify fraudulent cases) can lead to severe financial losses and legal consequences. Recall measures the proportion of correctly identified fraudulent cases to all actual fraudulent cases, which is crucial to ensuring the maximum detection of genuine fraud. This helps regulatory authorities take effective actions to prevent potential losses. A high recall rate means that the model can effectively capture more fraudulent cases, thereby reducing the risk of missed detections. Especially when financial fraud events are rare, recall provides a more accurate reflection of the model's detection capability. Therefore, recall not only enhances the model's practicality and reliability but also serves as an important metric for ensuring financial security and compliance.

Therefore, we first consider AUC and then simultaneously pay attention to accuracy and the recall rate as reference indicators for model evaluation.

3. Results

3.1. Base Estimator Training and Parameter Optimization

When tuning the parameters of the base estimators, we took the higher AUC. The Bayesian optimization method was used on the training set through cross-validation to select the optimal parameters. Table 6 displays the parameter tuning results for seven models (SVM, Random Forest, KNN, AdaBoost, XGBoost, LightGBM, and Extratrees) following cross-validation on the 2013–2016 data.

3.2. Stacking Ensemble

We used seven pre-trained base estimators to construct the stacking ensemble model. K-fold cross-validation was performed on each basic model, and the cross-validation results were combined to form the input features of the logistic regression model (final estimator). The training process of the stacking model is shown in Figure 2. Then, we selected the combination of different base estimators to perform stacking model ensemble. By comparing the AUC results of all model combinations, the Random Forest, AdaBoost, XGBoost, LightGBM, and ExtraTrees models were finally selected as the base estimators,

while the logistic regression model served as the final estimator. In addition, we performed grid search tuning on the logistic regression model, resulting in an adjustment of the inverse regularization strength to 30.

Table 6. Base learner parameters.

Method	Parameters
SVM	C: 9.75, gamma: 0.001
Random Forest	max_depth: 29; min_samples_leaf: 3; min_samples_split: 7; n_estimators: 607
KNN	n_neighbors: 25
AdaBoost	learning_rate: 0.12; n_estimators: 202,
XGBoost	learning_rate: 0.01; max_depth: 30; n_estimators: 958; subsample: 0.6
LightGBM	learning_rate: 0.025; max_depth: 16; n_estimators: 903; num_leaves: 251; subsample: 0.85; subsample_freq: 2
ExtraTrees	max_depth: 28; min_samples_split: 2; n_estimators: 764

SVM (C controls the trade-off between model complexity and error; gamma defines the influence range of the RBF kernel function in the support vector machine), Random Forest (max_depth: maximum depth of the decision tree; min_samples_leaf: minimum number of samples required at a leaf node; min_samples_split: minimum number of samples required to split an internal node; n_estimators: number of decision trees in the random forest), KNN (n_neighbors: number of nearest neighbors used for prediction), AdaBoost (learning_rate: learning rate; n_estimators: number of weak learners), XGBoost (learning_rate: learning rate; max_depth: maximum depth of each tree; n_estimators: number of base learners; subsample: proportion of samples used for training each tree), LightGBM (learning_rate: learning rate; max_depth: maximum depth of the tree; n_estimators: number of weak learners; num_leaves: maximum number of leaf nodes per tree; subsample: proportion of samples used for training each tree; subsample_freq: frequency of random sampling control), and Extratrees (max_depth: maximum depth of the decision tree; min_samples_split: minimum number of samples required to split a node; n_estimators: number of base learners).

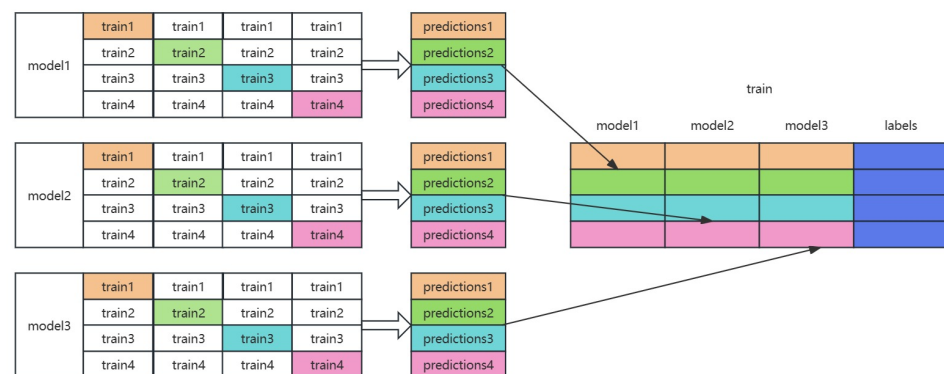


Figure 2. Using cross-validation in the training process of the stacking model.

3.3. Fraud Prediction Results and Comparison

We divided the training set and test set by year and optimized the parameters of these methods on the balanced training set by using the Bayesian optimization algorithm to obtain relatively optimal models. The relatively optimal stacking ensemble model was obtained by fusing the six best-performing models, namely, Random Forest, AdaBoost, XGBoost, ExtraTrees, LightGBM, and logistic regression. Additionally, we compared them with eight other classification methods, including SVM, Random Forest, KNN, AdaBoost, XGBoost, LightGBM, ExtraTrees, and MLP+Attention. We used data from 2013 to 2016 as the training set and data from 2017 to 2018 as the test set. Then, we used the test set samples for prediction and analysis. The comparison of the prediction results is shown in Table 7.

Table 7. Comparison of model prediction performance.

Method	Accuracy	Recall	AUC	Accuracy+Recall
MLP+Attention	0.6902	0.6500	0.7201	1.3402
KNN	0.7417	0.6011	0.7452	1.3428
SVM	0.6887	0.7567	0.7792	1.4454
Random Forest	0.6944	0.7684	0.7852	1.4628
AdaBoost	0.7451	0.7076	0.7902	1.4527
ExtraTrees	0.7066	0.7871	0.7985	1.4937
LightGBM	0.7022	0.7637	0.7987	1.4660
XGBoost	0.7162	0.7520	0.8002	1.4682
Stacking	0.6817	0.8246	0.8146	1.5063

The bolded values represent the best performance for that metric in our experimental results.

AUC (Area Under Curve), as a comprehensive metric for evaluating model performance, is not affected by the selection of specific thresholds. As shown in Table 7, the stacking model achieved an AUC of 0.8146, outperforming other mainstream machine learning models in the studies presented in the table. This indicates that the stacking ensemble model exhibited superior overall performance. This finding is consistent with previous research, which suggests that model ensembles can combine the strengths of multiple models to achieve better results. The recall rate of the stacking model reached 82.46%, which is significantly higher than that of other models in the table, indicating that the framework in this study performed exceptionally well in identifying potential fraudulent companies. The high recall rate ensures that fraudulent companies are less likely to go undetected, which holds substantial practical value in high-risk financial applications. When considering both recall and precision, the stacking model achieved a combined metric score of 1.5063, surpassing the best-performing model, Extratrees, with a score of 1.4937. The results from the research in the table show that the stacking model used in this study not only outperformed other mainstream models in terms of overall performance but also addressed the limitations of models with lower recall rates, making it better suited to meet the practical needs of financial fraud prediction.

3.4. Ablation Experiments of Framework Stages

To demonstrate the performance of each stage of the prediction framework, we conducted ablation experiments on feature selection, model selection, and hyperparameter optimization, as shown in Table 8. When no specific stage was applied (-), the model achieved an AUC of 0.8047. Introducing feature selection (FS) led to an increase in recall to 0.7953 and in AUC to 0.8102. We further applied model selection (MS) and used the better base estimators to obtain the stacking model, which yielded improvements in the AUC and recall metrics. After applying parameter optimization (PO), the overall process of the framework (FS+MS+PO) led to the best recall of 0.8246 and AUC of 0.8146.

Table 8. Performance of each stage in framework.

Stage	Accuracy	Recall	AUC	Accuracy+Recall
-	0.7082	0.7707	0.8047	1.4790
FS	0.6909	0.7953	0.8102	1.4862
FS+MS	0.6827	0.8116	0.8110	1.4944
FS+MS+PO	0.6817	0.8246	0.8146	1.5063

These results highlight the positive impact of each stage on the model's overall performance. The progressive integration of feature selection, post-processing, and model calibration not only improved AUC but also enhanced recall, showcasing the effectiveness of the proposed prediction framework.

4. Discussion

4.1. Predicting Fraud Companies on Different Stock Exchanges

The Shanghai Stock Exchange (SSE) and the Shenzhen Stock Exchange (SZSE) are the two main stock exchanges in China. There are some differences between companies listed on the SSE and the SZSE: for instance, the SSE predominantly lists large-cap stocks, with a majority of state-owned enterprises, while the SZSE focuses on small- and mid-cap stocks, with a majority of privately owned or joint venture companies. Additionally, companies listed on the SZSE have higher price-to-earnings ratios and valuations, demonstrating some growth advantages.

Therefore, we separated the companies listed on the Shanghai Stock Exchange and the Shenzhen Stock Exchange and used the proposed framework to perform financial fraud prediction analysis on the companies from each exchange separately. For the SSE companies, we selected five base models, namely, KNN, AdaBoost, XGBoost, LightGBM, and ExtraTrees, as the base estimators in our ensemble model. The hyperparameter for the final estimator (logistic regression) was adjusted with a regularization strength of 0.2. For the SZSE companies, we chose six base models, specifically SVM, Random Forest, AdaBoost, XGBoost, LightGBM, and ExtraTrees, as the base estimators in our ensemble. The hyperparameter for the final estimator (logistic regression model) was adjusted with a regularization strength of 10.

As shown in Figures 3 and 4, the indicators of the stacking model used for fraud prediction in this framework outperformed the other seven mainstream machine learning models. These results further demonstrate the excellent generalization ability of the proposed framework on this task. Generally, the market pays more attention to the financial condition of high-growth companies to avoid risks, and the model exhibits stronger prediction capabilities for companies listed on the SZSE, presenting more opportunities for practical applications. In addition, compared with using prediction solely based on companies from a specific stock exchange, incorporating all companies yielded better results. Therefore, when the model framework conducts unified prediction, incorporating more effective data can improve prediction performance and robustness.

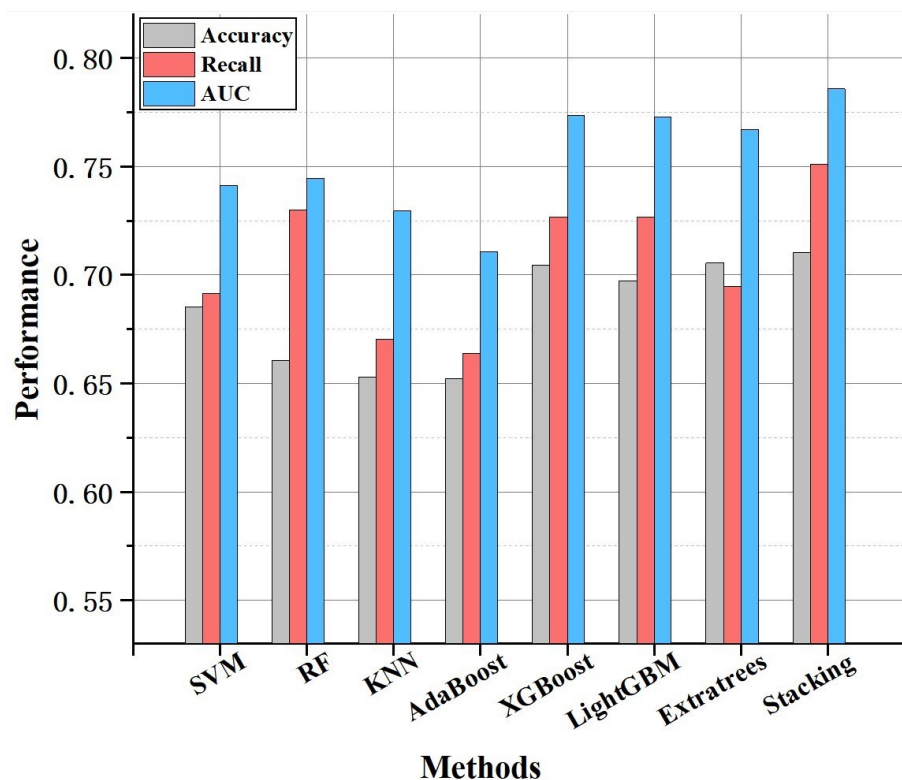


Figure 3. Fraud prediction with different models in stock exchange (SSE).

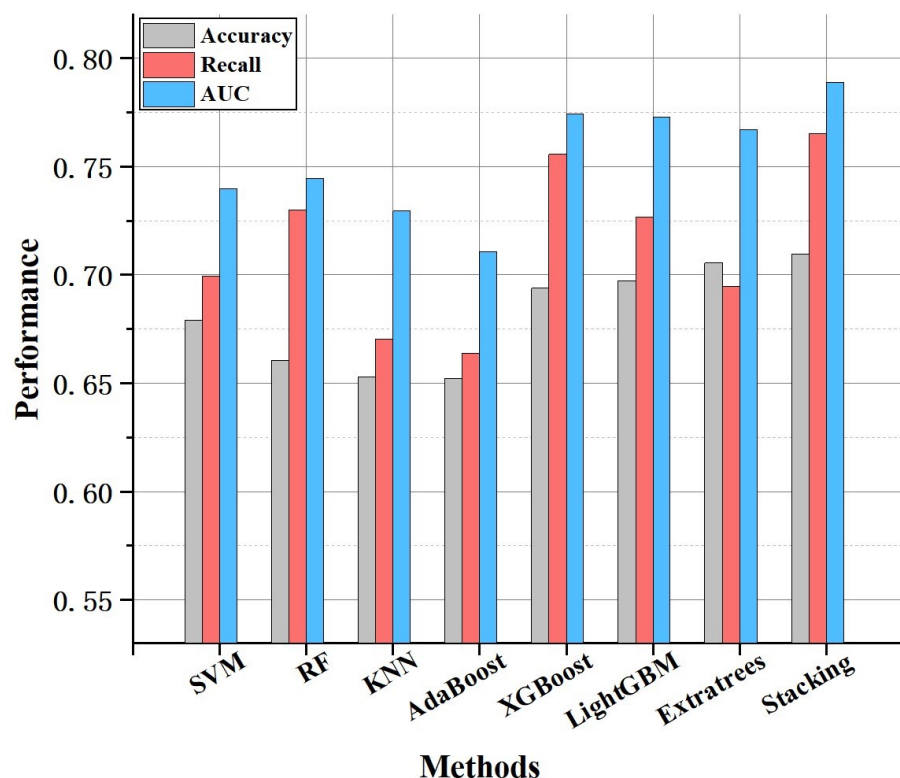


Figure 4. Fraud prediction with different models in stock exchange (SZSE).

4.2. Predicting Frauds Occurring in the Future

In order to predict the potential fraud risk of companies earlier and take corresponding preventive measures, further experiments were conducted in this study by using the proposed framework to predict the likelihood of financial fraud occurring in companies in the following two to three years. Specifically, the fraud occurrence in 2019 and 2020 was predicted based on the data from 2017. Firstly, when predicting the likelihood of fraud occurrence in the next two years, we selected KNN, AdaBoost, XGBoost, LightGBM, and Extratrees as the base estimators in our ensemble model after filtering. The hyperparameters of the final estimator (logistic regression) were tuned with a regularization strength of 1.7. Additionally, when predicting the likelihood of financial fraud occurring within the next three years, we ultimately chose SVM, Random Forest, XGBoost, and Extratrees as the base estimators in our ensemble model. The regularization strength of the final estimator (logistic regression) was set to 0.017.

The experimental results, as shown in Figures 5 and 6, demonstrate that the stacking model, which combines the predictions of multiple models within the proposed framework, outperformed the other seven machine learning models in terms of the AUC metric. Furthermore, it achieved the highest sum of accuracy and recall values.

Overall, when comparing different prediction horizons, the performance of the model tended to decrease as the prediction time span increased. As time went on, the effectiveness of the model diminished. However, when used in the same scenario, the stacking model still outperformed the other seven mainstream machine learning models, demonstrating the good generalization ability of the proposed framework.

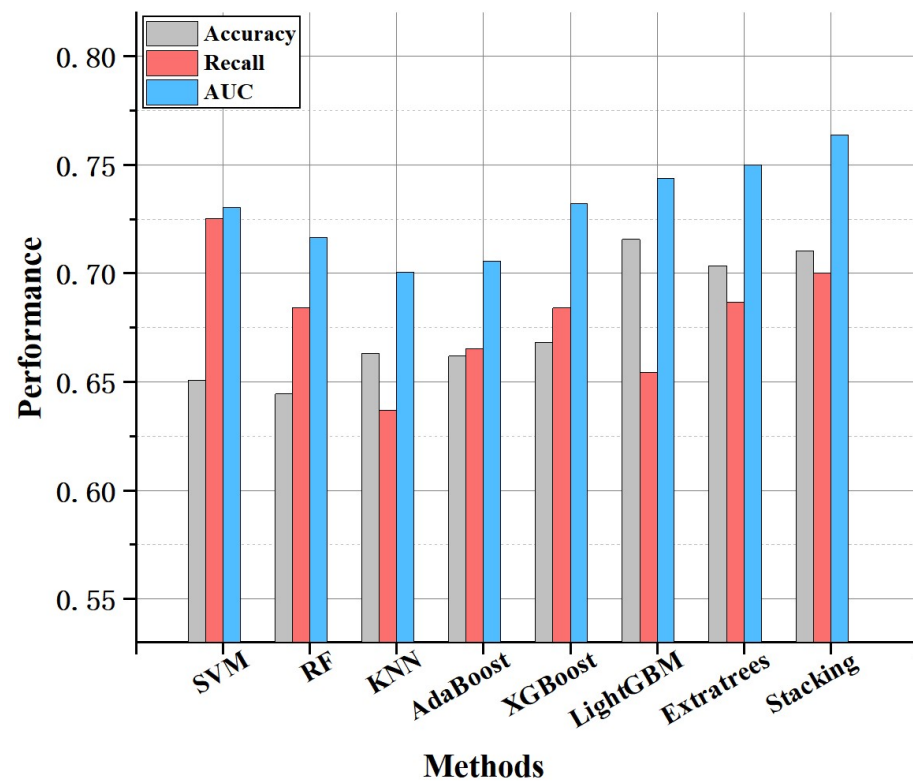


Figure 5. Fraud prediction for future period (next 2 years).

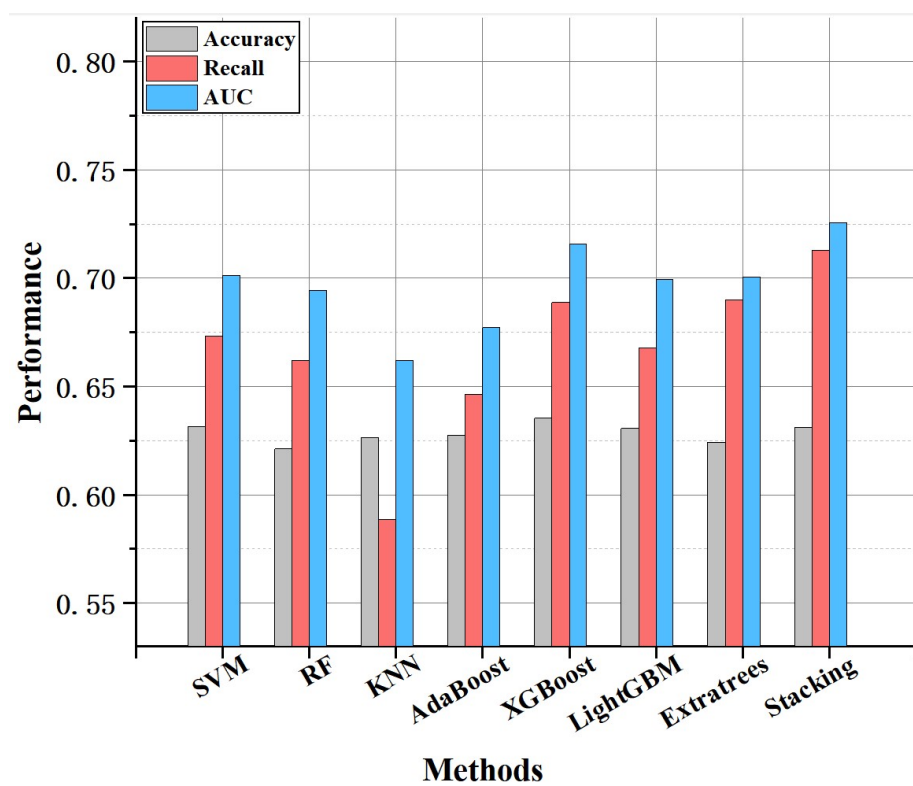


Figure 6. Fraud prediction for future period (next 3 years).

4.3. Predicting Specific Fraudulent Behavior

In this study, financial fraud is defined as the violation involving one or more of the following behaviors: “fictitious profits”, “misrepresentation of assets”, “false records (misleading statements)”, and “general improper accounting treatment”. In financial fraud analysis, predicting specific fraudulent behaviors helps gaining a deeper understanding of the operational conditions and motives of fraudulent companies, enabling regulatory agencies and investors to take more effective measures. Therefore, we conducted predictive experiments on individual fraudulent behaviors related to financial fraud.

By using the proposed framework, we predicted the occurrence of fictitious profits, misrepresentation of assets, false records (misleading statements), and general improper accounting treatment separately. We used the stacking ensemble model which combines the best-performing models, including Random Forest, AdaBoost, XGBoost, Extratrees, LightGBM, and logistic regression.

As shown in Figures 7–10, when individually predicting these four specific fraudulent behaviors related to financial fraud, the stacking model still demonstrated the best performance, achieving the highest AUC and recall rates. This result further supports the notion that the proposed framework has excellent generalization ability and outperforms other mainstream machine learning models in the task of identifying specific fraudulent behaviors. Among these four fraud types, the prediction performance was the highest on misrepresentation of assets. This may be because the misrepresentation of assets fraudulent behavior causes more noticeable anomalies in both financial and non-financial data, making it easier for machine learning models to detect it.

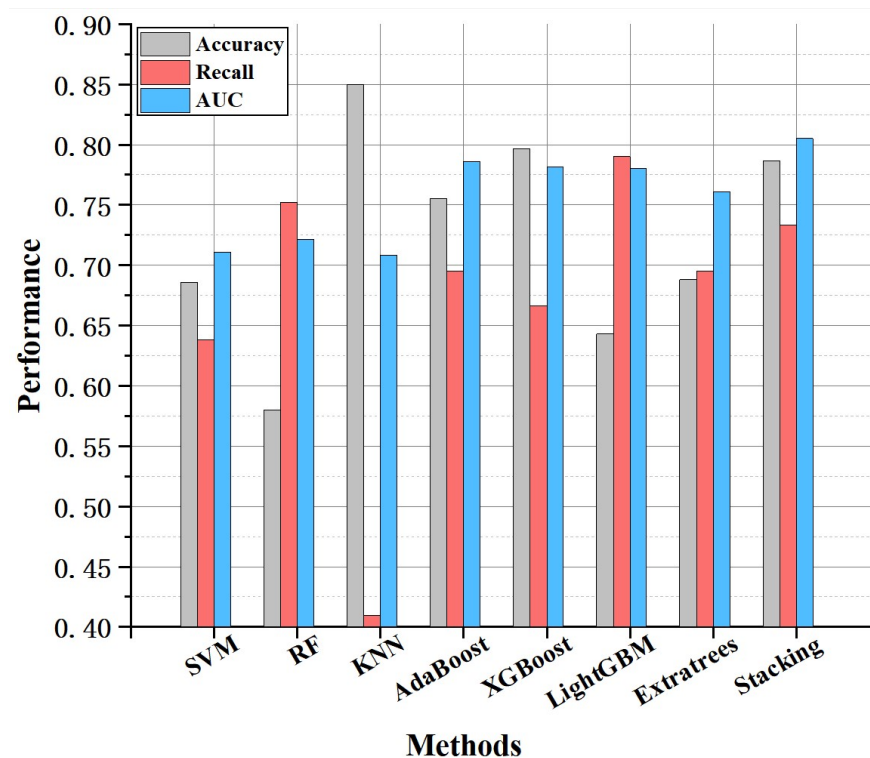


Figure 7. Prediction of specific fraudulent behavior (fabricated profits).

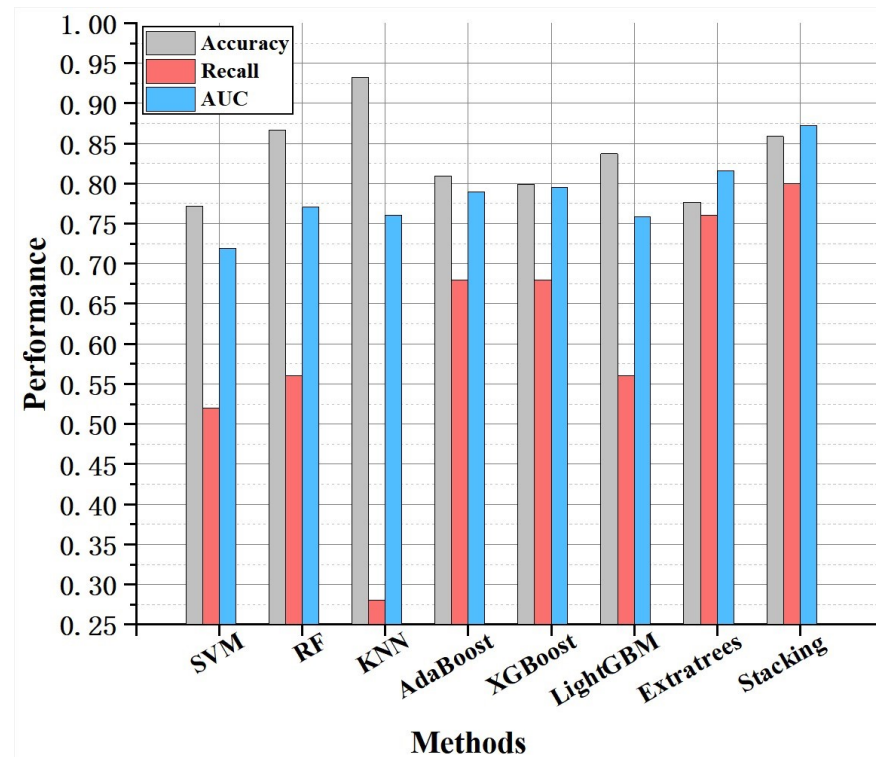


Figure 8. Prediction of specific fraudulent behavior (misreported assets).

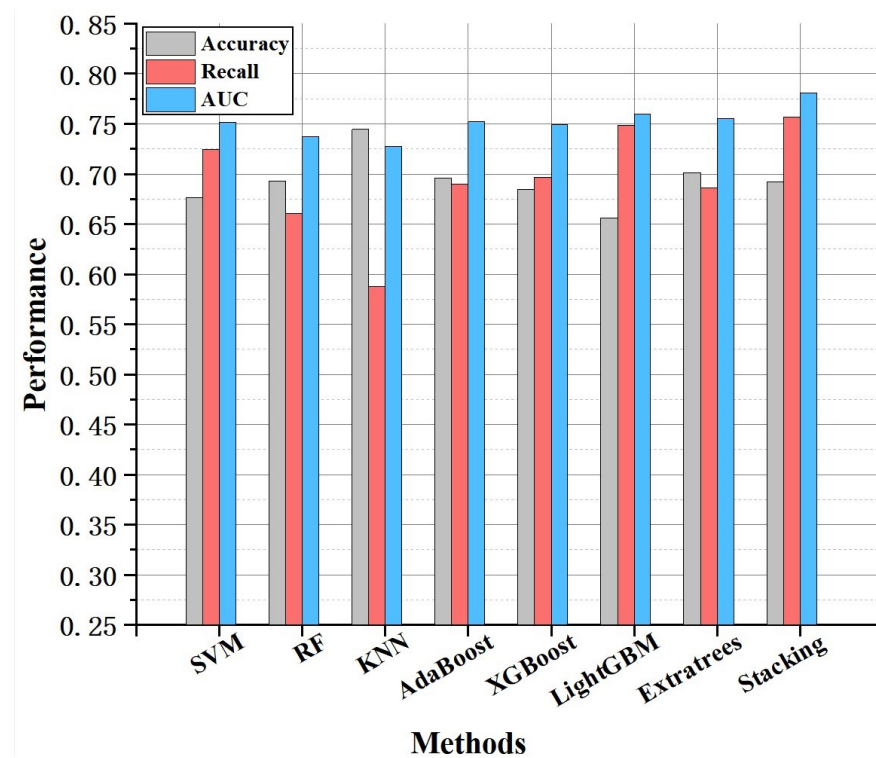


Figure 9. Prediction of specific fraudulent behavior (false records (misleading statements)).

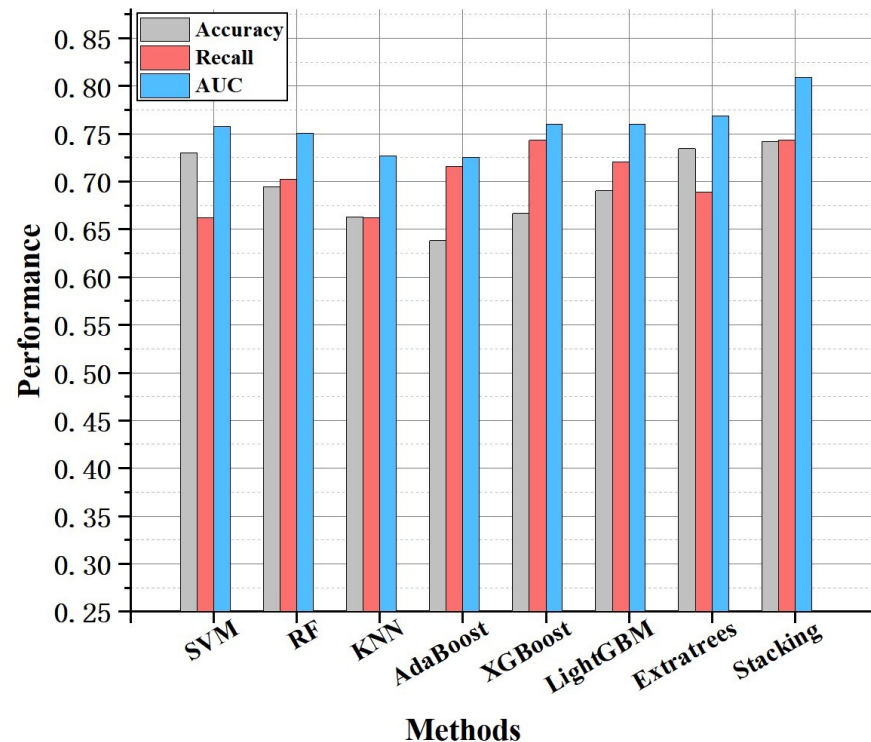


Figure 10. Prediction of specific fraudulent behavior (improper general accounting practices).

4.4. Analysis of Features

To provide more direct information for financial fraud prediction, this framework also analyzes the input features based on the base learners of the stacking model to identify features that are highly correlated with fraudulent behavior. Specifically, we obtained the feature importance rankings for each base learner after training. Among them, 6 features simultaneously appeared in the top 50 features of feature importance rankings for all five machine learning models. The specific meanings of these features are provided in Table 9.

Table 9. Shared important features of base estimators.

Features	Definition
Common Stock Earnings Yield A	After-tax dividend per share divided/current period's closing price
Top Ten Holders Rate	Proportion of shares held by the top ten shareholders
Ownership	1 for state-owned enterprises, else 0
Tangible Asset Ratio	$\text{Tangible Assets} / \text{Total Assets}$ ($\text{Tangible Net Assets} = \text{Total Assets} - \text{Intangible Assets} - \text{Goodwill}$)
Retained Earnings to Total Assets Ratio	$(\text{Retained Earnings} + \text{Undistributed Profits}) / \text{Total Assets}$
Earnings per Share Reserve	$\text{Ending Value of Earnings per Share Reserve} / \text{Ending Value of Paid-in Capital}$

Furthermore, we randomly selected several specific fraud cases for analysis and the further evaluation of the proposed framework. By using the stacking model, we predicted the occurrence of financial fraud in 2018 based on the data from 2017. The model detected fraudulent behavior for (i) Kangmei Pharmaceutical Co., Ltd. (600518), (ii) Chuying Agro-Pastoral Group Co., Ltd. (002477), and (iii) DHC Software Co., Ltd. (002065). Kangmei Pharmaceutical Co., Ltd. was accused of false entries, Chuying Agro-Pastoral Group Co., Ltd. was accused of fictitious profits, false presentation of assets, and false entries, and DHC

Software Co., Ltd. was accused of general accounting mishandling. Among them, Kangmei Pharmaceutical had a retained earnings-to-total assets ratio of 0.1416, which was relatively low. Chuying Agro-Pastoral had a retained earnings-to-total assets ratio of 5.50% and a return on equity of 0.45%, both relatively low. Its earnings per share reserve was only 6.37%, far below the average level. The fraudulent behavior of these three companies are described below.

(i) Kangmei Pharmaceutical Co., Ltd., abbreviated as “Kangmei Pharmaceutical” (SSE: 600518), operates in the pharmaceutical manufacturing industry with industry code C27 and is located in Puning, Guangdong Province, China, Guangdong Province. On the evening of 28 December 2018, Kangmei Pharmaceutical announced that the company had received an “Investigation Notice” from the China Securities Regulatory Commission (CSRC). On 17 May 2019, the CSRC reported that Kangmei Pharmaceutical had engaged in financial report falsification, involving suspected false statements and other illegal activities.

(ii) Chuying Agro-Pastoral Group Co., Ltd., was successfully listed on the Shenzhen Stock Exchange on 15 September 2010 (stock code: 002477). It operates in the forestry industry with industry code A03 and is located in Xinzheng, Henan Province, China. The violation records in the CSMAR database indicate that in the 2018 fiscal year, Chuying Agro-Pastoral inflated equity and debt investments, totaling CNY 6,975,744,631.86. Chuying Agro-Pastoral’s financial data in the “2018 Annual Report” were found to contain false information.

(iii) DHC Software Co., Ltd. was founded in January 2001, with its headquarters located in Beijing’s Zhongguancun. It went public on the Shenzhen Stock Exchange’s main board in August 2006, with stock code 002065. The company operates in the software and information technology services industry, with industry code I65. In 2018 and 2019, DHC Software had several trade-related transactions, with recognized revenues that did not comply with accounting standards. This resulted in DHC Software inflating operating income by CNY 65,226,300 and CNY 36,232,700 in its 2018 and 2019 annual financial reports, respectively, accounting for 0.77% and 0.41% of the total operating income for each respective year.

We also conducted an additional analysis by using the ExtraTrees model, which performed well within the framework, to plot partial dependence plots (PDPs) for the top five features in terms of importance (Table 10). The PDPs show the impact of a specific feature on the model’s predictions, helping us to understand the model’s response to individual features, identify nonlinear behavior, and isolate the effects of other features.

Table 10. The five most important features of the ExtraTrees model.

Features	Definition
Ownership	1 for state-owned enterprises, else 0
Report Reliability Index	Report reliability index published by Shenzhen Dibo Company
Legal Compliance Index	Legal compliance index published by Shenzhen Dibo Company
Largest Holder Rate	Proportion of shares held by the largest shareholder
Common Stock Earnings Yield A	After-tax dividend per share divided/current period’s closing price

The results are shown in Figures 11–15. The partial dependence plots obtained from the model analysis reveal a noticeable linear relationship with financial fraud for certain features: (a) The probability of financial fraud is lower when the property rights are classified as state-owned enterprises. State-owned enterprises are typically subject to stricter regulatory and auditing systems. Government departments and relevant regulatory bodies conduct more rigorous scrutiny and supervision of the financial reports and activities of state-owned enterprises. This is likely to reduce the occurrence of fraudulent behavior. (b) A lower reliability index in the reports is associated with a higher probability of financial fraud. A lower reliability index may indicate a higher risk of false reporting and financial data manipulation within the company, reflecting a lack of transparency in its financial information. Inaccurate, misleading, or intentionally fabricated financial information can be used to deceive

investors, creditors, and other stakeholders. (c) A lower legal compliance index corresponds to a higher probability of financial fraud. A lower legal compliance index may reflect a lack of robust regulations and systems within a company. The absence of effective internal controls, supervision, and compliance mechanisms may provide opportunities and room for financial fraud to occur. (d) When the proportion of shares held by the largest shareholder (the first major shareholder) is lower, the probability of financial fraud is higher. When the largest shareholder's control over the company is weakened, it creates greater opportunities and motivations for other stakeholders, such as minority shareholders and management, to engage in financial fraud. (e) As Common Stock Earnings Yield A decreases, the probability of financial fraud increases. A decrease in Common Stock Earnings Yield A may indicate a weaker economic condition for the company. In times of financial hardship, the company is likely to face greater pressure and may be more inclined to engage in financial fraud as a means to conceal losses and poor financial performance.

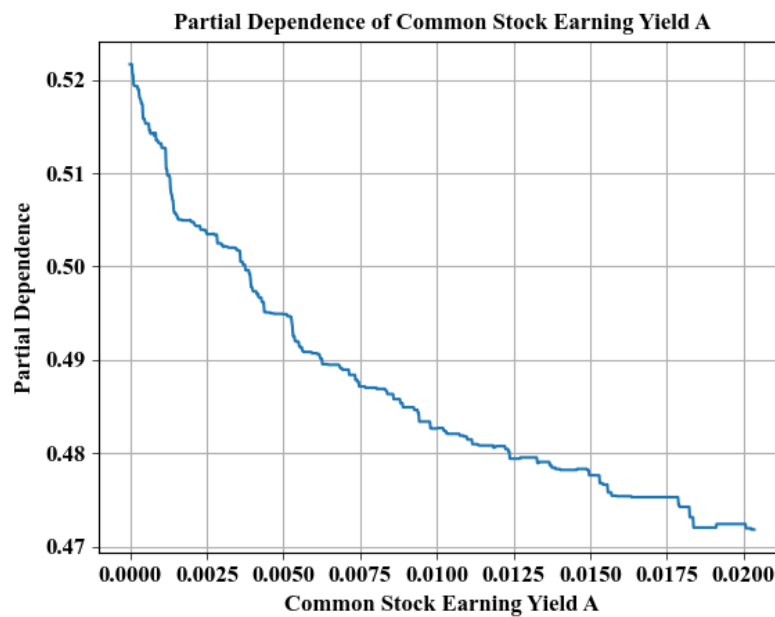


Figure 11. The PDP for the top five features of the ExtraTrees model (ownership).

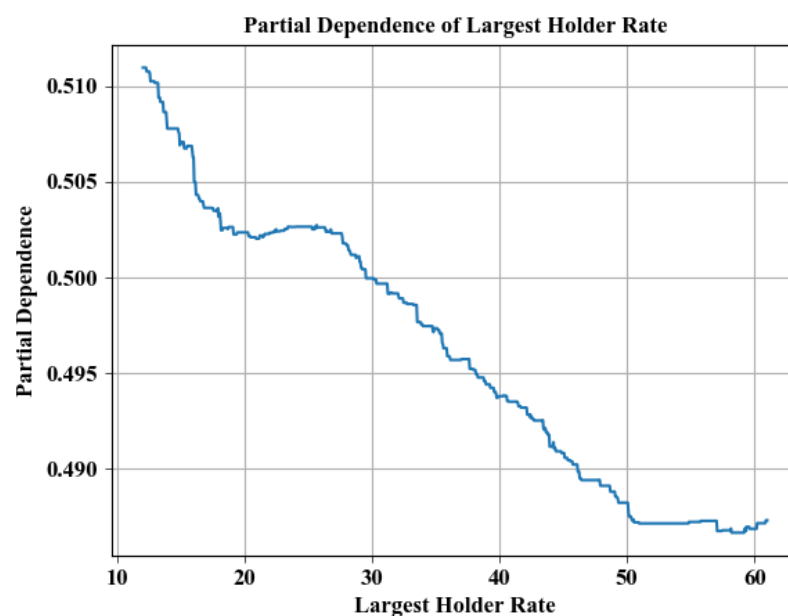


Figure 12. The PDP for the top five features of the ExtraTrees model (report reliability index).

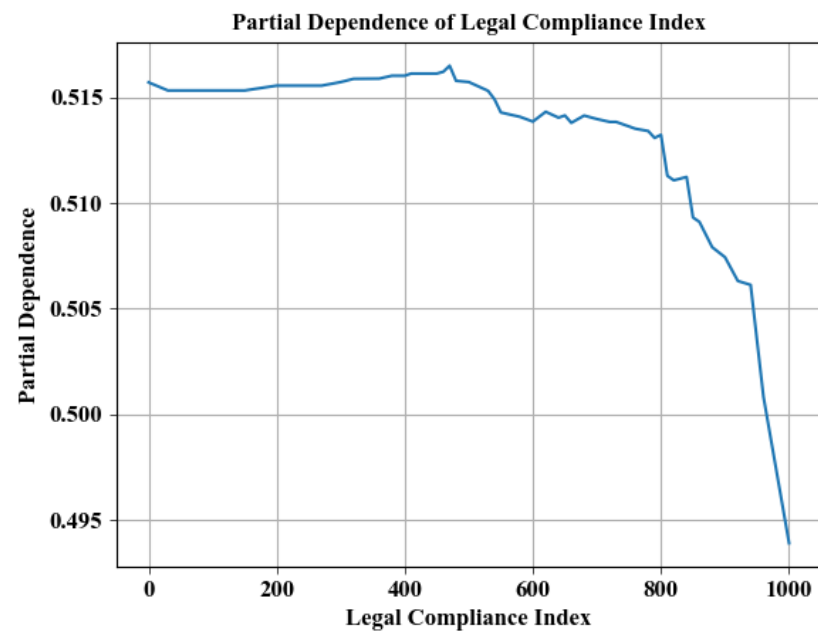


Figure 13. The PDP for the top five features of the ExtraTrees model (legal compliance index).

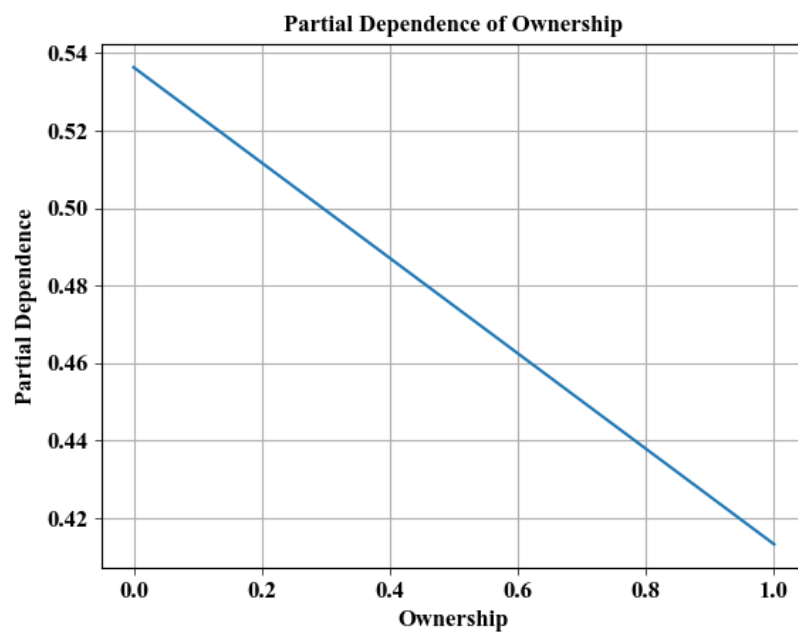


Figure 14. The PDP for the top five features of the ExtraTrees model (Largest Holder Rate).

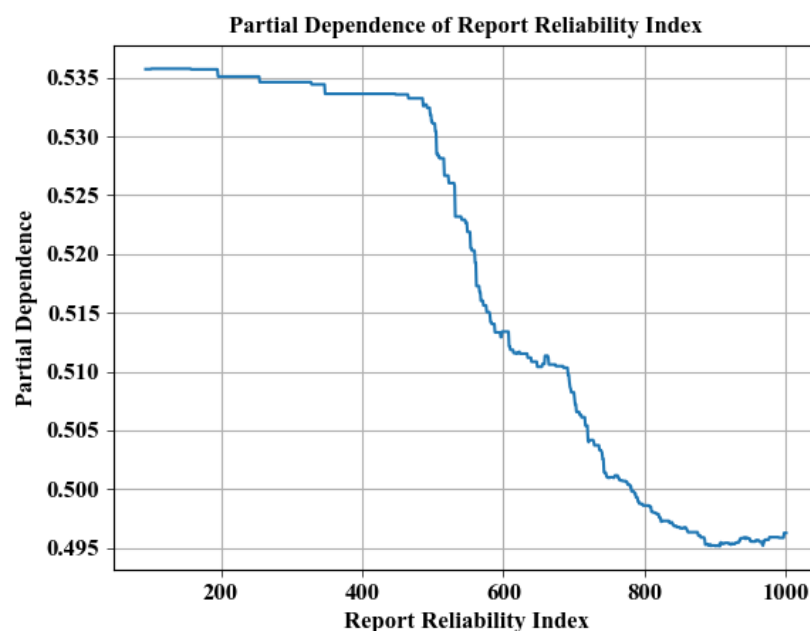


Figure 15. The PDP for the top five features of the ExtraTrees model (Common Stock Earnings Yield A).

5. Conclusions

In this paper, we have proposed a financial fraud prediction framework based on stacking ensemble learning, taking into account feature analysis, parameter optimization, model selection, and stacking ensemble, and performance evaluation. Unlike previous studies, we focus on financial fraud prediction tasks, aiming to analyze historical data and identify potential risk factors to provide early warnings before issues arise. This approach gives relevant departments more time and opportunities to monitor or take corrective actions, thereby reducing potential financial losses and legal risks. By implementing the financial fraud prediction framework proposed in this study, regulators can identify potential risk points before problems occur, enabling timely intervention. For example, if the predictive model detects anomalies indicating fraudulent risks in a company's financial statements, regulators can intervene early to conduct preliminary investigations, preventing further escalation and greater losses. For auditors, adopting this predictive framework enhances the accuracy and efficiency of audits. By leveraging the insights provided by the model, auditors can focus their efforts on identified high-risk areas, making their work more targeted and effective. Lastly, for investors, the proposed approach offers critical insights to help them avoid high-risk investments and protect their interests. By identifying potential financial fraud, investors can make more informed decisions. For instance, when evaluating a company, if the predictive model flags its financial data as high-risk, investors might choose to withhold investment in the company or demand greater transparency and explanations.

The objective of stacking lies in harnessing the collective power of multiple base models by inputting their predictions into a meta-model. This framework addresses the challenge of limited generalization ability in existing machine learning methods for financial fraud prediction tasks. The framework involves effective analysis and cross-validation to select financial and non-financial features. SVM, LightGBM, and other machine learning models are used as base models for training and parameter optimization. Finally, the best-performing base models and meta-model are selected for stacking ensemble. Experiments conducted by using data from listed companies in China have demonstrated that the financial fraud prediction framework based on stacking ensemble achieves a recall rate of 0.8246 and an AUC of 0.8146, surpassing the individual performance of mainstream machine learning models like SVM, LightGBM, and ExtraTrees. The study demonstrates that the

proposed framework performs exceptionally well in different types of companies, violation timeframes, and fraudulent behaviors, consistently outperforming current mainstream machine learning models. This showcases its strong generalization capability, which enables the model to adapt to diverse financial contexts. In future research and practical applications, the framework can be tailored to different fraud scenarios with only minor adjustments to the base models. Furthermore, ablative experiments have shown that the complete framework constructed through feature selection, parameter optimization, and model ensemble yields better results. The results and contributions of this study provide important guidance for regulators, managers, auditors, and investors in their decision-making processes, facilitating effective risk management and ensuring the stability of financial markets.

However, this study also has certain limitations. For example, the coverage of the dataset is limited, and there are many types of non-financial data that can still be explored for valuable information. The stacking ensemble model we used is more complex compared to general machine learning models, which makes conducting related research on large datasets more computationally intensive. On the other hand, our framework has limited overall interpretability. After utilizing our framework, to gain a clearer understanding of the logic behind the predictions, further research on the base models in the framework would be required, making the process more cumbersome.

There are several potential directions for further research. Firstly, it is worth considering the inclusion of additional types of data, such as social media sentiment data, textual data from annual and quarterly reports, and audio data from company meetings. By incorporating data from different dimensions, we can gain a more comprehensive understanding of company operations and more accurately identify potential fraudulent activities. Secondly, it is possible to further model the time-series relationship between the financial and non-financial information of companies. By analyzing the dynamic changes in financial data and other relevant factors, we can capture signs of fraudulent behavior and identify potential fraud signals. Furthermore, exploring the impact of interconnections between different companies on financial fraud behavior, such as relationships between suppliers and customers, can contribute to the development of more comprehensive fraud prediction models. Finally, with the advancement and application of Large Language Models, their application to financial fraud prediction will be a key area of future research. Large Language Models have the ability to handle and analyze data of different types and scales, uncovering patterns and anomalies hidden within the data and providing valuable assistance in analysis and decision making.

Author Contributions: Conceptualization, S.Z., H.W., J.R., D.H. and T.M.; methodology, S.Z., H.W. and D.H.; software, S.Z. and H.W.; validation, S.Z. and H.W.; formal analysis, S.Z. and H.W.; investigation, S.Z. and H.W.; resources, J.R. and D.H.; data curation, H.W. and T.M.; writing—original draft preparation, H.W. and T.M.; writing—review and editing, S.Z., J.R., E.W.T.N., D.H. and Y.L.; visualization, H.W.; supervision, J.R., E.W.T.N., D.H. and Y.L.; project administration, J.R., E.W.T.N. and D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Natural Science Foundation of China (Grant: 62376074), the Shenzhen Science and Technology Program (Grants: RKX20231110090859012, SGDX20230116091244004, KCXST20221021111404010, JSGGKQTD20221101115655027, KJZD20231023095959002), Shenzhen Humanities and Social Sciences Key Research Bases (Grant number KP191001) and Harbin Institute of Technology (Shenzhen) Joint Basic Education Cultivation Project “Application Project of Intelligent Assistive Teaching System for Secondary School Biology Curriculum Based on Multimodal Large Language Model”.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. ACFE. *Report to the Nations 2020 Global Study on Occupational Fraud and Abuse*; Association of Certified Fraud Examiners: New York, NY, USA, 2020. Available online: <https://legacy.acfe.com/report-to-the-nations/2020/> (accessed on 1 July 2023).
2. Kwok, B.K. *Accounting Irregularities in Financial Statements: A Definitive Guide for Litigators, Auditors and Fraud Investigators*; Routledge: Abingdon, UK, 2017.
3. Papík, M.; Papíková, L. Detecting accounting fraud in companies reporting under US GAAP through data mining. *Int. J. Account. Inf. Syst.* **2022**, *45*, 100559. [\[CrossRef\]](#)
4. Cressey, D. *Other People's Money; A Study of the Social Psychology of Embezzlement*; Patterson Smith: Montclair, NJ, USA, 1953.
5. Imoniana, J.O.; Murcia, F.D.R. Patterns of similarity of corporate frauds. *Qual. Rep.* **2016**, *21*, 143. [\[CrossRef\]](#)
6. Shoetan, P.O.; Oyewole, A.T.; Okoye, C.C.; Ofodile, O.C. Reviewing the role of big data analytics in financial fraud detection. *Financ. Account. Res. J.* **2024**, *6*, 384–394. [\[CrossRef\]](#)
7. Li, J.; Chang, Y.; Wang, Y.; Zhu, X. Tracking down financial statement fraud by analyzing the supplier-customer relationship network. *Comput. Ind. Eng.* **2023**, *178*, 109118. [\[CrossRef\]](#)
8. Meredith, K.; Blake, J.; Baxter, P.; Kerr, D. Drivers of and barriers to decision support technology use by financial report auditors. *Decis. Support Syst.* **2020**, *139*, 113402. [\[CrossRef\]](#)
9. Bao, Y.; Ke, B.; Li, B.; Yu, Y.J.; Zhang, J. Detecting accounting fraud in publicly traded US firms using a machine learning approach. *J. Account. Res.* **2020**, *58*, 199–235. [\[CrossRef\]](#)
10. Khan, A.T.; Cao, X.; Li, S.; Katsikis, V.N.; Brajevic, I.; Stanimirovic, P.S. Fraud detection in publicly traded US firms using Beetle Antennae Search: A machine learning approach. *Expert Syst. Appl.* **2022**, *191*, 116148. [\[CrossRef\]](#)
11. Wang, G.; Ma, J.; Chen, G. Attentive statement fraud detection: Distinguishing multimodal financial data with fine-grained attention. *Decis. Support Syst.* **2023**, *167*, 113913. [\[CrossRef\]](#)
12. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 507–520.
13. Dechow, P.M.; Ge, W.; Larson, C.R.; Sloan, R.G. Predicting material accounting misstatements. *Contemp. Account. Res.* **2011**, *28*, 17–82. [\[CrossRef\]](#)
14. Perols, J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Audit. A J. Pract. Theory* **2011**, *30*, 19–50. [\[CrossRef\]](#)
15. Abbasi, A.; Albrecht, C.; Vance, A.; Hansen, J. Metafraud: A meta-learning framework for detecting financial fraud. *Mis. Q.* **2012**, *36*, 1293–1327. [\[CrossRef\]](#)
16. Ravisankar, P.; Ravi, V.; Rao, G.R.; Bose, I. Detection of financial statement fraud and feature selection using data mining techniques. *Decis. Support Syst.* **2011**, *50*, 491–500. [\[CrossRef\]](#)
17. Hassanniakalager, A.; Perotti, P.; Tsoligkas, F. A Machine Learning Approach to Detect Accounting Frauds. *SSRN Electron. J.* **2022**. [\[CrossRef\]](#)
18. Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybern.-Part A Syst. Humans* **2009**, *40*, 185–197. [\[CrossRef\]](#)
19. Beaver, W.H. Financial ratios as predictors of failure. *J. Account. Res.* **1966**, *4*, 71–111. [\[CrossRef\]](#)
20. Fanning, K.; Cogger, K.O.; Srivastava, R. Detection of management fraud: A neural network approach. *Intell. Syst. Account. Financ. Manag.* **1995**, *4*, 113–126. [\[CrossRef\]](#)
21. Green, B.P.; Choi, J.H. Assessing the risk of management fraud through neural network technology. *Auditing* **1997**, *16*, 14–28.
22. Fanning, K.M.; Cogger, K.O. Neural network detection of management fraud using published financial data. *Intell. Syst. Account. Financ. Manag.* **1998**, *7*, 21–41. [\[CrossRef\]](#)
23. Cecchini, M.; Aytug, H.; Koehler, G.J.; Pathak, P. Detecting management fraud in public companies. *Manag. Sci.* **2010**, *56*, 1146–1160. [\[CrossRef\]](#)
24. Xu, F.; Zhu, Z. A Bayesian approach for predicting material accounting misstatements. *Asia-Pac. J. Account. Econ.* **2014**, *21*, 349–367. [\[CrossRef\]](#)
25. Papík, M.; Papíková, L. Application of selected data mining techniques in unintentional accounting error detection. *Equilib. Q. J. Econ. Econ. Policy* **2021**, *16*, 185–201. [\[CrossRef\]](#)
26. Kim, Y.J.; Baik, B.; Cho, S. Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Syst. Appl.* **2016**, *62*, 32–43. [\[CrossRef\]](#)
27. Hajek, P.; Henriques, R. Mining corporate annual reports for intelligent detection of financial statement fraud—A comparative study of machine learning methods. *Knowl.-Based Syst.* **2017**, *128*, 139–152. [\[CrossRef\]](#)
28. Brown, N.C.; Crowley, R.M.; Elliott, W.B. What are you saying? Using topic to detect financial misreporting. *J. Account. Res.* **2020**, *58*, 237–291. [\[CrossRef\]](#)
29. Humpherys, S.L.; Moffitt, K.C.; Burns, M.B.; Burgoon, J.K.; Felix, W.F. Identification of fraudulent financial statements using linguistic credibility analysis. *Decis. Support Syst.* **2011**, *50*, 585–594. [\[CrossRef\]](#)
30. Craja, P.; Kim, A.; Lessmann, S. Deep learning for detecting financial statement fraud. *Decis. Support Syst.* **2020**, *139*, 113421. [\[CrossRef\]](#)
31. Jan, C.L. Detection of financial statement fraud using deep learning for sustainable development of capital markets under information asymmetry. *Sustainability* **2021**, *13*, 9879. [\[CrossRef\]](#)

32. Papik, M.; Papikova, L. Detection models for unintentional financial restatements. *J. Bus. Econ. Manag.* **2020**, *21*, 64–86. [\[CrossRef\]](#)
33. Hamal, S.; Senvar, Ö. Comparing performances and effectiveness of machine learning classifiers in detecting financial accounting fraud for Turkish SMEs. *Int. J. Comput. Intell. Syst.* **2021**, *14*, 769–782. [\[CrossRef\]](#)
34. Cheng, C.H.; Kao, Y.F.; Lin, H.P. A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes. *Appl. Soft Comput.* **2021**, *108*, 107487. [\[CrossRef\]](#)
35. Gepp, A.; Kumar, K.; Bhattacharya, S. Lifting the numbers game: Identifying key input variables and a best-performing model to detect financial statement fraud. *Account. Financ.* **2021**, *61*, 4601–4638. [\[CrossRef\]](#)
36. Throckmorton, C.S.; Mayew, W.J.; Venkatachalam, M.; Collins, L.M. Financial fraud detection using vocal, linguistic and financial cues. *Decis. Support Syst.* **2015**, *74*, 78–87. [\[CrossRef\]](#)
37. Yao, J.; Pan, Y.; Yang, S.; Chen, Y.; Li, Y. Detecting fraudulent financial statements for the sustainable development of the socio-economy in China: A multi-analytic approach. *Sustainability* **2019**, *11*, 1579. [\[CrossRef\]](#)
38. Duan, W.; Hu, N.; Xue, F. The information content of financial statement fraud risk: An ensemble learning approach. *Decis. Support Syst.* **2024**, *174*, 114231. [\[CrossRef\]](#)
39. Khaksari, I.; Shoorvarzi, M.; Mehrazeen, A.; Massihabadi, A. Developing a model to predict fraudulent financial reporting. *Int. J. Nonlinear Anal. Appl.* **2024**, *15*, 93–105.
40. Rahman, M.J.; Jie, X. Fraud detection using fraud triangle theory: Evidence from China. *J. Financ. Crime* **2024**, *31*, 101–118. [\[CrossRef\]](#)
41. Bhattacharya, I.; Mickovic, A. Accounting fraud detection using contextual language learning. *Int. J. Account. Inf. Syst.* **2024**, *53*, 100682. [\[CrossRef\]](#)
42. Bertomeu, J.; Cheynel, E.; Floyd, E.; Pan, W. Using machine learning to detect misstatements. *Rev. Account. Stud.* **2021**, *26*, 468–519. [\[CrossRef\]](#)
43. Xu, X.; Xiong, F.; An, Z. Using machine learning to predict corporate fraud: Evidence based on the GONE framework. *J. Bus. Ethics* **2023**, *186*, 137–158. [\[CrossRef\]](#)
44. Pazarskis, M.; Lazos, G.; Koutoupis, A.; Drogas, G. Preventing the unpleasant: Fraudulent financial statement detection using financial ratios. *J. Oper. Risk* **2021**, *17*, 1–18. [\[CrossRef\]](#)
45. Chen, Y.J.; Wu, C.H.; Chen, Y.M.; Li, H.Y.; Chen, H.K. Enhancement of fraud detection for narratives in annual reports. *Int. J. Account. Inf. Syst.* **2017**, *26*, 32–45. [\[CrossRef\]](#)
46. Kotsiantis, S.; Koumanakos, E.; Tzelepis, D.; Tampakas, V. Forecasting fraudulent financial statements using data mining. *Int. J. Comput. Intell.* **2006**, *3*, 104–110.
47. Schapire, R.E. The strength of weak learnability. *Mach. Learn.* **1990**, *5*, 197–227. [\[CrossRef\]](#)
48. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [\[CrossRef\]](#)
49. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
50. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
51. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
52. Whiting, D.G.; Hansen, J.V.; McDonald, J.B.; Albrecht, C.; Albrecht, W.S. Machine learning methods for detecting patterns of management fraud. *Comput. Intell.* **2012**, *28*, 505–527. [\[CrossRef\]](#)
53. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.