Manufacturing Companies Financial Fraud Detection Based on Interpretable Machine Learning

Xiang Li[1,a,*], Xinyu Da[2,b], Wanxin Shi[3,c], Wenjun Liu[4,d]

{lixiang77@stu.scu.edu.cn[a,*], daxinyu@stu.scu.edu.cn[b], shiwanxin@stu.scu.edu.cn[c], liuwenjunjj@stu.scu.cn[d]}

College of Business, Sichuan University,Chengdu 610064, China[1,3,4]
College of Mathematics and Economics, Sichuan University,Chengdu 610064, China[2]

Abstract: Because of the industrial characteristics like complex purchasing and selling links in manufacturing companies, it is relatively easy for them to conduct financial fraud. Taking A-share manufacturing companies from 2009 to 2021 as an observation sample, this paper constructed a financial fraud detection feature set of manufacturing companies from five dimensions. By using 5 algorithms to build detection models and calculating the metrics, research found that XGBoost, LightGBM and Random Forest have better predictive performance. And the further parameterized features importance showed that indicators such as "ROA", "Income receivable" and "Internal control index" have a significant guiding role in identifying financial fraud of Manufacturing Listed Corporations.

Keywords: Machine Learning; Financial Fraud; Detection Framework; Binary Classification

1 INTRODUCTION

Manufacturing enterprises are the foundation of China's industrial development. However, they are facing all types of internal and external problems, and the characteristics of the listed manufacturing companies like complex purchasing and selling links lead to various financial fraud problems. The traditional financial fraud detection method is analyzing the existing financial statement data. However, the fraudulent means of listed companies are becoming more and more concealed these days. Therefore, it is significant and necessary to build a financial fraud detection model which combines finance and big data to identify corporate financial fraud better and improve the environment of capital market.

This paper starts from existing theories like the perspective of accounting information system theory, combining the characteristics of financial fraud in manufacturing, and utilizes interpretable machine learning algorithms to build a predictive model of financial fraud. The remaining parts are as follows: the second part is literature review, which summarizes the recent theoretical literature about the types, means and detection of corporate financial fraud. The third part is research design, which constructs a multidimensional financial fraud detection model. The fourth part is model construction and the last part presents the conclusions and recommendations.

ICPDI 2023, September 01-03, Chongqing, People's Republic of China

DOI 10.4108/eai.1-9-2023.2338768

The contributions and innovations are as follows. Firstly, this papaer used a five-dimensional framework, which includes indicators of the process of generating accounting information, and further excavated the information of implied financial fraud. Secondly, the conclusions can help enrich the application of big data and machine learning in financial fraud, as well as construct a detection model that adapt to the situation of China's capital market. Thirdly, this paper fully integrated the information and methods to compare various types of data and models to select the featured one that match the characteristics of the manufacturing industry. Finally, innovatively integrating big data into enterprise financial management helps companies to identify and predict their financial fraud and improve their financial management and corporate governance level.

2 LITERATURE REVIEW

At present, domestic and foreign scholars have already had some significant results in the analysis of financial data and the detection of financial fraud, whose direction can be divided into the following categories.

First，constructing the theory of financial fraud detection. Most of the researches focus on the composition of fraud drivers, and builds them into a linked system. Specifically, "Iceberg Theory" emphasizes the need for auditing to focus on the personalized content that is hidden in addition to the visible features, in terms of both structure and behavior. The "triangle theory" [1] believes that fraud is generated by the joint action of pressure, opportunity and excuse. "GONE" theory suggests that financial fraud is composed of four factors: greed, opportunity, need and exposure. The "five-dimensional fraud identification framework"[2] includes financial tax dimension, industry business dimension, corporate governance dimension, internal control dimension, and numerical characteristics dimension.

Second, introducing more fraud detection variables. Quantitative data and textual analysis are introduced[2]. On the data side, the quantitative data extends from basic financial indicators such as ratio structure to non-financial data such as inventory volume[4], and internal governance variables like board size[5], board structure, board stability[6] and board meeting frequency[7], etc. In terms of textual aspects, because the content of corporate disclosure is very likely to be the result of manipulation and careful concealment, some researches use deep machine learning methods to analyze management and audit reports[8].

Third, establishing more effective identification models and frameworks. The researches mainly focus on data algorithms and other aspects to improve predicting accuracy. From early models such as Zeta[9] and Logistic[10] to statistical analysis models developed based on data mining techniques, such as the classic Mscore[11], Fscore [12] and Cscore[13]models. With the development of big data, research has also begun to introduce artificial neural network models (ANN)[14] , machine learning based on Support Vector Machines (SVM)[15], and big data mining[16], Random Forest, Decision Tree[17] and introduction of Benford's law[18] and so on.

In summary, although existing studies have covered a lot, there is still room for innovation. Firstly, only a part of the current research selects variables based on two-factor, three-factor or four-factor theories, while those theories do not reflect the collusive relationship among accounting information. Therefore, based on the latest five-dimensional fraud identification framework, this paper develops a formation mechanism of financial reporting fraud, so as to improve the comprehensiveness and accuracy of the prediction. Secondly, this paper is aimed at manufacturing industry and its characteristic, which is more useful for identifying listed companies in China's manufacturing industry. Thirdly, this paper combines the machine learning method and other advanced technologies to utilize its predictive properties to establish a more effective financial fraud identification model.

3 RESEARCH DESIGN

3.1 Sample selection and data sources

In this paper, according to the relevant provisions of the Company Law and the Securities Law of the People’s Republic of China, the listed manufacturing companies which have been publicly penalized for fraudulent behaviors by the China Securities Regulatory Commission, the Shenzhen Stock Exchange and the Shanghai Stock Exchange, were chosen as fraud samples. Since the main ways of fraud include false statement of revenues, expenses, money funds, or costs[19], we chose the listed companies that fulfill the four types of violations in CSMAR database: fictitious profit, misrepresentation of assets, false records, and material omissions, as the companies in which financial fraud occurs. We selected the samples from 2009 to 2021 considering the impact of the 2005 Shareholding Reform and the 2008 Financial Crisis. The penalty for financial fraud has a lag of 2-3 years, therefore, we selected companies that have never been publicly penalized from 2009 to 2019 as non-fraud sample. Finally we collected 10,870 pieces of data and the following empirical analysis were conducted with the help of python.

3.2 Data pre-processing

Firstly, variables with missing values greater than 40% were deleted since filling in too much missing data would introduce more noise. Thereafter, missing values for all variables were filled in using multiple interpolation for each variable. Finally, z-score method was applied to normalize all numerical data.

3.3 Evaluation methodology

Since the financial fraud detecting model is a typical binary classification model, this paper used Accuracy, Precision, Recall, F-Score, AUC-score and Cohen’s kappa as evaluation metrics to comprehensively measure the model’s predictive performance.

4 FEATURE ENGINEERING

The selection of financial fraud characteristic variables is related to the predicting efficiency and accuracy of the final machine learning model. Huang[19] pointed out that more and more fraudulent companies have adopted complex transaction fraud techniques. Therefore, the identification of financial fraud needs to start from each link accordingly, that is, to “deconstruct” and “verify” the source and process of accounting information production.

Referring to the research of Ye[2], this paper constructed a feature set based on the five-dimensional identification framework of financial fraud. A big data perspective was combined in order to obtain more comprehensive data and variables, and machine learning methods were utilized to form a financial fraud identification index. The five dimensions include financial tax dimension, industry business dimension, corporate governance dimension, internal control dimension, and digital features dimension. The comprehensive financial fraud feature set includes 64 specific feature variables are shown in table 1.

TABLE 1. FIVE-DIMENSIONAL FEATURE SET

Dimension	**Label**	**Feature name**	**Dimension**	**Label**	**Feature name**
Financial Tax	X1	Current ratio	Financial Tax D	X34	Net profit growth rate
	X2	Quick ratio		X35	Growth rate of selling expenses
	X3	Interest coverage ratio		X36	Overhead growth rate
	X4	Asset liability ratio		X37	Sustainable growth rate
	X5	Equity multiplier		X38	Operating profit per share
	X6	Non-recurring gains and losses		X39	Liability per share
	X7	Weighted average return on net assets		X40	PE ratio
	X8	Net operating cash flows per share		X41	Tobin's Q
	X9	Basic earnings per share		X42	Enterprise value multiple
	X10	Current assets ratio		X43	Dividend yield
	X11	Fixed asset ratio		X44	Cash dividend cover multiple
	X12	Intangible assets ratio		X45	Retention rate
	X13	Current liabilities ratio		X46	Abnormal linkage between gross profit margin and closing balance of inventories
	X14	Profit from main operation percentage		X47	Income receivable
	X15	Composite tax rate	Industry Business	X48	Whether current asset turnover is off average
	X16	Accounts receivable turnover ratio		X49	Is capital intensity off average
	X17	Inventory turnover ratio		X50	Whether the growth rate of operating income is off average
	X18	Accounts payable turnover ratio		X51	Gross operating margin deviation from average
	X19	Capital intensity	Corporate Governance	X52	Nature of property rights
	X20	ROA		X53	Dual Role of the Board Chairman
	X21	ROE		X54	Size of Board of Directors (persons)
	X22	Return on invested		X55	Percentage of

		capital			independent directors (%)
	X23	Gross profit margin		X56	Shareholding ratio of the largest shareholder
	X24	Sales expense ratio		X57	Shareholding ratio of top ten shareholders
	X25	Management cost ratio		X58	Cumulative number of pledges by controlling shareholders as a percentage of shareholdings
	X26	Financial cost ratio	Internal Control	X59	Internal control index score
	X27	Net cash content of operating profit		X60	Establishment of an audit committee
	X28	Full cash recovery rate		X61	Big four accounting firms
	X29	Operating index		X62	Top ten accounting firms in China
	X30	Cash reinvestment ratio		X63	Quality of disclosure
	X31	Financial leverage	Digital Feature	X64	Characteristics of the company's net profit for three consecutive years
	X32	Business leverage		X65	Size characteristics of top five customers or suppliers
	X33	Combined leverage			

5 MODEL CONSTRUCTION AND ANALYSIS OF RESULTS

5.1 Imbalanced Processing

In an imbalanced dataset, the samples belonging to a few classes in the overlapping part of the classes are prone to be misclassified in large numbers[20].

The number of listed manufacturing companies that conducted financial fraud in this paper is relatively small compared with those without fraud. So it is necessary to do the imbalanced-processing and train the model with the processed data. Imbalanced processing methods are mainly divided into three kinds: undersampling, oversampling, and integrated sampling. The most representative algorithms include oversampling algorithm SMOTE[21], undersampling algorithm ClusterCentroids[22], and comprehensive sampling algorithms SMOTE+ENN and so on. In this paper, the oversampling algorithm SMOTE, undersampling algorithm ClusterCentroids, integrated sampling algorithm SMOTE+ENN and SMOTE+tomek were selected to process the data to obtain balanced sample data, and the optimal sampling method was selected by comparing the prediction results.

5.2 Model training results

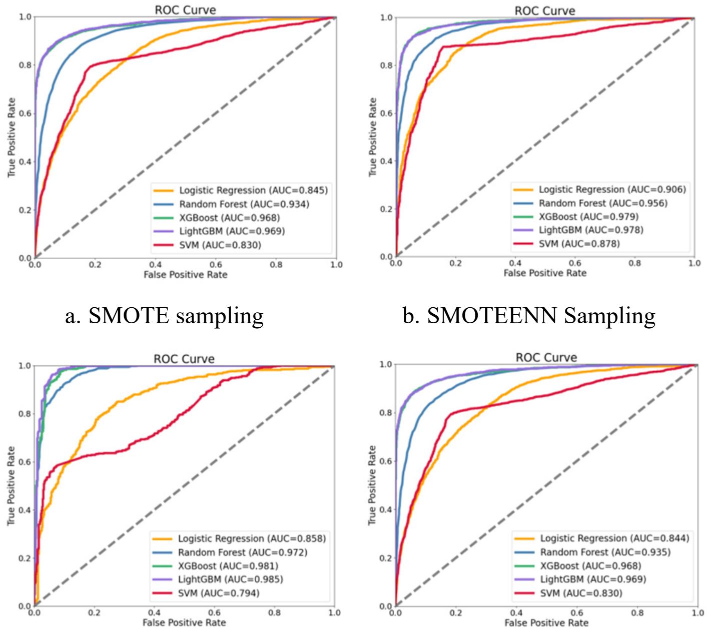
In this paper, we used different sampling methods such as logistic regression, random forest, support vector machine (SVM), XGboost and LightGBM to train the models, and finally got the evaluation indexes of different models under the four sampling methods. The index scores are shown in Table 2. Overall, the logistic regression algorithm performed relatively worse. The SVM algorithm performed poorly on the oversampled dataset, possibly because the algorithm's performance is only affected by the support vector, and the addition of repetitive samples may not be able to optimize the algorithm's results. Random Forest, XGboost, and LightGBM had significantly better predictive performance than the other two algorithms. The main purpose of financial fraud detection is to find as many fraudulent companies as possible, so we focused on the recall and F1-score of the model. Through comparison, we found that XGboost performs the best on the integrated sampled dataset using SMOTE+ENN, so we chose the XGboost+SMOTEENN model for further optimization and tuning.

TABLE 2. RESULTS OF THE MODEL ASSESSMENT INDICATORS

	Sampling method	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUC socre**	**Cohen's kappa**
Logistic Regression	RawData	0.8669	0.6897	0.3048	0.4244	0.8269	0.3620
	SMOTE	0.7597	0.7700	0.7405	0.7550	0.8447	0.5194
	SMOTE ENN	0.8373	0.8718	0.8679	0.8698	0.9060	0.6529
	Cluster Centroids	0.7810	0.7909	0.7638	0.7771	0.8583	0.5619
	SMOTETomek	0.7599	0.7703	0.7407	0.7552	0.8444	0.5198
Random Forest	RawData	0.8749	0.7112	0.3752	0.4913	0.8610	0.4276
	SMOTE	0.8450	0.9076	0.7683	0.8321	0.9337	0.6901
	SMOTE ENN	0.8824	0.9388	0.8690	0.9025	0.9557	0.7550
	Cluster Centroids	0.9086	0.9070	0.9105	0.9087	0.9716	0.8171
	SMOTE Tomek	0.8482	0.9107	0.7721	0.8357	0.9354	0.6964
SVM	RawData	0.8393	1.0000	0.1905	0.0038	0.8140	0.0031
	SMOTE	0.5179	0.9712	0.0369	0.0711	0.8300	0.0358
	SMOTE ENN	0.4348	0.9710	0.1006	0.1823	0.8783	0.0733
	Cluster Centroids	0.6562	0.6088	0.8743	0.7177	0.7945	0.3123
	SMOTE Tomek	0.5179	0.9712	0.0369	0.0712	0.8297	0.0358
XGboost	RawData	0.8914	0.7342	0.5105	0.6022	0.8841	0.5417
	SMOTE	0.9064	0.9448	0.8633	0.9022	0.9681	0.8129
	SMOTEENN	0.9273	0.9652	0.9170	0.9405	0.9787	0.8474
	Cluster Centroids	0.9352	0.9193	0.9543	0.9364	0.9809	0.8705
	SMOTETomek	0.9091	0.9472	0.8665	0.9051	0.9676	0.8182
LightGBM	RawData	0.8893	0.7455	0.4743	0.5797	0.8882	0.5196
	SMOTE	0.9105	0.9478	0.8688	0.9066	0.9694	0.8209

	SMOTE ENN	0.9281	0.9607	0.9197	0.9402	0.9783	0.8486
	Cluster Centroids	0.9495	0.9403	0.9600	0.9308	0.9846	0.8990
	SMOTE Tomek	0.9115	0.9450	0.8738	0.9080	0.9687	0.8230

This paper plotted the ROC curves of five models under four different sampling methods to show a clearer differences in the prediction abilities of several types of models. The vertical coordinate of the ROC curve is the True Positive Rate, which is also known as the Recall Rate. The horizontal coordinate is the False Positive Rate, and the area under the curve is the AUC-score. The closer the curve is to vertical axis, the better the model's ability to identify financial anomalies in listed companies. As can be seen in Figure 1, the AUC values of the two integrated algorithms, LightGBM and XGBoost, are much higher than others, and the difference between the two is almost negligible.



a. SMOTE sampling

b. SMOTEENN Sampling

c. Cluster Centroids sampling

d. SMOTETomek sampling

Figure 1. Comparison of ROC curves

5.3 Parameterization

For the selected SMOTEENN+XGBoost model, the weight, cover and gain feature importance scores obtained from training were output. The features with very low scores were excluded: X61 “Whether the Auditor is Big four International Auditing Firms”, X27 “Net Cash Content of Operating Profit”, and X42 “Enterprise Value Multiple”. It indicated that although companies may choose non-Big four auditing firms to cover up counterfeiting, the profitability and value multiples may generate counterfeiting motives. These three features were not well separated, which have a high probability to occur in non-counterfeiting samples as well.

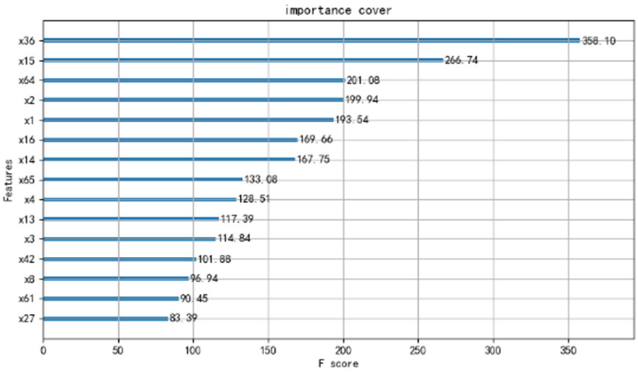
Therefore, the remaining 58 features were retained and tuned on the selected SMOTEENN+XGBoost model.

The performance of the model on the test set before and after the tuning is shown in table 3. It shows that the performance of the optimized model was improved in all indicators. The number of misjudgments for both Type I and Type II errors was reduced, and the model generalization ability was enhanced.

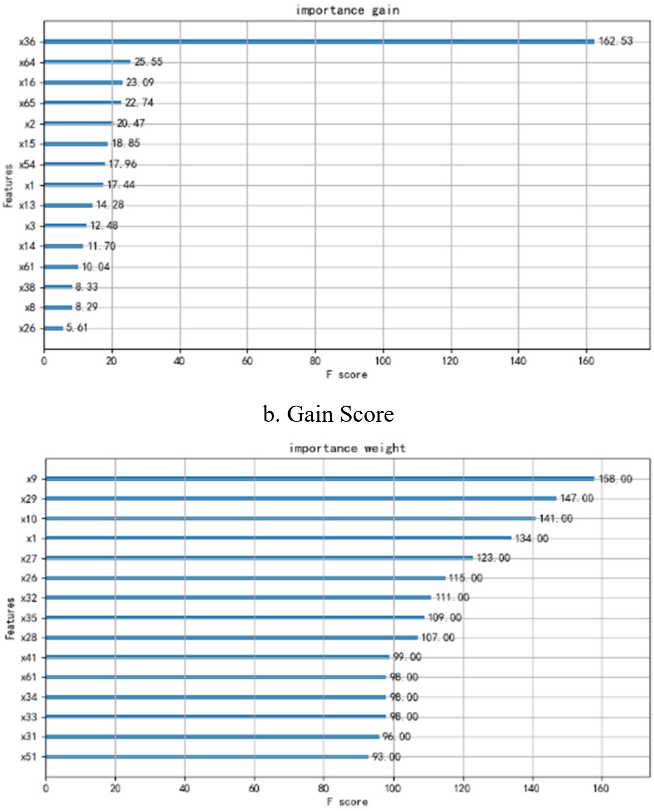
TABLE 3. RESULTS BEFORE AND AFTER MODEL OPTIMIZATION

	Accuracy	***Precison***	***Recall***	***F1score***	***AUC***
pre-optimizing	0.9273	0.9652	0.9170	0.9405	0.9787
post-optimizing	0.9300	0.9674	0.9224	0.9429	0.9812

Figure 2 shows the top fifteen feature importance scores. Among them, X20 "ROA" and X46 "Annual Gross Profit Margin and Inventory Ending Balance Linkage" are ranked in the top two of both cover and gain. The reason may be that investors tend to pay attention to the effect of profit realization related to input assets when assessing the achievement of corporate profit target and judge it through the indicators of earnings per share, return on net assets and return on total assets. In order to attract more investors, listed companies may engage in financial fraud, so ROA plays an important role in identifying financial fraud. Generally speaking,, inventory and gross profit margin is a negative correlation, so the Abnormal Linkage Between Gross Profit Margin and Closing Balance of Inventories is important for fraud detection. The importance score of X65 "Revenue Receivable in the financial tax dimension" is high, which confirms that inflating or recognizing revenue in advance are the most common means of financial fraud. The importance score of X59 "Internal Control Index Score" in the Internal Control dimension is higher, which is probably because the Internal Control Index Score reflects characteristics such as the difficulty and cost of conducting financial fraud in listed companies.



a. Cover Score



b. Gain Score

c. Weight score

Figure 2 Ranking of feature importance scores (top 15)

6 Conclusion

In this study, we took China's A-share listed companies in the manufacturing industry as an example, and combined the five-dimensional fraud identification framework proposed by Ye Qinhua et al.[2] to construct a more comprehensive research framework for financial fraud detection in listed manufacturing companies. The experimental results showed that XGboost and LightGBM model performed best on the dataset processed by SMOTE+ENN integrated sampling as well as ClusterCentroids undersampling. By deriving the importance of the features, it was found that the indicators of “ROA”, “Linkage between annual gross profit margin and inventory closing balance”, “Accounts receivable”, “Internal control index” have a significant guiding role in identifying financial fraud in the manufacturing industry. From the feature dimension, a number of linkage features in financial and tax dimension were at the top of the importance ranking, which confirmed the linkage anomaly analysis proposed by Yeh Chin-Hua's study. Meanwhile, among the top 15 importance scores, “Internal control index” and “Gross operating margin deviation” came from the internal control dimension and industry business dimension respectively, which proved that the proposed five-dimensional financial fraud identification framework has a reference significance for constructing the financial fraud characterization project.

Acknowledgments. Sponsors:National Undergraduate Training Program for Innovation and Entrepreneurship(202310610160)

REFERENCES

[1] Albecht，W. S.，C. Albecht. 2004. Fraud Examination & Prevention.South－Western.

[2] YE Q H,YE F,HUANG S Z. Construction of Financial Fraud Identification Framework - Based on Accounting Information System Theory and Big Data Perspective[J]. Accounting Research,2022,No.413(03):3-16.

[3] Purda, L., D. Skillicorn. 2015. Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. contemporary Accounting Research, 32 (3) : 1193- 1223.

[4] Ye K T,Liu J Y. Non-financial information and identification of corporate financial fraud[J]. Accounting Research,2021(09):35-47.

[5] LIU L G,DU Y. An empirical study of the relationship between corporate governance and accounting information quality[J]. Accounting Research,2003(02):28-36+65.

[6] YANG Q X,YU L,CHEN N. Board Characteristics and Financial Fraud - Empirical Evidence from Chinese Listed Companies[J]. Accounting Research,2009(07):64-70+96.

[7] Anderson, Roald C.; Mansi, Satar A.; Reeb, DavidM.2004.Board Characteristics, Accounting Report Integrity, and the Cost of Debt. Journal of Accounting & Economics, 37 (3):315~342.

[8] MATIN R, HANSEN C,HANSEN C,et al.Predicting distresses using deep learning of text segments in annual reports[J].Expert Systems with Application,. 2019,132: 199-208.

[9] ALTMAN E I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy[J].The Journal of Finance, 1968 ,23(4) :589-609.

[10] OHLSON J. Financial ratios and the probabilistic prediction of bankruptcy[J].Joumal of Accounting Research, 1980, 18 (1):109-131.

[11] Beneish, M. D. 1999. The Detection of Earnings Manipulation. Financial Analysts Journal, 55 (5) : 24-36.

[12] Dechow, P.M, W . Ge , C.R.Larson,R.G.Sloan.2011.Predicting Material Accounting Misstatements. Contemporary Accounting Ｒesearch，28 (1) : 17- 82.

[13] Khan, M. and Watts, R.L. (2009) Estimation and Empirical Properties of a Firm-Year Measure of Accounting Conservatism. Journal of Accounting and Economics, 48, 132-150.

[14] Green, Brian Patrick; Choi, Jae Hwa. Auditing: a Journal of Practice & Theory. Spring97, Vol. 16 Issue 1, p14-28. 15p. 2 Diagrams, 4 Charts.

[15] Cecchini,M., H. Aytug, G. J. Koehler, P. Pathak. 2010. Detecting Management Fraud in Public Companies. Management Science, 56 (7): 1146-1160.

[16] Durtschi, C.,W. A.Hillison, C. Pacini. 2004. The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data.Journal of Forensic Journal of Forensic Accounting, 5 (1) :17- 34

[17] WANG G, WANG K M,ZHOU Y Y, et al. Establishment of a financial crisis early warning system for domestic listed companies based on three decision tree models [J]. Mathematical Problems in Engineering , 2020: 8036154.

[18] Amiram, D., Z. Bozanic, E. Rouen. 2015. Financial Statement Errors: Evidence from the Distributional Properties of Financial Statement Numbers. Review of Accounting Studies, 20 (4) : 1540-1593.

[19] Huang S.Z.,Huang J.C.2004.An overview of financial statement fraud behavioral characteristics and early warning signals. Accounting Newsletter, 23: 4-9.

[20] LI Y X,CHAI Y,HU Y Q et al. A review of classification methods for unbalanced data[J]. Control and Decision Making,2019,34(04):673-688.

[21] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over- sampling technique[J]: 321-357.

[22] Cheng X F, Li J, Li X F. An imbalanced data classification algorithm based on under sampling[J].