Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 8, No. 3, 2023

Financial Statement Fraud Detection based on Integrated Feature Selection and Imbalance Learning

Yinhe Chen[1, *]

[1]School of Economics and Management, Hebei University of Technology, Tianjin, China
[*]**Corresponding author:** cyh_7928@126.com

Abstract: Based on the data mining technology, this paper proposes an integrated feature selection method to construct the financial statement fraud detection feature system of listed companies, uses the SMOTE algorithm to solve the class unbalanced distribution problem, and combines the machine learning algorithm models to construct the financial statement fraud detection model. Based on the real financial statement data of Chinese listed companies, the empirical analysis is conducted to provide support for the auditors. The integrated feature selection framework proposed in this paper improves the problem of poor generalization of the single feature selection method, and the SMOTE effectively strengthens and improves the ability of the model to detect financial statement fraud of listed companies.

Keywords: Financial statement fraud, Fraud detection, Feature selection.

1. Introduction

With the development of machine learning, scholars use machine learning algorithms to build models to detect financial statement fraud of listed companies and assist auditors in their work. There are two challenges in financial statement fraud detection: on the one hand, there is a high correlation and redundancy between financial statement features, which reduces the efficiency and effect of the model; On the other hand, financial statement fraud has a serious problem of unbalanced class distribution, which will seriously affect the fraud detection ability of the classification model. To solve the above problems, firstly, this paper proposes an integrated feature selection method to eliminate redundant indicators and optimize the overfitting risk of the single feature selection method. Secondly, an overfitting method SMOTE is used to carry out class balance processing. Finally, constructing the financial statement fraud detection model through advanced machine learning algorithms. The effectiveness of the model in detecting financial statement fraud is verified by real data of Chinese listed companies.

2. Literature References

The study of financial statement fraud often carried out the empirical analysis of financial statement fraud enterprises, so as to analyze the signals of financial statement fraud in listed companies for auditors and investors.

Albrecht and Romney et al. [1] first proposed the "red flag" in 1986, by analyzing the features of fraudulent companies, such as board size and director change. Kinney and McDaniel [2] found that companies in financial distress are more likely to commit financial fraud. Gozman and Currie [3] found that if incentives are present, the likelihood of financial statement fraud increases, which usually takes the form of overstating gains or concealing losses. Persons[4] constructed a stepwise regression model to analyze the fraud characteristics of listed companies, concluded that industry would affect financial statement fraud, and found that fraud companies had higher financial leverage, lower capital turnover, and higher current asset ratio than non-fraud companies. Beasley [5] used logistic regression on 150 sample companies to study the relationship between board members and financial statement fraud, concluding that the size of the board was positively correlated with the possibility of fraud.

Feature selection is necessary for the construction of the feature system [6]. Feature selection affects the performance of financial statement fraud detection by eliminating redundant features or features with little predictive information [7-8].

3. Method

3.1. Integrated Feature Selection

Existing studies often use the embedded method for feature selection. Embedded refers to the selection of the feature set that best improves the accuracy of the given model. However, the embedded method may cause overfitting. In order to improve the generalization of feature selecting, this paper proposes an integrated feature selection method to rank the feature importance, which combines four types of ensemble algorithms commonly used in feature importance calculation, including random forest, gradient boosted decision tree (GBDT), eXtreme gradient boosting (XGBoost) and light gradient boosting machine (LightGBM).

The process of calculating the feature importance of the input features by random forest is as follows [9]:

a) The training set is randomly sampled to form in-bag data, which is used to construct the random forest model, and the unextracted data forms out-of-bag data (OOB).

b) Use the corresponding OOB data to test the performance of each decision tree i in the constructed random forest, and obtain the data error of OOB $error_i^{'j}$.

c) Randomly change the value of feature j in OOB and test the $error_i^{'j}$ again with the new OOB data.

d) Calculate the feature importance of j according to $VI(x^j) = \sum_{i=1}^{N}(error_i^{'j} - error_i) \Big/ N$, where N is the number of decision trees.

GBDT's process of evaluating feature importance is as follows [10] :

a) Calculate the change of the Gini index $\Delta Gini = Gini_m - Gini_l - Gini_r$ before and after branch division at node *m* of the decision tree *i*.

b) The importance of feature *j* at node *m* is $VI(x_m^j) = \Delta Gini * (n_m / n)$, where n_m is the sample size of node *m*.

c) If the node of feature *j* is in the set *M*, then the importance of feature *j* in the decision tree *i* is $VI\left(x_i^j\right) = \sum_{m \in M} VI\left(x_m^j\right)$. d) Suppose there are *N* decision trees, then $VI\left(x^j\right) = \sum_{i=1}^{N} VI\left(x_i^j\right)$.

XGBoost's and LightGBM's feature evaluating processes are like GBDT, but exist some differences, especially in efficiency.

3.2. SMOTE

Synthetic Minority Oversampling Technique (SMOTE) is an oversampling method [11]. The generation process of SMOTE algorithm is as follows.

a) For each sample *x* in the minority class, the Euclidean distance is used as the standard to calculate the distance to all samples in the minority class, and its k-nearest neighbors are obtained.

b) A sampling ratio is set according to the sample imbalance ratio to determine the sampling rate *N*. For each minority sample *x*, a number of samples are randomly selected from its k-nearest neighbors, and the selected neighbors are assumed to be x_n.

c) For each randomly selected neighbor x_n, a new sample is constructed with the original sample according to $c = a + rand\ (0, 1) * |a - b|$.

4. Empirical Results

4.1. Dataset

This paper selects non-financial Chinese listed companies from 2010 to 2020 as the research samples, obtaining 30,409 samples. A total of 1,221 listed companies that were punished by the CSRC for financial fraud were used as fraud samples, which mainly involve five types of violations: fictitious profits, false assets, false records, major omissions, and false disclosure. In order to comprehensively obtain relevant indicators, 296 indicators related to the operating conditions, internal governance, and financial statements are pre-selected as feature sets.

4.2. Evaluation Metrics

Firstly, the confusion matrix of the binary classification problem is given, as shown in Table 1. Where *TP* represents the number of positive classes classified as positive, *FN* represents the number of positive classes misclassified as negative, *FP* represents the number of negative classes misclassified as positive, and *TN* represents the number of negative classes classified as negative.

Table 1. Confusion matrix of binary classification

Actual \ Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Recall is the percentage of instances where the true class is positive that are correctly classified, which means the detection ability of fraud sample.

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

The Receiver Operation Curve (ROC) is the curve between the true positive class rate and the false positive class rate at all critical values. *AUC* is the area under the ROC and ranges from 0.5 to 1, which evaluates the overall performance of the model.

4.3. Results and Comparison

Firstly, in order to verify the effect of the integrated feature selection method, this paper compares the model effects of the four types of feature selection methods used separately and the integrated method, and the results are shown in Table 2. The results show that the integrated feature-selecting method has the strongest ability to identify financial statement fraud, and can combine the advantages of the single feature-selecting method well, with better *AUC* and *Recall*.

Table 2. Effect comparison of feature-select methods

	RF	GBDT	XGB	LGB	Integrated method
AUC	0.6923	0.6975	0.7007	0.6925	0.7087
Recall	0.4138	0.4220	0.4301	0.4111	0.4598

In order to obtain a financial statement fraud detection model with better performance, this paper compares the effect of the existing advanced machine learning models, including Random Forest, GBDT, XGBoost, and LightGBM, and the results are shown in Table 3. It shows that SMOTE overfitting effectively solves the imbalanced distribution problem and significantly improves fraud detection ability. Among the four machine learning algorithms selected, GBDT performs best both in *AUC* and *Recall*, which means that it has a stronger fraud detection ability.

Table 3. Effect comparison of different models

Model	Recall	AUC
RF	0.2776	0.6325
RF+SMOTE	0.6055	0.7610
GBDT	0.4573	0.7101
GBDT+SMOTE	0.7224	0.7672
XGB	0.4598	0.7087
XGB+SMOTE	0.6646	0.7310
LGB	0.4611	0.7102
LGB+SMOTE	0.6947	0.7501

5. Summary

Aiming at the problem of financial statement fraud detection, this paper proposes an integrated feature-selecting method to construct a financial statement fraud detection feature system and uses SMOTE algorithm to oversample the data solving the imbalance distribution of financial statement fraud data. Through the real data verification of Chinese listed companies, it is found that the model proposed in this paper has a good ability to detect financial statement fraud, which can effectively provide decision support for auditors.

References

[1] Albrecht W, Romney M. Red-flagging management fraud: A validation. Advances in Accounting, 1986, 3: 323-333.

[2] Kinney W R, McDaniel L S. Characteristics of firms correcting previously reported quarterly earnings. Journal of Accounting and Economics, 1989, 11(1): 71-93.

[3] Gozman D, Currie W. The role of Investment Management Systems in regulatory compliance: a Post-Financial Crisis study of displacement mechanisms. Journal of Information Technology, 2014, 29(1): 44-58.

[4] Persons O S. Using financial statement data to identify factors associated with fraudulent financial reporting. Journal of Applied Business Research, 1995, 11(3): 38.

[5] Beasley M. An Empirical Analysis of the Relation between the Board of Director Composition and Financial Statement Fraud. The Accounting Review, 1996, 71(4): 443-465.

[6] Hu L, Gao W, Zhao K, et al. Feature selection considering two types of feature relevancy and feature interdependency. Expert Systems with Applications, 2018, 93: 423-434.

[7] Ravisankar P, Ravi V, Raghava Rao G, et al. Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems, 2011, 50(2): 491-500.

[8] 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. Applied Soft Computing, 2021, 108(3): 107487.

[9] Breiman, L. Random Forests. Machine Learning ,2001, 45: 5–32.

[10] Friedman J. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 2001, 29(5): 1189-1232.

[11] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321–357.