Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detecting Financial Statement Fraud with Interpretable Machine Learning

Zhongzhu Liu (zhongzhuliu@126.com)
Huizhou University
Rongguang Ye
Guangdong University of Technology
Rongye Ye
Huizhou University

Research Article

Keywords: Machine Learning, Financial Statement Fraud , imbalanced datasets

Posted Date: June 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-640038/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Detecting financial statement fraud with interpretable machine learning

Zhongzhu Liu[1,*,+], **Rongguang Ye**[2,+], **and Rongye Ye**[1,+]

[1]School of Mathematics and Statistics, Huizhou University, Guangdong Province, China
[2]School of Applied Mathematics, Guangdong University of Technology, Guangdong Province, China
[*]zhongzhuliu@126.com
[+]these authors contributed equally to this work

ABSTRACT

In this study, we explored a stable and explainable model in the detection of financial fraud. To effectively handle imbalanced datasets, we selected the Smote oversampling algorithm with the highest AUC value and compared it with Borderline Smote and ADASYN algorithms. Using the MCB method, we found that the Adaptive Lasso algorithm had higher stability than SCAD, MCP, Stepwise, and SQRT Lasso algorithms. Moreover, the AUC value was improved by WoE encoding and IV value testing of the features. Finally, we ranked the fraud factors based on the importance of the features, and the partial dependence function was used to make the model interpretable. By comparing the AUC and KS values, the integrated models XGBoost, LightGBM, and RF showed better ability to identify financial fraud compared with traditional models such as SVM and LR.

Introduction

The detection of financial fraud is a global challenge. The China Securities Regulatory Commission has handled 59 illegal acts of financial fraud cases of listed companies since 2020, accounting for 23% of information disclosure cases. With the development of information technology, financial fraud mainly presents the following characteristics: Firstly, the counterfeiting model is very complex, and is mainly manifested in the use of fictional business to implement systematic financial data fraud, for example, Wisdom Shanghai School, a subsidiary of Aerospace Communications, was a fictional business linked to purchase and transportation for more than two years. Secondly, the forms of counterfeiting are diverse, for example, Yu Diamond and other companies falsely increased their profits through the circulation of their funds, false sales, and losses in subsidiaries. Besides, they failed to disclose a total of 1 billion CNY in external guarantees and related transactions as required by the law. In all such cases, serious fraudulent cases account for more than 1/3, with large amounts of fraud involved. In , Luckin Coffee admitted that their financial data were falsified, and the total amount of fraud reached 2.2 billion CNY between the second quarter and fourth quarter. Besides, other related expenses were also falsely increased. Most listed companies continue to engage in financial fraud, making the auditors' work complex. Therefore, the need to effectively and scientifically detect financial fraud has contributed to significant research in this study area.

Previous research mainly focused on the use of machine learning methods and data mining techniques such as regression analysis, Feedforward Neural Networks, Support Vector Machines, and Big Data Distributed Systems to detect fraud in financial statements[1]. However, these studies did not make any specific analysis on the interpretability of the characteristics after selecting the fraud factors, and could also not effectively explain the specific impact of the fraud factors on the prediction results. Previous studies have also not presented any detailed analysis of the fraud factors. The algorithm used in the selection was also used for instability detection, and the instability of the algorithm was in an unknown state. However, to effectively explain the selected fraud factors and detect the instability of the algorithm is an important issue at present. This study identified the characteristics of fraud factors, analyzed the interpretability of the selected fraud factors, and detected the instability of the algorithm used in feature selection, to provide solutions on how to deal with the interpretability of the fraud factors and algorithmic performance. There exists a challenge in instability analysis, thus, the interpretability of financial fraud factors may provide auditors and the Securities Regulatory Commission with an effective way of detecting financial fraud. Data mining technology plays an important role in financial fraud detection. Chi-Chen Lin et al.[2] utilizes the logistic regression model and data mining technology to study the importance of financial fraud factors, to discover and extract hidden objective facts and information.

The purpose of this study was to use the MUC curve to determine the instability of five regression algorithms used in feature selection of invalid fraud factors and identify a more stable algorithm to eliminate fraud factors that have a weak impact on the results. The AUC and KS values of the integrated models, LGBM, XGB, RF, SVM, and LR were compared with those of the

traditional models, SVM and KS. The model with a good generalization ability and high prediction accuracy was selected. Chi-Chen Lin et al.[2] utilized logical regression to rank selected fraud factors according to their importance. However, there was no specific analysis made on how the fraud factors affected the test results, which is important for the interpretability of the features in the model obtained by machine learning and data mining.

The second study's purpose was to make an interpretable analysis of the selected feature variables. Here, we used WoE coding combined with the IV value test for feature engineering and processing on the data samples[3]. In addition, we used LightGBM and XGBoost models to rank the features based on the importance of the fraud factors. The PDPbox library in Python was used to draw local dependence graphs to determine the impact of fraud factors on the detection results. These methods were used to build a highly explanatory and relatively stable model, to efficiently detect financial fraud. In addition, the data samples selected from most previous studies were of good quality, and there is no detailed comparative analysis on the extreme imbalance of positive and negative proportion of sample data in this kind of problem, and the treatment of missing values is not discussed. However, in original sample data obtained in real life, solving the above problems is often an important step. Thus, to make up for this deficiency and build a model that can easily adapt to such a complex situation, multiple interpolation methods based on the random forest were used to fill the missing values present in the data selected in this study. Different oversampling methods were scientifically compared, and the results of this study can provide a reference point for future follow-up research.

Literature review

Data mining methods, including hidden Markov model, feedforward neural network, Bayesian belief network, Gradient Boosted Tree, random forest, genetic algorithm, and text mining technology are widely used to identify financial fraud. The following section will review the research findings of financial fraud detection based on data mining methods.

Kirkos et al.[4] reports that losses caused by financial fraud in American enterprises are estimated to be more than $400 billion, annually. W.Xu et al.[5] compared the detection ability of credit card fraud between SVM and the ID3+BP hybrid model. The results showed that SVM performed better compared with the ID3+BP hybrid model in detecting financial fraud. The study by X.Li et al.[6] showed that the support vector machine model had a higher accuracy of 86.612%, in detecting financial fraud, and the logical regression accuracy was 83.036%. J. Liang et al.[7] reported that the FGABPN method had a high detection accuracy of financial fraud.

Huang et al.[8] developed a new fraud detection method based on Zipf's 's law. The purpose of this method was to help financial auditors effectively review a large number of data samples and detect any hidden fraud records. A.A. Rizki and I. Surjandari et al.[9] established the SVM and artificial neural network models after feature selection of fraud factors, to detect whether there was fraud in financial statements. The results showed that feature selection of financial fraud factors helped to improve the accuracy of the SVM model, with an accuracy of 88.37%, while the artificial neural network had the highest feature selection accuracy of 90.97%.

V.Bhusari and S.Patil[10] used the HMM model to detect credit card fraud. The study revealed that HMM had high coverage of fraud detection, 84% of which were fraudulent, and a 77% false-positive rate which was very low. Calderon and Cheh[11] adopted a diversified approach and used a neural network method to conduct in-depth research in the field of financial audit and risk assessment, based on deep learning theory. Besides, they also made an in-depth expansion of the neural network modeling theory and studied the use of neural network methods for financial risk assessment.

R.Bauder and R.da Rosa et al.[12] showed that the detection performance of LOF was the best, while that of Autoencoders, KNN, and 5-neighbors was poor in detecting medical insurance fraud. Huang S. Y. et al.[13] studied the effectiveness of the Growing hierarchical self-organizing map method in financial fraud detection and showed that the method had a good application prospect, by comparing GHSOM with other classification methods, such as neural network, SVM, GHSOM+LDA, BP neural network, SOM+LDA and so on. Q.Dengjue and G. Mei et al.[14] showed that a combination of V-KSOM was better compared with the SOM method alone, and V-KSOM performed better in detecting financial fraud.

K.Behera and S.Panigrahi et al.[15] used Fuzzy Clustering and neural network analysis to detect credit card fraud, and the results showed that the combined use of the Fuzzy Clustering method improved the TP by 93.90%. In addition, the FP of this method was less than 6.10%. Feroz et al.[16] found that using a neural network to train samples can achieve better results, and this is because the neural network model can "learn" what is relatively important. Compared with the traditional data mining methods, the neural network model adopts an adaptive learning process to accurately judge the importance of the detection target. Therefore, the neural network model is relatively stable in detecting financial fraud. H. L. Etheridge[17] and other researchers have used the neural network method to detect financial fraud, and have reported high performance in detection. Ramamoorti[18] analyzed the model architecture of multi-layer perceptron and compared this model with the model used by Delphi. They found that internal auditors greatly benefited from using the neural network model for risk assessment. S. Subudhi used the ADASYN method for oversampling, and DT and SVM models to predict the results. The results showed that the sensitivity of the SVM model was 94.74% and the sensitivity of the DT model was 94.52%. Q. Deng[19] found that the

financial statement data released by the studied company contained false indicators. I.Sadgali determined the TPR, accuracy, and sensitivity of the hybrid model, and found that this method performed better than the traditional model method. Busta et al.[20] used the neural network model to distinguish between "normal" and "manipulated" financial data, through an in-depth study of the data distribution of hidden financial information. Data samples were analyzed based on Benford's law, which points out that naturally occurring numbers have a specific pattern distribution. A total of six types of neural network models were analyzed and compared to determine the most effective model. The results showed that the neural network model achieved an average accuracy of 70.8%.

C. Yan et al.[21] utilized the outlier detection method based on the nearest neighbor in their study and found that the improved algorithm had higher accuracy, reduced the time complexity, and reduced the interference of the model to the K value. I. Benchaji and S. Douzi et al.[22] utilized the K-means Clustering and GA models to detect credit card fraud. The results showed that the use of the K-means Clustering model and genetic algorithm improved the recognition of credit card fraud detection, thus effectively reducing the number of false prompts. Hajek P et al.[23] compared the effectiveness of six models, including DT, SVM, Bayesian classifiers, LR, ensemble classifiers, and neural networks in detecting financial fraud. The results showed that the Bayesian belief network was superior to the other five models. Huangjun Zhou et al.[24] proposed the use of big data mining technology based on distributed algorithms to identify and analyze fraud in the supply chain. Spakr and Hadoop evaluated a deep learning method based on a convolutional neural network, which was reported to significantly reduce the time required for the recognition process and also improved the accuracy of financial fraud detection.

Methodology

This study utilized the Teddy database to collect data sets of Chinese listed companies between 2012 and 2017, including 363 financial-related variables, some of which are listed in Table 1. Among them, 11219 samples were financial fraud companies, while 91 samples were non-financial fraud companies. First, we eliminated the variables whose variance was 0 or the missing rate was more than 50%. The multiple interpolation methods based on the random forest is a new, high-performance missing value filling algorithm, with good performance in high-dimensional data[25]. In this study, the random forest multiple interpolations were used to interpolate the remaining missing data. Besides, we first considered whether the missing values were randomly missing before interpolation. Then, we counted the number of missing values contained in each sample as additional variables. The ratio of positive and negative samples adopted in this study was about 1:123. When the samples in the data set are highly unbalanced, this affects the performance of the classifier. Buda's research suggests that oversampling is a good way to solve the highly imbalanced data[26]. GBDT (Gradient Boosting Decision Tree) is an integrated classifier, with high flexibility to handle various data types and high performance[27]. In this study, the most suitable method for this data was selected from a variety of oversampling methods through the GBDT classifier, among which the oversampling methods included Smote, ADASYN, Borderline Smote.

Following the processing illustrated above, there were 94 features in the dataset, thus, it was necessary to reduce their dimensions (reduce dimensions) to improve the computing speed and model optimization. Yang Li et al proposed an MCB (Model Confidence Bounds) method to detect the instability of the algorithm, and effectively compare common features in the selection methods. Therefore, we used the MCB method to select the feature selection method most suitable for the data set using the R language[28].

Feature construction plays a very significant role in improving the performance of classifiers. Therefore, in this study, the decision tree sub-box was first carried to the feature, and then the WOE coding to the feature. To verify whether the constructed feature was qualified, the IV value test was used to filter the unqualified coding feature. In addition, we used two integrated learning models, XGBoost and LightGBM, to determine whether the company had fake data since the two models perform particularly well in the classification of tasks. Finally, we obtained the order of feature importance through the two models and drew a Partial dependence diagram to analyze the influence of the features.

XGBoost

XGBoost innovates the loss function, and its second-order Taylor expansion can make a better approximation of the loss function of the model. Besides, XGBoost is a distributed lifting tree model, which supports large-scale computing[29], and in practical applications, most of the inputs are sparse. In this regard, XGboost uses the sparse sensing algorithm, which can accept a large number of sparse data and efficiently perform the calculations. Studies show that sparse sensing algorithm is dozens of times faster than the traditional methods. In addition, XGBoost not only punishes the number of leaves but also adds weight punishment when pruning. By improving regularization, it can reduce the variance of the model, prevent overfitting, control the complexity of the model, and make the training model more simple. Compared with the linear model, XGBoost can also deal with the characteristics of different dimensions, and also deal with outliers in the data and nonlinear decision boundary problems.

Table 1. Some variables related to financial data

No.	Variables	No.	Variables	No.	Variables
1	PUBLIC DATA	14	CAPITAL RESER	27	PUR FIX ASSETS OTH
2	NOTES RECEIV	15	SURPLUS RESER	28	INVEST INCOME
3	PREPAYMENT	16	RETAINED EARNINGS	29	FINAN EXP
4	INVENTORIES	17	REFUND OF TAX	30	SELL EXP
5	AVAIL FOR SALE FA	18	GAIN INVEST	31	COGS
6	FIXED ASSETS	19	FOREX EFFECTS	32	BIZ TAX SURCHG
7	INTAN ASSETS	20	DISP FIX ASSETS OTH	33	NOPERATE EXP
8	DEFER TAX ASSETS	21	ADMIN EXP	34	INCOME TAX
9	NOTES PAYABLE	22	DILUTED EPS	35	NOPERATE INCOME
10	ADVANCE RECEIPTS	23	REVENUE	36	MINORITY GAIN
11	PAYROLL PAYABLE	24	BASIC EPS	37	END DATE
12	TAXES PAYABLE	25	OPERATE PROFIT	38	CASH C EQUIV
13	DEFER REVENUE	26	ASSETS IMPAIR LOSS	39	OTH RECEIV

LightGBM

LightGBM utilizes a histogram algorithm to convert continuous features into discrete Bins, which not only reduces the Gain of each split node but also reduces the use of memory. LightGBM uses GOSS (Gradient base One-Side Sampling) technology, to enhance its advantage in dealing with large amounts of data. In most cases, the performance of the model trained by the GOSS algorithm is better than that of the ordinary random sampling algorithm. On the other hand, the GOSS method also increases the diversity of the base learner, which inherently improves the generalization ability of the model. In addition, it uses the growth strategy of Leaf-wise to reduce the amount of computation and prevent over-fitting by controlling the maximum depth of the tree. Besides, LightGBM also supports efficient parallel computing of features and data. These advantages make LightGBM greatly reduce computing time while ensuring high performance, and it is also particularly outstanding in classification and prediction tasks[30]. For this reason, this study utilized the LightGBM model to obtain efficient results.

SVM

Support vector machine is a classification and prediction model, which is divided into linear support vector machine and nonlinear support vector machine. In linear support vector machine, it is divided into soft interval classifier and hard interval classifier. Nonlinear support vector machine leads to kernel function for simplified operation, and the data of nonlinear distribution is mapped to high-dimensional space to become linear distribution. Support vector machine with kernel function is greatly improved in computing ability, which makes it an efficient classifier. At present, SVM is used to detect financial fraud and credit card fraud, classified forecasting, and other fields. Therefore, SVM has good prediction and classification abilities in these fields[31].

Random Forest (RF)

Random forest is a classification and prediction model containing multiple decision trees, with a good ability to deal with samples with higher dimensions, and it is widely used in classification and prediction methods[32]. The random forest has the advantages of the decision tree, it is easy to set the parameters of the model, the speed is fast in the learning process, the time complexity of calculation is low, and the prediction accuracy for classification problems is high.

Logistic Regression (LR)

Logistic Regression algorithm is a commonly used classification algorithm, which belongs to the traditional machine learning model. It is a two-classification algorithm with good classification ability. At present, it is widely used in industry, and it has the advantages of easy implementation and relatively mature technology[33]. Therefore, this paper utilized the logistics regression model to predict financial fraud.

Results

Data processing and feature selection

We eliminated the missing rate of more than 50% of the data, because the data set selected contained a certain proportion of missing values, while other missing values were filled with multiple interpolations based on random forest. In addition, the positive and negative samples selected in this study were extremely uneven, thus, we used the oversampling method to

Table 2. AUC values of oversampling methods

Oversampling method	AUC(%)
Base	78.25
Smote	80.63
Borderline Smote	75.00
ADASYN	80.12

make up for this deficiency. Before oversampling, we randomly divided the data set into 70% training set and 30% test set, and to avoid oversampling, we set the sampling ratio to 3:10. We then compared different oversampling methods, by first establishing the GBDT classifier to calculate the AUC value on the test set, following the oversampling method, and compared it with the non-oversampled samples (this processing can be implemented in the Python tool). We selected the AUC value as the evaluation index because the proportion of positive and negative samples was not balanced. The AUC value can evaluate the model detection ability more effectively. Table 2 gives a comparison of different oversampling methods, and the Smote oversampling method had the highest AUC value of 80.63%, which was 2.38% higher than that of the unoversampled data sets.

After processing, the data dimension was still high at 94 dimensions. To avoid dimension-related problems, we used the model confidence algorithm (MCB) to draw the MUC curve and select a more stable feature selection method (including Adaptive Lasso, SCAD, MCP, Stepwise, LAD, SQRT Lasso)[28]. The MUC (Model Uncertainty Curve) curve is shown in figure 1, which is similar to the ROC curve, and the larger the area under the curve, the more stable the algorithm is and the better the effect. We observed that the Adaptive Lasso algorithm had the strongest stability, therefore, it was used to select the features of the data set[34]. Table 3 lists ten features deleted by Adaptive Lasso, including three discrete variables about dates.

The interpretability of the feature variables is very important to understand the process of machine learning. We encoded data samples with WOE, but WOE coding can only be used for discrete variables. Therefore, we first used the decision tree to discretize all the features, and then construct the features. However, not all of the features were effective. IV can measure the prediction ability and contribution ability of feature variables and can carry out important analysis of the selected features, therefore, we used IV values to filter the features. Research shows that when, this feature has good prediction ability and contribution ability, therefore, we filtered the features of $IV < 0.3$ and $IV > 0.5$. The features shown in Table 4 were screened by IV values, with a total of 16 items, which can be obtained from the table. SELL_EXP features showed the strongest ability to predict and contribute to the model.

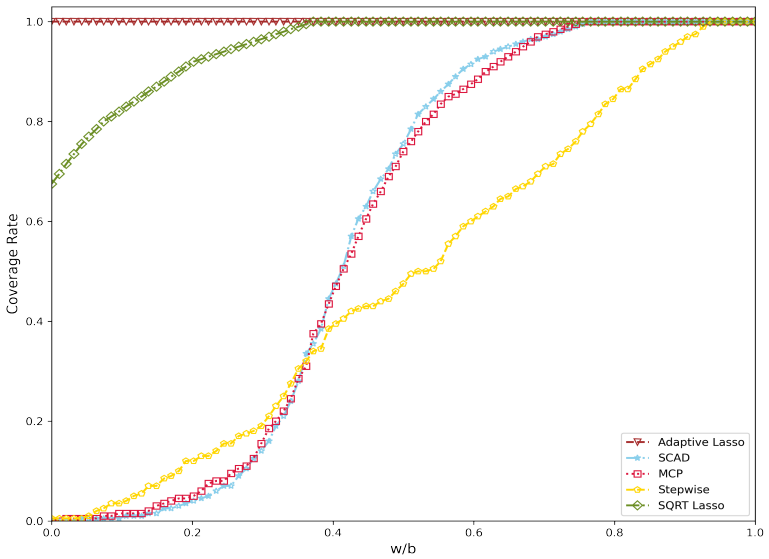

Figure 1. MUC Curves of five feature selection methods

Table 3. Features deleted by Adaptive Lasso algorithm

No.	Feature	No.	Feature
1	PUBLISH DATE	6	DILUTED EPS
2	END DATE REP	7	BASIC EPS
3	END DATE	8	OPERATE PROFIT
4	C FR OTH INVEST A	9	T COMPR INCOME
5	N CE END BAL	10	INVEST INCOME

Table 4. Features retained after WOE coding

Feature	IV	Feature	IV
SELL EXP	0.4649	T PROFIT	0.3841
C INF FR FINAN A	0.3401	PUR FIX ASSETS OTH	0.3763
C OUTF FR INVEST A	0.3953	C FR OTH OPERATE A	0.3584
C PAID DIV PROF INT	0.3145	C PAID FOR OTH OP A	0.4176
CAPITAL RESER	0.3672	OTH PAYABLE	0.3409
TAXES PAYABLE	0.3092	CIP	0.3250
INVENTORIES	0.3791	OTH RECEIV	0.3414
PREPAYMENT	0.3831	CASH C EQUIV	0.3238

Results of the model

In solving the model, the reason why the selected test set remained unchanged was that the change in the test set after data oversampling made the model select the oversampled data, leading to a higher final AUC value. The prediction results of the model are presented in Table 5. LightGBM showed the best performance, with an AUC value of 86.03% in the test set, and 64.95% KS value, indicating that it had a strong ability to distinguish between models. In addition, the AUC value was significantly higher than the unsampled AUC value of 78.25% as shown in Table2. The performance of XGBoost was also high, with an AUC value of 83.21% and a KS of 54.90%. The performance of other traditional machine learning models was average. We compared the model after WoE coding and IV value test with the model without WoE coding, and the results show that the AUC value of the encoded model in the LightGBM model increases by 1.35%, and the KS value increases by 0.05%. The results presented in Table 5 show that the two integrated models of LightGBM and XGBoost were outstanding in this study, thus, we selected the top 10 features of these two models for observation. *INT PAYABLE* was the most important factor, with an important ratio of 3.99%, as shown in Table 6. The next important factor was *NULL NUM*, and *NULL NUM*, which was listed as the most important factor by XGBoost as shown in Table 7, with an important ratio of 4.09%. Therefore, the missing values in the data set were valuable rather than randomly missing. Surprisingly, seven factors appeared in the TOP10 factor of XGboost and LightGBM. The third important factor in LightGBM was NFRV, and other TOP10 factors have been listed in Tables 6 and 7.

Interpretability of features

Tables 6 and 7 enumerate the TOP10 characteristics of the two models. To make an interpretable analysis of the characteristic variables, we used the original segmented training set and test set. Because LightGBM performed best in this study, we used dependency graphs in the Pdpbox library of Python to analyze the TOP3 factors of LightGBM and Xgboost. Among them, an increase in the positive direction value of the longitudinal axis represented an increase in the prediction probability of the positive sample, and a decrease in the negative direction value represented an increase in the prediction probability of the

Table 5. Performance evaluation

Classifier compared	AUC(%)	KS
XGBoost	83.61	0.5490
LightGBM	86.03	0.6495
LightGBM(no woe code)	84.68	0.5978
LR	71.80	0.3534
SVM	76.72	0.4973
RF	75.56	0.4245

Table 6. LightGBM model feature importance ranking (TOP10)

Feature	Importance ratio(%)	Rank
INT PAYABLE	3.99	1
NULL NUMA	2.88	2
N CF FR INVEST A	2.67	3
NOPERATE EXP	2.62	4
OTH NCA	2.57	5
C PAID OTH FINAN A	2.52	6
C PAID INVEST	2.47	7
DISP FIX ASSETS OTH	2.42	8
RETAINED EARNINGS	2.37	9
C INF FR INVEST A	2.32	10

Table 7. XGBoost model feature importance ranking (TOP10)

Feature	Importance ratio(%)	Rank
NULL NUM	4.09	1
INT PAYABLE	3.35	2
C INF FR INVEST A	2.75	3
N CF FR INVEST A	2.73	4
DISP FIX ASSETS OTH	2.62	5
NOPERATE EXP	2.52	6
RETAINED EARNINGS	2.50	7
INCOME TAX	2.29	8
PAID IN CAPITAL	2.27	9
CASH C EQUIV WOE	2.02	10

negative sample.

Fig. 2 (The increase in the positive value of the vertical axis represented an increase in the prediction probability of positive samples, while a decrease in the negative value represented an increase in the prediction probability of negative samples.) shows that the probability of data fraud in *INT PAYABLE* before 108 is relatively small, but the probability of data fraud tends to be flat when greater than 108. When *NULL NUM* reached about 240, the probability of data fraud was the highest. For the *N CF FR INVEST A* factor, when it approached 0 from a negative value, the probability of data fraud rapidly increased, and when this factor was positive, there was a certain probability of fraud than a negative value. Finally, as the value of this factor increased, the higher the probability that the sample was not fake. Among the four important factors, *INT PAYABLE* played the most important role, by increasing the probability of fraud by 4%.

Conclusion

This study investigates whether companies are faking data when data is seriously unbalanced. Data imbalance is common, and in this study, the relatively best oversampling method from Borderline Smote, ADASYN, and Smote is selected based on the AUC index. The highest AUC of Smote oversampling is 80.63%. In addition, this study also proposes a suitable feature selection method using the MUC algorithm to avoid data damage caused by the arbitrary application of the feature selection algorithm. To facilitate the identification of data fraud in this paper, we encode all the features by WOE and select the safe features based on the IV value. As shown in Table7, the coding features appear on the TOP10 ranking of the most important XGBoost features, thus, this step is very necessary.

In this study, LightGBM and XGBoost, and other machine learning classifiers are used to classify data sets. The results of their respective AUC and KS values show that the performance of LightGBM is the highest, and the effect of, Logistic Regression is the second on Xgboost. LightGBM and XGBoost are used to rank the importance of features, where *NULL NUM* ranks very high. These rules may help to improve the chances of identifying fraudulent corporate data. Besides, this study analyzes the impact of four important factors on classification, and the process has certain guiding significance for practitioners or fraud (financial) examiners of the China Securities Regulatory Commission.

Compared with earlier research, not only the stability detection of the sampling method and dimensionality reduction method are considered but also feature engineering to improve the performance of subsequent classifiers and make an interpretable

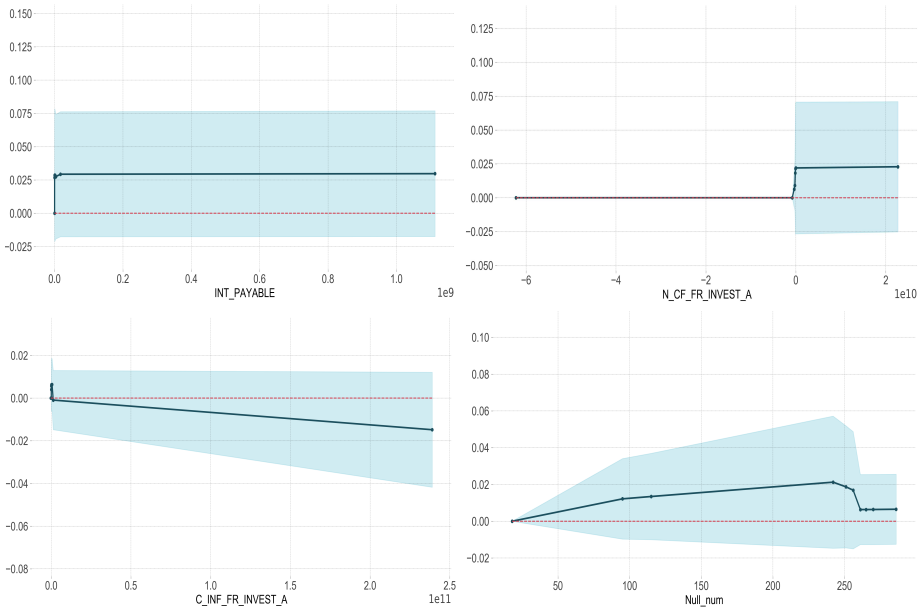

Figure 2. Partial dependence diagram of four factors

analysis of features. For follow-up research, larger and high-quality sample data for deep learning mining, and deep learning technology can be used. In addition, this data can be combined with natural language processing technology to identify financial texts, based on BERT technology for in-depth text mining analysis, and to better predict whether the listed companies have falsification in their financial statements.

Acknowledgements

The work is supported by the fund of science and technology plan project in Huizhou (No. 2020SD0402030).

Author contributions statement

Zhongzhu Liu conceived the experiment(s), Rongguang Ye, Rongye Ye conducted the experiment(s) and analysed the results. All authors reviewed the manuscript.

References

1. Sadgali, I., Sael, N. & Benabbou, F. Performance of machine learning techniques in the detection of financial frauds. *Procedia computer science* **148**, 45–54 (2019).
2. Lin, C.-C., Chiu, A.-A., Huang, S. Y. & Yen, D. C. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Syst.* **89**, 459–470 (2015).
3. Weed, D. L. Weight of evidence: a review of concept and methods. *Risk Analysis: An Int. J.* **25**, 1545–1557 (2005).
4. Kirkos, E., Spathis, C. & Manolopoulos, Y. Data mining techniques for the detection of fraudulent financial statements. *Expert. systems with applications* **32**, 995–1003 (2007).
5. Xu, W. & Liu, Y. An optimized svm model for detection of fraudulent online credit card transactions. In *2012 International Conference on Management of e-Commerce and e-Government*, 14–17 (IEEE, 2012).
6. Li, X. & Ying, S. Lib-svms detection model of regulating-profits financial statement fraud using data of chinese listed companies. In *2010 International Conference on E-Product E-Service and E-Entertainment*, 1–4 (IEEE, 2010).
7. Liang, J. & Lv, W. Research on detecting technique of financial statement fraud based on fuzzy genetic algorithms bpn. In *2009 International Conference on Management Science and Engineering*, 1462–1468 (IEEE, 2009).
8. Huang, S.-M., Yen, D. C., Yang, L.-W. & Hua, J.-S. An investigation of zipf's law for fraud detection (dss# 06-10-1826r (2)). *Decis. Support. Syst.* **46**, 70–83 (2008).

9. Rizki, A. A., Surjandari, I. & Wayasti, R. A. Data mining application to detect financial fraud in indonesia's public companies. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*, 206–211 (IEEE, 2017).

10. Bhusari, V. & Patil, S. Study of hidden markov model in credit card fraudulent detection. In *2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*, 1–4 (IEEE, 2016).

11. Calderon, T. G. & Cheh, J. J. A roadmap for future neural networks research in auditing and risk assessment. *Int. J. Account. Inf. Syst.* **3**, 203–236 (2002).

12. Bauder, R., da Rosa, R. & Khoshgoftaar, T. Identifying medicare provider fraud with unsupervised machine learning. In *2018 IEEE international conference on information Reuse and integration (IRI)*, 285–292 (IEEE, 2018).

13. Huang, S.-Y., Tsaih, R.-H. & Yu, F. Topological pattern discovery and feature extraction for fraudulent financial reporting. *Expert. systems with applications* **41**, 4360–4372 (2014).

14. Deng, Q. & Mei, G. Combining self-organizing map and k-means clustering for detecting fraudulent financial statements. In *2009 IEEE International Conference on Granular Computing*, 126–131 (IEEE, 2009).

15. Behera, T. K. & Panigrahi, S. Credit card fraud detection: a hybrid approach using fuzzy clustering & neural network. In *2015 Second International Conference on Advances in Computing and Communication Engineering*, 494–499 (IEEE, 2015).

16. Feroz, E. H., Kwon, T. M., Pastena, V. S. & Park, K. The efficacy of red flags in predicting the sec's targets: an artificial neural networks approach. *Intell. Syst. Accounting, Finance & Manag.* **9**, 145–157 (2000).

17. Etheridge, H. L. & Brooks, R. C. Neural networks: A new technology. *The CPA J.* **64**, 36 (1994).

18. Ramamoorti, S., Bailey Jr, A. D. & Traver, R. O. Risk assessment in internal auditing: a neural network approach. *Intell. Syst. Accounting, Finance & Manag.* **8**, 159–180 (1999).

19. Deng, Q. Detection of fraudulent financial statements based on naïve bayes classifier. In *2010 5th International Conference on Computer Science & Education*, 1032–1035 (IEEE, 2010).

20. Busta, B. & Weinberg, R. Using benford's law and neural networks as a review procedure. *Manag. Auditing J.* (1998).

21. Yan, C. & Li, Y. The identification algorithm and model construction of automobile insurance fraud based on data mining. In *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, 1922–1928 (IEEE, 2015).

22. Benchaji, I., Douzi, S. & El Ouahidi, B. Using genetic algorithm to improve classification of imbalanced datasets for credit card fraud detection. In *International Conference on Advanced Information Technology, Services and Systems*, 220–229 (Springer, 2018).

23. Hajek, P. & Henriques, R. Mining corporate annual reports for intelligent detection of financial statement fraud–a comparative study of machine learning methods. *Knowledge-Based Syst.* **128**, 139–152 (2017).

24. Zhou, H. *et al.* A distributed approach of big data mining for financial fraud detection in a supply chain. *CMC-COMPUTERS MATERIALS & CONTINUA* **64**, 1091–1105 (2020).

25. Stekhoven, D. J. & Bühlmann, P. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).

26. Subudhi, S. & Panigrahi, S. Effect of class imbalanceness in detecting automobile insurance fraud. In *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, 528–531 (IEEE, 2018).

27. Hancock, J. T. & Khoshgoftaar, T. M. Gradient boosted decision tree algorithms for medicare fraud detection. *SN Comput. Sci.* **2**, 1–12 (2021).

28. Li, Y., Luo, Y., Ferrari, D., Hu, X. & Qin, Y. Model confidence bounds for variable selection. *Biometrics* **75**, 392–403 (2019).

29. 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*, 785–794 (2016).

30. Zhang, S., Hu, Y. & Tan, Z. Research on borrower's credit classification of p2p network loan based on lightgbm algorithm. *Int. J. Embed. Syst.* **11**, 602–612 (2019).

31. Rochman, E. M. S. *et al.* Classification of thesis topics based on informatics science using svm. In *IOP Conference Series: Materials Science and Engineering*, vol. 1125, 012033 (IOP Publishing, 2021).

32. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).

33. Menard, S. *Applied logistic regression analysis*, vol. 106 (Sage, 2002).

34. Zou, H. & Hastie, T. Addendum: regularization and variable selection via the elastic net. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **67**, 768–768 (2005).