Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2021.020104
ISSN 2790-0983 Vol. 2, Issue 1: 26-37

Machine Learning-Based Study on the Identification of Misstatements in Annual Reports of Listed Companies - A Financial Restatement Prediction Perspective

Ga dan Cai Rang *

Qinghai Normal University, Qinghai, China

2627203173@qq.com

corresponding author

Keywords: ML, Listed Companies, Annual Report Misstatement Identification, Financial Restatement Prediction

Abstract: The significance and importance of financial reporting as the basis and foundation for relevant decision makers to make judgments cannot be overstated, and both capital market participants and regulators attach great importance to the disclosure of financial reports. Although relevant regulatory authorities such as China's Securities Regulatory Commission and Accounting Standards Board have issued corresponding standards and regulations to strictly regulate the disclosure behaviour of companies' financial reports, due to insufficient supervision or relatively low costs of non-compliance, listed companies have committed FF in violation of relevant laws and regulations in order to preserve their own interests, causing unmeasurable losses to stakeholders. The main objective of this paper is to launch a study on the identification of misstatements in listed companies' annual reports based on ML under the perspective of financial restatement prediction. Based on the research on FF patterns and FFI in the first part, we firstly select the set of primary features based on the theory of FF motive, then perform Mann-Whitney test on the set of primary features to obtain the set of original features, and then use Bortua algorithm to select the final set of FI features from the set of original features. The data of the fraud samples and non-fraud samples were loaded into the four types of financial fraud identification models (FFIM) in the order of the original FI features and the FI features constructed by the BA. The model identification results showed that the combination of the FFI features constructed by the BA and the RFM had a good identification effect, and the overall evaluation indexes of G-mean and F-value were 75.9% and 78.3% respectively.

1. Introduction

Due to the difference in the level of accounting information technology between different

Copyright: © 2021 by the authors. This is an Open Access article distributed under the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (https://creativecommons.org/licenses/by/4.0/).

enterprises and the lack of sufficient research on data-based auditing, data-based auditing is only used in large enterprises with sound accounting information technology in audit practice, and often in internal audits. In external audits, only some large accounting firms will adopt it, but the scope of use is also very limited. Taking FF audits as an example, auditors tend to use data-based audits in the identification and assessment of FF risks, with the help of data mining techniques such as classification, clustering and correlation analysis to find the risk points of corporate fraud [1-2].

In a related study, Thomas et al. examined the impact of audit committee linkages through a network of directors on the quality of financial reporting, particularly misstatements in annual financial statements [3]. Using network analysis, multiple dimensions of connectedness were examined and it was found that companies with well-connected audit committees were less likely to misstate their annual financial statements after controlling for operating performance and corporate governance characteristics. In addition, it was shown that audit committee linkages through the network of directors moderated the negative impact of the linkages between the board of directors and the misstated company on the quality of financial reporting.Recep et al. summarised the fundamentals of ASR assessment and then proposed a new approach based on alternative hypothesis recommendations for detecting and correcting these errors generated by automated speech recognition (ASR) systems [4]. The proposed method consists of a series of processes such as identifying incorrect words, selecting words that can be corrected, and identifying candidate words to replace these words. Tests carried out by creating different test environments resulted in a significant improvement in performance in Turkey, with an average performance increase of 4.60%.

This paper begins with an introduction to the theory related to FF patterns, FFI and data mining techniques. FF pattern is a qualitative study on the definition, motivation and characteristics of FF; FFI is a study on the identification of corporate FF based on theories such as statistics, ML and data mining; data mining is the process of discovering useful patterns and trends from large data sets, which has been applied to FFI by a large number of researchers and scholars with good results. Secondly, the empirical data required for the construction of the FFIM are prepared, including the collection of a sample of fraudulent firms, a sample of non-fraudulent firms and the construction of a FI feature set. Finally, a FI model for financial statements of listed companies is constructed based on ML related algorithms.

2. Design Research

2.1. Error Reporting Identification Design Solution

The information mining of data sets is done by writing programs that are run internally by software, and the whole process is like a black box, where sample data is input into the black box and the results are output after the black box is mined [5-6]. But what kind of samples and data sets are input into the "black box" is as important as how the "black box" is constructed, and it can even be said that the data preparation work before data mining accounts for more than two-thirds of the whole data mining process. Take FFI as an example, the whole data mining process needs to go through six stages: fraud problem understanding, sample data understanding, data preparation, FI model construction, FI model performance evaluation, and fraud audit program deployment [7-8], as shown in Figure 1.

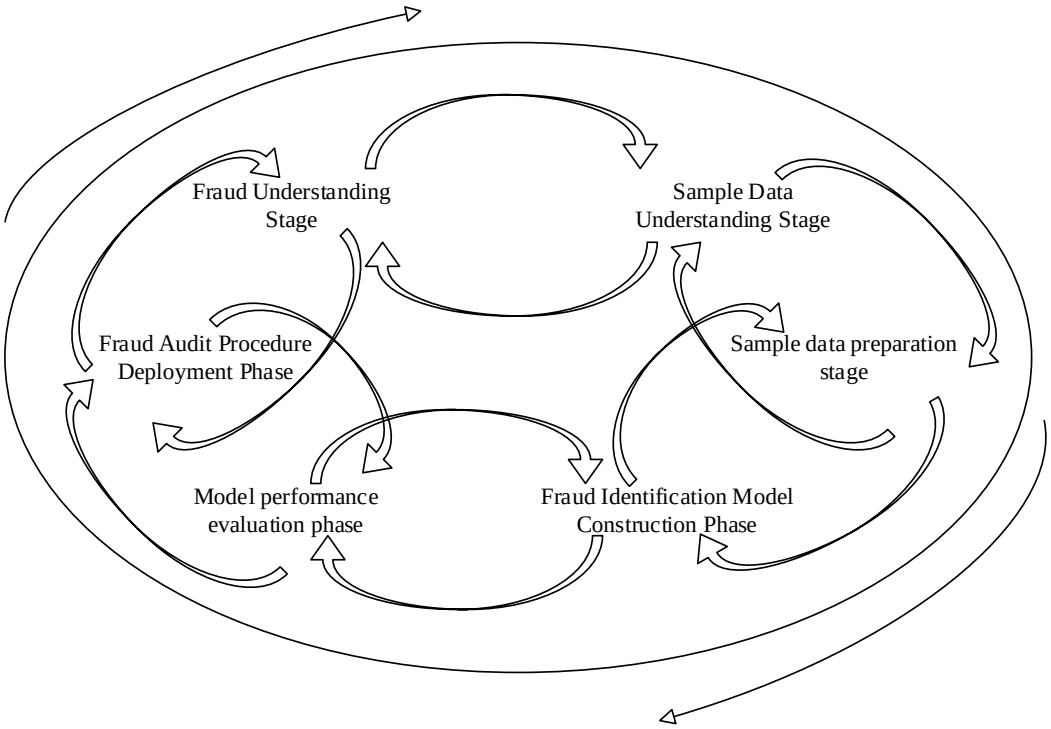

Figure 1. Design scheme for the application of data mining in error reporting identification

The sequence of these six phases is not simply front-to-back and unrelated to each other; they are adaptive to each other because the task segment of the latter phase is usually dependent on the results of the former phase, and whether the results of the former phase can meet the task requirements of the latter phase is subject to continuous trial and error between the two [9-10]. In addition, the lifecycle of data mining is also an iterative and escalating process, as lessons are learnt from each completed data mining exercise and used as input for a new project in the next data mining process. The following paper will briefly describe the various stages of the data mining process for FFI [11-12].

1) Fraud understanding stage

There are three steps in the fraud problem understanding phase. Firstly, the objectives and constraints of the business or study are identified based on the business or study; secondly, these objectives and constraints are translated into a formula for defining the data mining problem; and finally, an initial strategy for achieving the objective is identified. To achieve this objective, the auditor must reasonably assess the risk of fraud by the audited entity, which, in terms of the outcome of the problem, can be solved using the dichotomous classification algorithm in data mining.

2) Sample data understanding stage

In the sample data understanding phase, auditors need to collect data, familiarise themselves with the data, evaluate the data, determine the executable data, and ultimately determine the initial fraud-identifying feature categories by understanding the financial characteristics, internal governance characteristics, manager characteristics, and operating environment characteristics of the enterprise in order to construct the initial fraud feature set.

3) Sample data preparation stage

The sample data preparation phase requires a significant investment of effort and entails pre-processing the data within the initial set of fraud-specific features identified in the sample understanding phase, including cleaning up the raw data, dealing with missing data, FS and other aspects of preparing the final data set. Fraud FS is based on the initial fraud feature set. The types of features relevant to FI are clarified during the construction of the initial FI feature set, which sets the direction for data collection. As much data as possible related to the types of FI features are collected during the data collection process, but no consideration is given to how well these data will be identified in FI. In the fraud FS phase, the best FI features are determined based on the principles of FS and the FS algorithm, and the best FI features should be as simple and easy to understand as possible while maintaining good identification performance.

(4) FI model construction stage

Data mining techniques such as classification, regression, correlation analysis and clustering can be chosen for the construction of FFIMs. In terms of the nature of the FI problem and the main research directions, classification is the most commonly used data mining method in the field of FI, while some scholars are also engaged in the research of association rule mining, but the application of classification in FI is more difficult than the latter. Similarly, this paper investigates the identification of corporate FF based on classification algorithms. The reason for choosing classification algorithms is that FFI is essentially a dichotomous problem, and classification algorithms have a sound theoretical system for dealing with dichotomous problems, and also have good performance in classification.

5) Model performance evaluation stage

After the FF modelling phase is completed, one or more FFIMs will be generated. Before applying these models to the actual problem, the effectiveness of the FI models must be evaluated and it must be determined whether the models are able to achieve the objectives set out in the fraud problem understanding phase.

6) Fraud audit process deployment phase

The use of data mining techniques to identify FF is useful to auditors, and the FFIM constructed in this paper based on data mining techniques is largely complementary to the audit procedures performed by auditors in identifying and assessing the risk of fraud in the audited entity.

In summary, this paper is based on the study of corporate FFI, firstly, it introduces the definition of FF and FF audit, defines the two major categories and four forms of FF and the definition and objectives of FF audit, based on which it introduces the three mainstream theories of FF motivation; then it introduces the data mining techniques used in this paper to identify corporate FF. Finally, a framework for the application of data mining techniques in FFI is designed in conjunction with the data mining process.

2.2. ML

ML is broadly defined as "the act of using experience to improve the performance of a system by means of computation". In everyday practice there is a vast amount of data that contains information that can be learnt by computers, and ML is a 'learning algorithm' that learns the internal logic from this vast amount of data and generates models based on this internal logic. In practical terms, ML is a method of selecting appropriate algorithms to guide a computer to learn from a large amount of data, uncovering the logical information present in it, training a model based on it, and then using the model to make predictions about the problem under study [13-14].

There are different classifications of ML algorithms according to different ways of categorisation, and the algorithms covered in this paper are classified according to the type of data, and are divided into three main categories: the first category, supervised learning, refers to the category of data to be learnt that is known at the time of training, and the main algorithms in this category are random forest (LR) and integration methods. The second category, unsupervised learning, is the opposite, targeting datasets of unknown category and uncovering the intrinsic logical connections through learning training, and consists broadly of clustering algorithms, principal component analysis, etc. The third category, reinforcement learning, is somewhere in between, where an intelligent being "tries out" different actions based on the choices it makes in the environment and thus receives different rewards to guide its next behaviour. The aim of this learning is to maximise the cumulative rewards available to the intelligence [15-16].

ML algorithms can learn from high-dimensional, large amounts of non-linear data and explore the laws from these 'experiences' to generate models, overcoming the potential corruption of traditional linear regression methods when dealing with high-dimensional data by downscaling, and are flexible enough to solve the problems of traditional statistical methods in dealing with unstructured information. At the same time, the ML algorithm is not limited by the number of independent variables, and will not affect the model results because of the excessive number of variables [17-18].

2.3. Text Feature Acquisition Methods

FS is a very important step in the process of pattern recognition by ML. The pre-processed text has the problems of high dimensionality, redundant features and more features affecting the classification accuracy. Therefore, FS of the text to be classified can effectively reduce the text dimensionality and improve the classification speed and accuracy. The details are as follows.

1) Information Gain

The FS method based on Information Gain is used to calculate how much information a feature item can provide in the classification process, and to determine the importance of the feature item in the classification process. The formula is as follows.

$$IG(t) = -\sum_{i=1}^{m} P(C_i)\log(C_i) + P(t)\sum_{i=1}^{m} P(C_i \mid t)\log P(C_i \mid t) + P(\bar{t})\sum_{i=1}^{m} P(C_i \mid \bar{t})\log P(C_i \mid \bar{t}) \tag{1}$$

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{2}$$

The sum of the information entropy before and after the feature item in the Information Gain reflects the interaction between the classification prediction and the feature item. When the information gain of a word is greater, the greater the classification ability of the word in the classification process.

2) Mutual Information

The Mutual Information method is derived from information theory and can be used to measure the correlation between feature items and classification items. In text classification, mutual information between text and words is defined by the degree of co-occurrence of text and words, which is defined by the following formula.

$$MI(t,c) = \log \frac{P(t \mid c)}{P(t)} = \log \frac{P(t \mid c)}{P(t) \times P(c)} \tag{3}$$

In the mutual information FS method, the greater the mutual information between the word and the text, the greater the role of the word in the text classification.

(3) Cardinality test

The chi-square test is a method used to calculate the degree of independence between feature terms and categories. The chi-square test method is derived from the hypothesis test method in probability theory, in which it is first assumed that the feature items and category items are independent of each other, and then the actual value between the feature items and category items is calculated, and category items are considered to be related, and the calculation formula is as follows.

$$MI(t,c) = \log \frac{P(t \mid c)}{P(t)} = \log \frac{P(t \mid c)}{P(t) \times P(c)} \tag{4}$$

A, B, C and D in the formula indicate the four cases between feature item t and category item c, which are the number of texts of category item t containing feature item t, the number of texts of category item t not containing feature item t, the number of texts of non-category item c containing feature item t, and the number of texts of non-category item c not containing feature item t.

3. Experimental Study

3.1. FFI Features Selection

1) FS method

The following will introduce FS according to the research process of FS.

2) The selection process of FS algorithms

As shown in Figure 2, Feature selection (FS) generally goes through three processes: firstly, a subset of features to be judged is generated by subset search; secondly, the superiority of the generated subset to be judged is evaluated by a selected evaluation function; finally, a threshold is set for the evaluation function, and the search can be stopped when the value of the evaluation function reaches this threshold, and the optimal subset of features is output.

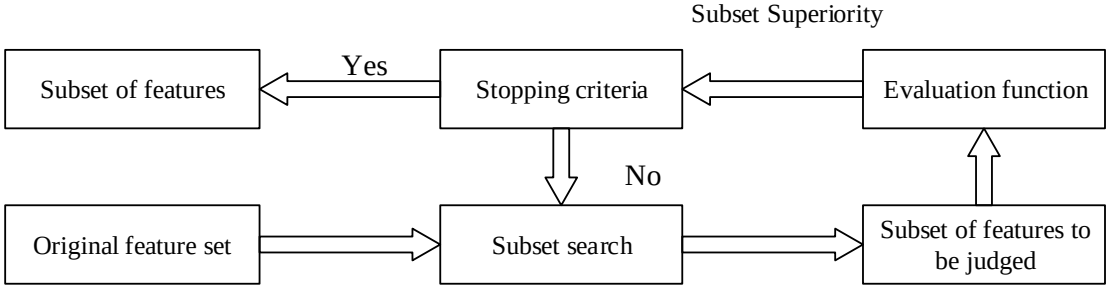

Figure 2. Basic framework for FS

3) Subset search

(1) Complete search (Complete)

It is feasible when calculation will no longer be feasible. In order to apply the complete search method in the original feature set with a large number of features, the complete search method

started to abandon the method of traversing all possible feature subsets in the initial feature set and reduce the number of exhaustive enumeration by introducing branching bounds, priority queues, etc. Therefore, the complete search formed two types of search patterns, an exhaustive search and a non-exhaustive search.

(2) Heuristic search (Heuristic)

Heuristic search is divided into algorithms such as forward search, backward search, two-way search and instance-based search.

(3) Randomised algorithms

Heuristic forward and backward search algorithms tend to fall into the trap of local optimisation when searching for the optimal CS and tend to compensate for the inability to escape from local optimisation.

4) Subset evaluation

Subset evaluation is the second key aspect of FS, in which the CS superiority is evaluated according to the set evaluation function, and the subset search process is stopped or not according to the superiority of the candidate subset (CS). The subset evaluation process is essentially an evaluation of the difference between the current CS and the true classification of the training data set. When the difference between the division of the training data set by the newly generated CS and the true division of the training data set reaches a minimum, the generation of the CS is stopped.

3.2. Boruta Algorithm (BA)

The BA consists of the following steps.

(a) In the first step, shaded features are created by replicating all true features in the dataset (the original dataset is extended by at least 5 shaded features.

In the second step, the newly added shaded features are randomized to eliminate correlations between them.

In the third step, train a RF classifier on the extended dataset and collect the computed z-scores.

In the fourth step, the maximum Z-score value (MZSA) of the shaded features is found, and then a hit is recorded for each true feature with a Z-score score higher than MZSA, indicating that this true feature is more important than the shaded features.

In the fifth step, the same bilateral test as MZSA is performed for the true features.

In the sixth step, attributes that are significantly more important than the shaded features are considered important.

In the seventh step, remove all shaded attributes.

In step eight, the process is repeated until importance has been assigned to all attributes or the algorithm has reached the limits set previously for the RF run.

At the end of training, the BA can also output a ranking of the original features, indicating the importance level of the feature, which is also a useful metric in FS.

4. Experiment Analysis

4.1. Analysis of the Model Recognition Effect of the BA Screened Feature Set Samples

The initial set of features screened by the BA was used as the set of features for FR because the original set of features for FR had a large number of dimensions, the average results were used to represent the overall results of the model (see Table 1).

Table 1. Average results of the test sample model runs

Classification models	Specificity	Sensitivity	Accuracy	G	F
DTs	66.7%	67.0%	66.9%	66.4%	66.0%
LR	66.1%	76.6%	71.3%	70.8%	71.6%
RF	72.0%	77.8%	74.9%	74.5%	74.4%
Support vector machines	70.3%	73.8%	72.0%	71.8%	71.4%

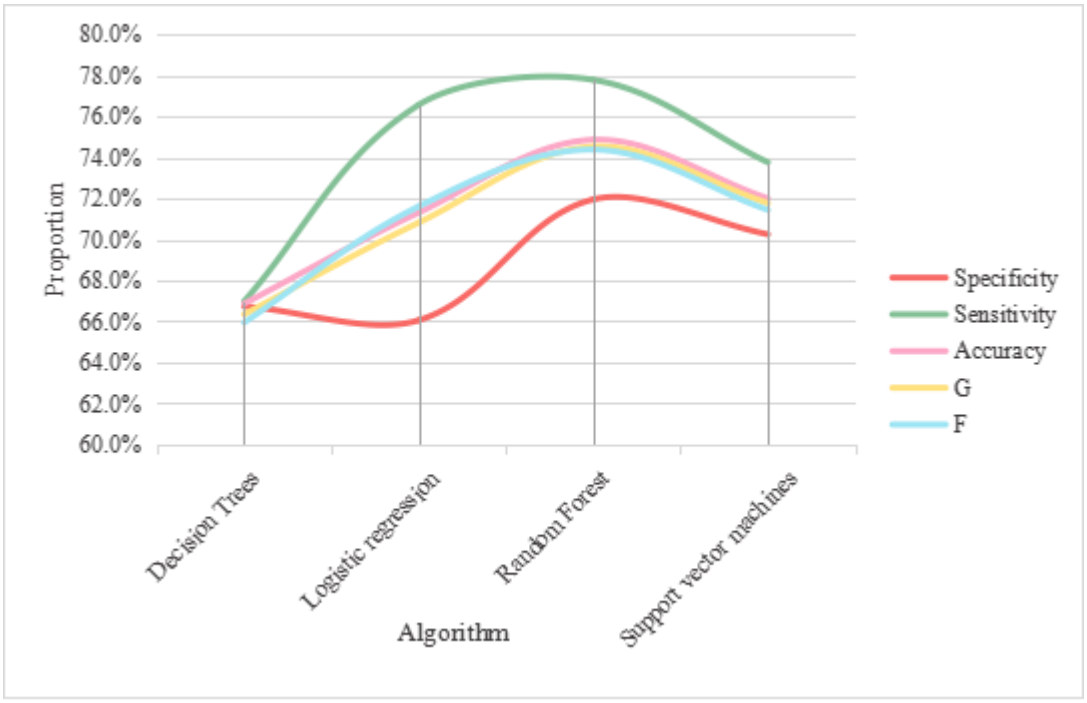

Figure 3. Analysis of the average run results of the test sample models

The results of the above four FI models in Figure 3 show that the set of features screened by the BA has better identification results in the random forest model (RFM), with F-value reaching 74.53% and 74.4%. The BA reduced original set to 18, which reduced the dimensionality of the original set of FI features, however, the overall identification results of the BA filtered feature set were not as effective as the original set of FI features. Therefore, the BA is unable to reduce the dimensionality of the FR features while maintaining the efficiency of the fraud recognition (FR) model.

Table 2. Test sample confusion matrix

Classification models	TP	FN	FP	TN
DTs	32	16	16	32
LR	37	11	16	32
RF	37	11	13	35
Support vector machines	35	13	14	34

From the confusion matrix of the tested samples, the number of correctly identified fraudulent statements is 32 and 16 incorrectly identified fraudulent statements, the average sensitivity of the model is about 66.74%; the number of correctly identified non-fraudulent statements is 32 and 16

incorrectly identified non-fraudulent statements, the average specificity of the model is about 66.99%; the average accuracy of the model is about 66.99%. The average specificity of the model is about 66.99%; the average accuracy of the model (accuracy) is 66.87%, and the values of the overall evaluation index G mean and F value are 66.35% and 65.95% respectively.

4.2. Analysis of ML Effects

In order to investigate whether the addition of social structure signals and social sentiment signals from the SSE e-interactive and interactive platforms can further improve the identification of FFs, this paper first conducted ML on the financial ratio signals from annual reports, non-financial ratio signals and social structure signals from the stock bar of Orient Fortune, as a basis for comparison of the subsequent ML results under the complete signal system. The results of the four classifiers are shown in Table 3,

Table 3. Effectiveness of social structure signals for FFI

	SVM	CART	RF	Adaboost
Accuracy	60.3%	70.7%	70.7%	70.7%
Recall	55.6%	77.8%	74.1%	81.5%
F1-score	56.6%	71.2%	70.2%	72.1%
AUC	60.0%	71.2%	76.5%	71.4%

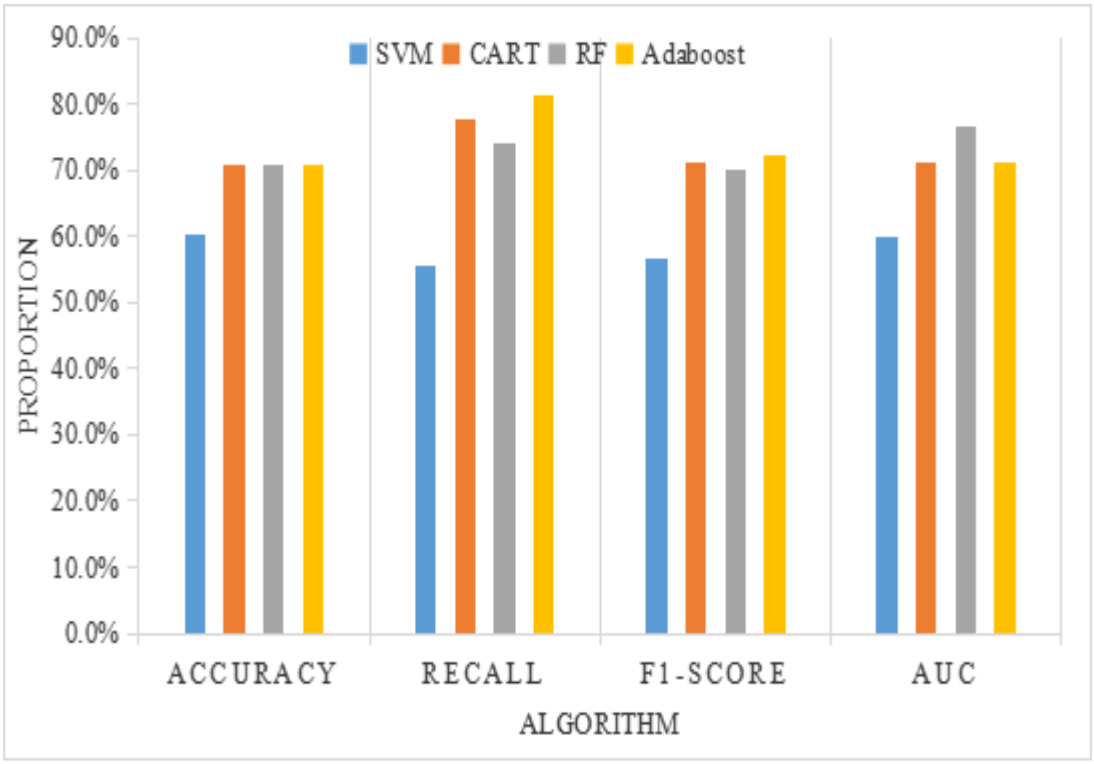

Figure 4. Analysis of the effectiveness of social structure signals for FFI

Further, the social structure signals and social sentiment signals from SSE e-interactive platform and interactive platform were combined with the financial ratio signals, non-financial ratio signals and social structure signals from the share bar of Oriental Fortune website to form the whole signal

system, and the same classifier was used to learn the final results as shown in Table 4.

Table 4. Effectiveness of FFI under the complete signalling system 96 fraudulent firms & 96 non-fraudulent firms

	Accuracy	Recall	F1-score	AUC
SVM	68.97%	66.67%	66.67%	0.6880
CART	70.69%	81.48%	72.13%	0.7139
RF	77.59%	77.78%	76.36%	0.8154
Adaboost	75.86%	81.48%	75.86%	0.7622

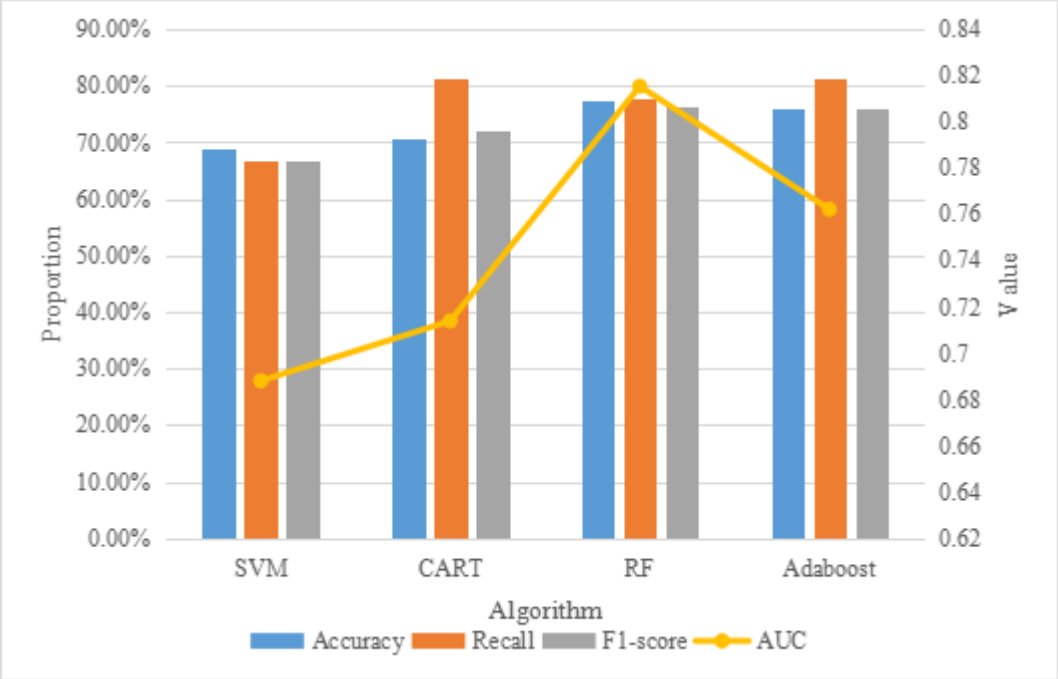

Figure 5. Analysis of the effectiveness of FFI with a complete signalling system

Comparing Figure 4 with Figure 5, it can be seen that the majority of the classification performance metrics of the four ML algorithms are further improved after adding the data from the SSE eInteractive and Interactive Platform. Thus, the effectiveness and superiority of the complete FFI signalling system constructed in this paper for the identification of FF of listed companies in China is verified.

5. Conclusion

In the context of trade globalization and the era of big data, the organization and transaction forms of enterprises have become more and more complex and varied, and the business and financial data involved in the operation activities of large enterprises have become more and more numerous, and information-based auditing and big data auditing have become more and more concerned by scholars engaged in audit theory research as well as audit practitioners. Data mining has gradually come into the view of audit theoretical research scholars and audit practitioners. Research on the application of data mining technology in FF auditing can enrich the research on the application of FF auditing theory in the era of big data, and explore the application of data mining

technology in auditing in greater depth. At the same time, it also helps to narrow the gap between big data audit research and audit practice. In China, research on data mining technology in the field of auditing has just taken off, but most of the research is rather scattered and has not yet formed a complete system. Therefore, an in-depth study of data mining technology in today's big data audit environment is of great significance in improving audit theories and methods and establishing a complete theoretical system of big data auditing.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Jay Henderson, Tanya R. Jonker, Edward Lank, Daniel Wigdor, Ben Lafreniere: Investigating Cross-Modal Approaches for Evaluating Error Acceptability of a Recognition-Based Input Technique. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6(1): 15:1-15:24 (2020).Candace E. Peacock, Ben Lafreniere, Ting Zhang, Stephanie Santosa, Hrvoje Benko, Tanya R. Jonker: Gaze as an Indicator of Input Recognition Errors. Proc. ACM Hum. Comput. Interact. 6(ETRA): 1-18 (2020). https://doi.org/10.1145/3517262

[2] Thomas C. Omer, Marjorie K. Shelley, Frances M. Tice: Do Director Networks Matter for Financial Reporting Quality? Evidence from Audit Committee Connectedness and Restatements. Manag. Sci. 66(8): 3361-3388 (2020). https://doi.org/10.1287/mnsc.2019.3331

[3] Recep Sinan Arslan, Necaattin Barisģi, Nursal Arici, Sabri Koçer: Detecting and correcting automatic speech recognition errors with a new model. Turkish J. Electr. Eng. Comput. Sci. 29(5): 2298-2311 (2021). https://doi.org/10.3906/elk-2010-117

[4] Rebecca Brooke Bays, Mary Ann Foley, Annelise Cohen: Is it all in the details? Description content and false recognition errors. Cogn. Process. 21(2): 185-196 (2020). https://doi.org/10.1007/s10339-019-00945-8

[5] A. B. Dhivya, M. Sundaresan: Tablet identification using support vector machine based text recognition and error correction by enhanced n-grams algorithm. IET Image Process. 14(7): 1366-1372 (2020). https://doi.org/10.1049/iet-ipr.2019.0993

[6] Farhad Abedini, Mohammad Reza Keyvanpour, Mohammad Bagher Menhaj: Correction Tower: A General Embedding Method of the Error Recognition for the Knowledge Graph Correction. Int. J. Pattern Recognit. Artif. Intell. 34(10): 2059034:1-2059034:38 (2020). https://doi.org/10.1142/S021800142059034X

[7] Nacereddine Hammami, Isah Abdullahi Lawal, Mouldi Bedda, Nadir Farah: Recognition of Arabic speech sound error in children. Int. J. Speech Technol. 23(3): 705-711 (2020). https://doi.org/10.1007/s10772-020-09746-3

[8] Roghayeh Mojarad, Ferhat Attal, Abdelghani Chibani, Yacine Amirat: Automatic Classification Error Detection and Correction for Robust Human Activity Recognition. IEEE Robotics Autom. Lett. 5(2): 2208-2215 (2020). https://doi.org/10.1109/LRA.2020.2970667

[9] Se-In Jang, Geok-Choo Tan, Kar-Ann Toh, Andrew Beng Jin Teoh: Online Heterogeneous Face Recognition Based on Total-Error-Rate Minimization. IEEE Trans. Syst. Man Cybern. Syst. 50(4): 1286-1299 (2020). https://doi.org/10.1109/TSMC.2017.2724761

[10] Leandro Miranda, José Viterbo, Flávia Bernardini: A survey on the use of ML methods in context-aware middlewares for human activity recognition. Artif. Intell. Rev. 55(4): 3369-3400 (2020). https://doi.org/10.1007/s10462-021-10094-0

[11] Esma Uzunhisarcikli, Erhan Kavuncuoglu, Ahmet Turan Özdemir: Investigating classification performance of hybrid deep learning and ML architectures on activity recognition. Comput. Intell. 38(4): 1402-1449 (2020). https://doi.org/10.1111/coin.12517

[12] Syed Saqib Raza Rizvi, Muhammad Adnan Khan, Sagheer Abbas, Muhammad AsadUllah, Nida Anwer, Areej Fatima: Deep Extreme Learning Machine-Based Optical Character Recognition System for Nastalique Urdu-Like Script Languages. Comput. J. 65(2): 331-344 (2020). https://doi.org/10.1093/comjnl/bxaa042

[13] Miroslav Stampar, Kresimir Fertalj: Applied ML in recognition of DGA domain names. Comput. Sci. Inf. Syst. 19(1): 205-227 (2020). https://doi.org/10.2298/CSIS210104046S

[14] Saswati Debnath, Pinki Roy: Audio-visual speech recognition based on ML approach. Int. J. Adv. Intell. Paradigms 21(3/4): 211-224 (2020).

[15] Leila Boussaad, Aldjia Boucetta: Extreme Learning Machine-Based Age-Invariant Face Recognition with Deep Convolutional Descriptors. Int. J. Appl. Metaheuristic Comput. 13(1): 1-18 (2020). https://doi.org/10.4018/IJAMC.290540

[16] R. Vinodini, M. Karnan: Face detection and recognition system based on hybrid statistical, ML and nature-based computing. Int. J. Biom. 14(1): 3-19 (2020).

[17] Soumia Faouci, Djamel Gaceb, Mohammed Haddad: Offline Arabic handwritten character recognition: from conventional ML system to deep learning approaches. Int. J. Comput. Sci. Eng. 25(4): 385-398 (2020).

[18] Maria Chiara Fastame: Are subjective cognitive complaints associated with executive functions and mental health of older adults? Cogn. Process. 23(3): 503-512 (2020).