Dean&Francis

ISSN 2959-6130

Identifying Key Factors in Financial Statement Fraud Detection in China's Stock Market through Logistic Regression and SVM

Weilun Chen

School of Business Administration,
Southwestern University of Finance
and Economics, Chengdu, Sichuan,
611130, China
42237010@smail.swufe.edu.cn

Abstract:

In the complex capital market, since false disclosures can undermine investors' confidence, the detection of financial statement fraud becomes extremely important. False disclosures can weaken investors' confidence and affect the operation of the market. Research on the detection of financial fraud in China's A-share market is relatively weak, especially in terms of regulatory enforcement and market behavior response. Therefore, this study aims to utilize machine learning to investigate the fraudulent behaviors in the financial reports of Chinese listed companies. This paper uses logistic regression (LR) and support vector machine (SVM) to classify and distinguish financial statements. The data comes from CSMAR and Wind databases, while the penalty records are from the China Securities Regulatory Commission. The final research results show that LR outperforms SVM in terms of accuracy and interpretability. The accuracy rate is 96. 8% and the F1 value is 80%. The market reaction analysis further reveals that after the penalty announcement is released, there will be a short-term upward trend in the company's stock price, but a long-term poor performance. This actually reflects the excessive reaction of investors and that it will take some time for the stock price to correct in the later stage. The results of this study provide empirical evidence for future applications of artificial intelligence in detecting stock market fraud and also contribute to the understanding of the Chinese capital market. They offer some references for regulatory authorities, investors, and enterprises.

Keywords: Financial fraud; Machine learning; Logistic regression; Support vector machine; Market reaction

ISSN 2959-6130

1. Introduction

Due to the explosive growth of Chinese listed companies in recent years, the detection of financial statement fraud has become particularly important and extremely challenging [1, 2]. In China's A-share market, accounting manipulation and information disclosure violations by companies are very common. It is well known that these behaviors will undermine investors' confidence and reduce the efficiency of capital allocation [3, 4]. This project aims to identify fraudulent companies based on the financial indicators of each company using machine learning, to study the market reactions triggered by law enforcement incidents [5-7].

2. Data Description

2.1 Data Source

Most of my data is derived from the CSMAR and Wind databases. The dependent variable 'fraud' is manually labeled based on CSRC announcements of administrative penalties for false disclosure [8]. The indicators and variable selection follow prior studies on earnings manipulation and fraud detection [9, 10]. The main independent variables include profitability indicators (return on assets, return on equity, gross margin), Leverage and liquidity (debt ratio, current ratio, quick ratio), Growth (revenue growth, asset growth), Governance (board size, independence, dual tenure of the CEO, type of auditor) Market (trade volume, volatility, stock returns).

2.2 Data Cleaning and Preprocessing

Due to the extensive data collection, missing values and coding errors are common problems. Firstly, empty or irrelevant columns were removed, and missing values were handled by deleting rows that contained missing target variables. At the same time, the median was used to fill in the numerical gaps. Non-numeric fields were converted to numeric types through forced conversion, and the "GBK" decoding errors caused by Chinese column names were corrected. Secondly, in this study, the "Wason Normalization" method was employed to reduce the influence of extreme values, and the outliers were corrected at the 1% and 99% levels. Finally, all numerical variables were standardized through z-score normalization. Since fraud cases accounted for less than 5% of all samples, "RandomOverSampler" was applied to balance the dataset.

2.3 Train-Test Split

To make the model more reliable, the researchers experimented with different data division ratios, including 70% training and 30% testing, 80% and 20%, and 60% and 40%. When dividing the data, ensure that the ratio of positive and negative samples in each subset (training set and testing set) is the same. This is done to ensure the fairness of model evaluation.

In addition, data augmentation or resampling operations are only performed on the training set, and the testing set remains unchanged to prevent "cheating" and to allow the test results to truly reflect the model's performance in reality.

To avoid multicollinearity, the study excluded highly correlated variables (with a correlation coefficient |r| > 0.9). Finally, the study retained 25 core variables, which covered financial performance, leverage ratio, governance structure, and growth status.

To further clarify which features were more important, the study used the recursive elimination method and combined it with the coefficients of logistic regression for analysis.

2.4 Summary of Data Statistics

Table 1. Statistics for Financial Variables

Variable	Mean	Std	Min	Max
ROA	0. 057	0. 064	-0. 412	0. 391
Debt Ratio	0. 46	0. 18	0. 05	0. 89
Current Ratio	1. 89	1. 73	0. 12	14. 2
Revenue Growth	0. 13	0. 27	-0. 76	2. 04

After research, it was found that fraudulent cases typically exhibit lower return on assets (ROA) and return on equity (ROE), along with a higher debt ratio. These fraudulent companies may also change their auditing firms and conduct frequent related-party transactions. This not only reflects financial instability but also exposes management problems within the company, which are likely related to fraudulent activities (Table 1).

3. Model Construction

This study compared two models, logistic regression

(LR) and support vector machine (SVM), for fraud classification. These models are widely used in the relevant literature on enterprise fraud prediction and can balance interpretability and accuracy. Parameter adjustment was carried out by following the grid search method of the previous machine learning framework.

3.1 Logistic Regression

Logistic regression is a generalized linear model that can predict the probability of an event occurring (with a value of 1 if the event happens) based on a set of predictor factors.
Its decision boundary is defined as:

$$P(y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\cdots+\beta_nX_n)}} \qquad (1)$$

To minimize the logarithmic loss function, the training process was carried out using the stochastic gradient descent method.

In the examples studied, fraudulent companies (1) were regarded as the positive class, while non-fraudulent companies (0) were considered the negative class.

3.2 Support Vector Machine (SVM)

The objective of using Support Vector Machines (SVM) is to find a method that can distinguish between fraudulent and non-fraudulent enterprises with the largest margin of separation.
For non-linear cases, we apply the Radial Basis Function (RBF) kernel.

$$K(x_i+x_j)=exp(-\gamma\text{¡}\neg x_i+x_j\text{¡}\neg^2 \qquad (2)$$

3.3 Evaluation Metrics

Due to the uneven distribution of data across various categories, merely looking at the accuracy rate might be misleading. Therefore, the following indicators were employed (Table 2).

Table 2. Evaluation metrics for logistic regression and support vector machines

Metric	Formula	Meaning
Precision	TP / (TP + FP)	Accuracy of predicted frauds
Recall (Sensitivity)	TP / (TP + FN)	Ability to detect frauds
Specificity	TN / (TN + FP)	Ability to detect non-frauds
F1 Score	2 × (Precision × Recall) / (Precision + Recall)	Balanced trade-off
Accuracy	(TP + TN) / Total	Overall correctness

To make the assessment results more reliable, the study added the key indicator of "specificity", as accurately identifying normal companies is equally important.

3.4 Model Validation and Robustness

This study conducted cross-validation and compared the average F1 score. In addition, different training-test ratios and resampling intensities (1:1, 1:2, 1:3) were tested to confirm their stability. All experiments were conducted using the same random seed to ensure the reproducibility of the results.

4. Model Results and Analysis

The logistic regression model achieved the highest accuracy and interpretability, confirming the results of previous studies on financial fraud. Although the stability of the support vector machine model is relatively poor, it confirms the significance of nonlinear effects. This result is consistent with recent research emphasizing integrated and explainable artificial intelligence models for fraud detection.

4.1 Logistic Regression Results

Table 3. Metrics of Logistic Regression Model

Metric	Value
Precision	0. 968
Specificity	0. 671
F1 Score	0. 800
Accuracy	0. 752

Despite the presence of class imbalance, the predictive performance of logistic regression is still quite excellent, and its interpretability enables the identification of the key drivers of financial fraud. The most important positive predictive factors (indicating a higher likelihood of fraud) include a high debt ratio, sudden income growth, frequent auditor changes, and low operating cash flow. On the contrary, stable profitability, high audit quality, and a large number of independent board members become negative predictive factors, and these characteristics are typically associated with a lower fraud risk and stronger corporate governance (Table 3).

4.2 Comparison and Discussion

Table 4. Comparison of Model Performance Across Algorithms

Model	Precision	Recall	Specificity	F1	Accuracy
Logistic Regression	0. 968	0. 745	0. 671	0. 800	0. 752
SVM	0. 490	0. 590	0. 520	0. 560	0. 510

Logistic regression outperforms SVM in all metrics (Table 4).

Moreover, the coefficient-based interpretability of LR allows financial analysts to understand why a firm is flagged as potentially fraudulent.

5. Investment Strategy and Empirical Analysis

This research design followed the standard procedures in financial literature [2][4]. Evidence of short-term abnormal returns indicates the theory of behavioral overreaction [3][4][8]. These results are also similar to those of research in emerging markets, where investor sentiment led to subsequent rebounds [8].

The event was based on the public announcement date of the administrative penalty or investigation decision made by the China Securities Regulatory Commission. The study employed the event study method to analyze the reaction of stock prices before and after the occurrence of this event. Event window: [-15, +15] trading days

Estimation window: [-120, -16] trading days

$$AR_{i,t} = R_{i,t} - E\left(R_{i,t}\right) \tag{3}$$

$$CAR_i(t_1, t_2) = \sum_{t=t_1}^{t_2} AR_{i,t} \tag{4}$$

Calculate the abnormal return (AR) and cumulative abnormal return (CAR) in the following manner:

The expected return () is estimated using market models:

$$R_{i,t} = \alpha_i + \beta_i R_{m,t} + ?_{i,t} \tag{5}$$

The study selected 42 listed companies that were punished by the China Securities Regulatory Commission during the period from 2019 to 2022, covering multiple industries.

The stock price data was obtained from the Wind Database.

For each firm, we computed: 1. Daily returns $R_{i,t}$; 2. Market index return $R_{m,t}$ (CSI 300); 3. AR and CAR for [-15, +15] window

5.1 Empirical Results

Table 5. Market Reaction Analysis

Window	Mean CAR (%)	t-stat	Significance
[-10, -1]	+3. 24	2. 19	p < 0. 05
[0, +1]	-1. 08	-1. 95	p < 0. 10
[+2, +10]	+4. 73	2. 47	p < 0. 05
[+11, +15]	+1. 11	0. 89	ns

The analysis results show that before the announcement was made, the cumulative excess return rate (CAR) slightly increased, which might indicate the existence of insider information or that investors had already anticipated the outcome. After the announcement was released (from day 0 to +1 day), the stock price declined, reflecting the initial pessimistic reaction of investors. However, within the following two weeks, the stock price rebounded, indicating a short-term excessive reaction in the market, followed by a correction (Table 5).

5.2 Trading Strategy Design

Based on the observed patterns, a reverse trading strategy was proposed (Tables 6, 7).

Table 6. Short-term Trading Strategy Table

Step	Action	Timing
1	Identify firms under CSRC investigation	Before announcement (t = -10 ~ -1)
2	Buy and hold	From t = 0 to +15
3	Sell	After 15 trading days

Table 7. Simulated portfolio backtesting yields:

Metric	Value
Average daily return	0. 38%
Cumulative 15-day return	5. 7%
Sharpe ratio	1. 21

Although these returns seem quite attractive, the liquidity restrictions and the small sample size bias may cause the results to be inaccurate.

5.3 Long-Term Effect

To examine the sustainability of short-term gains, this analysis will extend the observation period to 6 months and 1 year after the announcement is made (Table 8).

Table 8. Post-Announcement Excess Return Performance Over Different Time Horizons

Period	Mean Excess Return	Trend
+1 month	+2. 1%	rebound
+3 month	+0. 9%	fade
+6 month	-2. 8%	reversal
+12 month	-5. 6%	continued underperformance

Although non-compliant enterprises may experience a short-term rebound after being punished, their long-term investment returns are worse than those of normal enterprises. The results confirm this theory that market reactions can be excessive.

5.4 Interpretation and Discussion

This result can be explained by behavioral finance:
At first, investors will overreact to negative regulatory news, causing the stock prices of fraudulent companies to be lower than their intrinsic value. However, as uncertainty disappears, the stock prices will return to normal, resulting in temporarily fluctuating returns. Once the fundamental damage situation of the enterprise becomes clear (such as damaged reputation, restricted financing), the long-term prices will fall again.
Therefore, the short-term reverse investment strategy, first, investors will overreact to negative regulatory news, causing the stock prices of fraudulent companies to be lower than their intrinsic value. However, as uncertainty disappears, the stock prices will return to normal, resulting in temporarily fluctuating returns. However, once the fundamental damage situation of the enterprise becomes clear (such as damaged reputation, restricted financing), the long-term prices will fall again.

5.5 Risk and Limitation Analysis

This study has several risks and limitations. Firstly, since the cases covered by the China Securities Regulatory Commission only include those that have been discovered, the cases that have not been identified cannot be included in the observation scope, resulting in data bias. Secondly, the sample size is relatively small, with only 42 incidents, and the general applicability still requires further research. Thirdly, many fraudulent enterprises are mostly small companies with low liquidity. Moreover, the release time of the enforcement announcements is uncertain, and the possibility of information leakage also increases. Finally, during the testing process, transaction costs were not considered, which means that the actual returns may be much lower than the simulation results.

6. Conclusion

This study demonstrates that integrating machine learning with behavioral finance analysis can enhance fraud detection and regulatory insight. This highlights the potential application of artificial intelligence-based early warning

ISSN 2959-6130

systems in financial regulation. However, there are still issues such as a small sample size and limited features. The research results also show that data preprocessing and resampling are very important for addressing the class imbalance problem and improving the model performance. Without these treatments, due to the majority of non-fraudulent samples, the performance of logistic regression (LR) and support vector machine (SVM) would be very poor.

In the data of this study, logistic regression performed better, with an accuracy rate of 0.968, an F1 value of 0.800, and an accuracy rate of 0.752. It was also able to clearly identify some key risk indicators, such as a high debt ratio and frequent changes in auditors.

In contrast, the accuracy rate of SVM was only about 0.51, and it is susceptible to parameter settings and noise. In the future, its performance can be improved by adjusting the function or using a mixed model. The market reaction analysis shows that within 10 to 15 trading days after the penalty announcement is released, the stock will experience a short-term rebound. However, after 6 to 12 months, its performance will be poor, indicating that there is an excessive reaction in the Chinese capital market, and it will make corrections in the following months. Although the trading strategy of buying before the announcement and selling shortly after in the simulation can generate positive returns, its practical application is limited by liquidity, timing, and transaction costs.

References

[1] Beneish M D. The detection of earnings manipulation. Financial Analysts Journal, 1999.

[2] Chen G, Firth M, Gao D N. The information content of earnings announcements: Evidence from the Chinese stock market. China Economic Review, 2005.

[3] Dechow P M, Ge W, Schrand C. Understanding earnings quality: A review of the proxies, their determinants and their consequences. Journal of Accounting and Economics, 2010.

[4] Craja P, Kim A, Lessmann S. Deep learning for detecting financial statement fraud. Decision Support Systems, 2020.

[5] Chen Y, Wu Z. Financial fraud detection of listed companies in China: A machine learning approach. Sustainability, 2022.

[6] Achakzai M A K, et al. Detecting financial statement fraud using dynamic features and machine learning. Journal of International Financial Markets, Institutions & Money, 2023.

[7] Li J, et al. Financial fraud detection for Chinese listed firms: Managers' abnormal tone. Decision Support Systems, 2024.

[8] Cai S, et al. Explainable fraud detection of financial statement data via knowledge graph. Expert Systems with Applications, 2024.

[9] Xu W, Wu Z, Cai J. Revisiting the stock market reactions to private securities litigation against informational misconducts in China. International Review of Economics & Finance, 2024.

[10] CSRC. Official announcements & rules — enforcement and disclosure resources. China Securities Regulatory Commission, 2022.