

Constructing a Fraud Risk Early Warning Model for Chinese Listed Companies Driven by Heterogeneous Data

Yunbai Chen^{1,a}

¹School of Computer Science, Nanjing Audit University, China

^a1043721923@qq.com

Abstract. In this paper, we propose a model for early warning of fraud risk among Chinese listed companies, leveraging the capabilities of heterogeneous data analysis. Given the complexity and multidimensionality of corporate fraud, our model integrates multiple data sources including financial reports, market behavior, and the sentiment of annual report texts, representing diverse sets of heterogeneous data. This approach employs advanced data processing techniques to handle and amalgamate heterogeneous data, ensuring robustness and accuracy. Utilizing machine learning algorithms, the model not only detects potential fraud signals but also quantifies the level of risk, providing stakeholders with a dynamic predictive tool. This research offers a comprehensive data-driven approach to fraud detection in the corporate sector, underscoring the importance of meticulous risk assessment using various data streams. It marks a critical step in proactive fraud management in an increasingly complex financial environment.

Keywords: Heterogeneous Data, Fraud, Listed Companies

1. Introduction

In today's rapidly evolving financial environment, the detection and prevention of fraudulent activities in listed companies are becoming increasingly crucial. The integrity of financial markets largely depends on investor trust in the accuracy of corporate financial statements and disclosed information. Over the past few decades, a series of high-profile corporate scandals have emerged, causing significant economic losses to investors and eroding public confidence in financial markets.

Against this backdrop, constructing an effective fraud detection and early warning model for listed companies is of paramount importance. Traditional fraud detection methods have primarily relied on financial ratios and qualitative analysis. However, the complexity of modern financial fraud requires regulatory and auditing bodies to adopt more advanced technologies to better leverage the power of big data and machine learning. The auditing industry must adapt to new technologies, and auditors need to develop new skills and capabilities to enhance the reliability and accuracy of financial reporting [1]. Recent technological advances have enabled us to collect and analyze a vast array of different datasets, known as heterogeneous data, which include not only structured financial information but also unstructured data from various sources such as social media, news reports, and corporate annual reports.

Integrating heterogeneous data sources offers a unique opportunity to develop more powerful and comprehensive fraud detection models. By analyzing patterns and anomalies in different types of data, these models can identify potential risks that traditional analysis might miss. Furthermore, the application of machine learning algorithms allows for continuous learning and adjustment, thereby enhancing the predictive power of the model over time.

However, constructing fraud detection and early warning models using heterogeneous data is not without challenges. This process involves complex data collection, preprocessing, integration, and analysis. Moreover, ensuring the accuracy, relevance, and timeliness of the data analyzed is crucial. Additionally, the interpretation of results must be handled with care to avoid false positives, which could cause unnecessary panic or harm corporate reputations.

Despite these challenges, the potential benefits of such models are immense. They can provide early indications of financial misconduct, enabling companies, regulatory bodies, and investors to take timely action to investigate and mitigate risks. This not only protects investors but also enhances the overall transparency and integrity of financial markets.

In this paper, we propose a novel framework for constructing a heterogeneous data-driven fraud risk early warning model for listed companies. We discuss the methods for integrating different data sources, the application of machine learning techniques in analysis, and the challenges encountered in this process. Our research contributes to the field of financial fraud detection and prevention, holding significant implications for stakeholders in the financial sector who are seeking to harness the power of data-driven approaches in combating corporate fraud.

2. Research on Constructing a Fraud Risk Early Warning Model for Chinese Listed Companies

2.1 Heterogeneous Data Required for Model Construction

Effectively and accurately collecting heterogeneous data is crucial for constructing a successful early warning model for fraud risk in listed companies. Heterogeneous data mainly falls into two categories: structured and unstructured data. Structured data includes quantifiable and easily measurable variables. Stock returns, financial indicators with specific industry characteristics, and general financial indicators are all part of this category. Stock return, for instance, has been identified by Zhao Jun as a significant driver of financial fraud in Chinese listed companies, suggesting a potential link between stock performance and fraudulent activities[2]. Analyzing stock return data could be instrumental in constructing effective fraud early warning indicators. Industry-specific financial indicators, such as risk factors or performance benchmarks for particular industries, can be very helpful in detecting fraud in companies within those industries. General financial indicators include aspects like revenue growth, debt ratios, and profitability metrics, which can also reveal traces of corporate fraud. According to Khaksar et al. there is a significant positive correlation between financial statement updates and fraud detection[3]. The role of structured data is to provide a solid, quantifiable foundation for the model, offering clear, objective measures critical for identifying financial discrepancies or anomalies.

However, as pointed out by Hoberg and Lewis, a notable drawback of structured financial and stock market data is that fraudulent companies may manipulate current performance indicators and accounting transactions to align with their past performance data or that of their peers, making the fraud less detectable [4]. This implies that sole reliance on statistical indicators may not fully identify fraudulent activities in listed companies. This limitation has led to the decision to utilize the advantages of heterogeneous data for better construction of the fraud risk early warning model for Chinese listed companies. This process inevitably benefits from the inclusion of unstructured data.

Unstructured data, typically textual, requires advanced processing techniques to extract meaningful information. Annual report text data of companies offers insights into the quality of management and potential supervisory red flags. It is a rich source for sentiment and readability analyses, revealing discrepancies between financial data and narrative disclosures. Tetlock used news data to predict stock prices using negative words [5], and studies by Lin Le and Xie Dere also found that the tone of performance discussions positively responds to future earnings [6]. Bushee and others have argued that information asymmetry due to the complexity of disclosed information by listed companies often leads to decision imbalances [7]. Quantifying and analyzing unstructured data is challenging, but crucial for a comprehensive fraud detection approach. It provides context and depth to structured data, enabling a more detailed understanding of a company's financial health and integrity. Similarly, while Loughran and McDonald's research found that using text-based metrics like readability and tone could be effective in identifying fraud [8], some studies remain skeptical of whether text-based metrics can truly capture the intent of managerial misconduct.

Both statistical and text-based indicators have their strengths and weaknesses. The combination of structured and unstructured indicators could potentially compensate for each other's deficiencies in fraud detection capabilities, thereby enhancing the overall fraud risk early warning ability of the model.

2.2 Data Processing

To ensure data quality and the accuracy of the final model, this study identified and addressed missing and outlier values in the structured data and standardized the data format. Various statistical methods were used to analyze and process the stock returns of listed companies, constructing several effective statistical indicators. For industry-specific and general financial data, various data processing methods were employed to build corresponding effective indicators. For unstructured data, such as the text data in annual reports of listed companies, text cleaning, tokenization, and removal of stop words were employed for preprocessing to effectively extract text sentiment indicators and readability metrics. Subsequent sentiment analysis was conducted to construct sentiment tone indicators, ultimately performing feature extraction to integrate the extracted features into a structured form for further analysis or modeling. Both processed structured and unstructured data eventually took on a structured form. This study applied Z-score normalization to all types of indicators, reducing scale differences between various variables, facilitating unified analysis, and constructing the fraud risk early warning model.

Based on the heterogeneous data, this study constructed three main categories of indicators oriented towards early warning of fraud in listed companies: statistical indicators, industry-specific indicators, and text sentiment indicators.

This study is inspired by the work of Bollen and Pools, and we will employ the Probit analysis method to construct a fraud risk early warning model for listed companies [9]. In this approach, if a listed company sample is flagged for regulatory violations in a given year, the dependent variable is marked as 1; otherwise, it is 0. There are two advantages to using Probit analysis for model construction. First, a multivariate setup allows us to determine which indicator variables are more valuable for measuring the fraud risk of listed companies. Second, we can use the fitted values from the Probit to construct a scalar f -score for measuring fraud risk, a calculation method similar to that used by Dechow et al [10]. This method ultimately segments the fraud risk intervals to achieve the purpose of early warning.

In conducting the Probit analysis, we can test whether the three categories of indicators - statistical, industry-specific, and text sentiment - provide incremental information for the fraud of listed companies. Moreover, the coefficient estimates of the Probit analysis provide a method for each company to construct a quantitative indicator f -score, i.e., the ratio of the company's fitted probability of violation under Probit analysis to the unconditional probability of violation. If a company's fitted probability ratio is greater than 1, it implies that considering specific factors, the probability of violation for this company is higher than the overall average. This might indicate the presence of certain high-risk features or conditions for the company. Conversely, if the ratio is less than 1, it suggests that the company has a lower probability of violation after considering relevant factors, possibly indicating strong compliance or lower risk features. If the ratio is close to 1, this might suggest that the company's risk of violation is similar to the general case, or that the factors considered in the model have no significant impact on predicting violations.

Through such methods, we can construct a model based on heterogeneous indicators that quantifies the fraud risk levels of individual listed companies. This model can provide valuable insights for investors, regulatory bodies, and company management, helping them identify potential risk points and regulatory focuses, and thereby adopt appropriate risk management and compliance strategies. Especially for specific countries or markets, it is easier to build more targeted fraud risk early warning models, which can better identify fraudulent behaviors within those specific markets.

3. Summary

This study demonstrates the potential of heterogeneous data-driven methods in constructing an early warning model for fraud risk in listed companies. By integrating various data sources, including both structured and unstructured data, a more detailed and comprehensive analysis of potential fraud indicators can be conducted. The application of advanced machine learning

techniques further enhances the predictive accuracy of the model, making it a valuable tool for stakeholders in business and financial sectors.

The model can analyze a wide range of data types, from structured data in financial statements to sentiment in corporate annual reports, offering a broader perspective on potential fraud risks. In today's complex financial environment, such a comprehensive approach is essential, as traditional methods of fraud detection may have shortcomings. The model has achieved notable success in identifying abnormal patterns that may indicate fraudulent activities, enabling proactive measures to be taken to reduce such risks.

However, this study is not without limitations. Dependence on existing data sources may introduce biases, and the model's current configuration may not capture all forms of financial fraud. Additionally, the dynamic nature of fraudulent practices means that the model needs to be continually updated and refined to maintain its effectiveness.

Future research should focus on expanding the scope and depth of data sources, incorporating more real-time data, and exploring the integration of updated machine learning algorithms. Further investigation could also examine the applicability of the model across different industries and geographic regions, assessing its robustness and scalability.

The significance of this study goes beyond academic research. For regulatory bodies, the model offers a new avenue for monitoring and supervising listed companies. For auditors and internal risk managers, it provides a powerful tool for ongoing risk assessment. For investors, it enhances the transparency and credibility of financial information, aiding in more informed decision-making.

In summary, this paper makes significant progress in the field of financial fraud detection. The proposed warning model utilizes the power of heterogeneous data, offering a more effective, dynamic, and comprehensive approach to identifying fraud risks in listed companies. While there is room for improvement and expansion, it demonstrates the potential of data-driven strategies in combating financial fraud. This represents a step towards a safer and more transparent financial market, solidifying the foundation of investor confidence and corporate integrity.

References

- [1] Balios D, Kotsilaras P, Eriotis N, et al. Big data, data analytics and external auditing[J]. *Journal of Modern Accounting and Auditing*, 2020, 16(5): 211-219.
- [2] Zhao, J. Financial report fraud behaviors and prevention measures in Chinese listed companies. *Market Research*, 2017, (51-52).
- [3] Khaksar J, Salehi M, Lari DashtBayaz M. The relationship between auditor characteristics and fraud detection[J]. *Journal of Facilities Management*, 2022, 20(1): 79-101.
- [4] Hoberg G, Lewis C. Do fraudulent firms produce abnormal disclosure?[J]. *Journal of Corporate Finance*, 2017, 43: 58-85.
- [5] Tetlock P C. Giving content to investor sentiment: The role of media in the stock market[J]. *The Journal of finance*, 2007, 62(3): 1139-1168.
- [6] Lin, L., Xie, D. Do investors listen to the tone? -- An empirical study from the perspective of management tone. *Journal of Finance and Economics*, 2016, 42(7): 28-39.
- [7] Bushee B J, Gow I D, Taylor D J. Linguistic complexity in firm disclosures: Obfuscation or information?[J]. *Journal of Accounting Research*, 2018, 56(1): 85-121.
- [8] Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks[J]. *The Journal of finance*, 2011, 66(1): 35-65.
- [9] Bollen N P B, Pool V K. Suspicious patterns in hedge fund returns and the risk of fraud[J]. *The Review of Financial Studies*, 2012, 25(9): 2673-2702.
- [10] Dechow P M, Ge W, Larson C R, et al. Predicting material accounting misstatements[J]. *Contemporary accounting research*, 2011, 28(1): 17-82.