Analysis of an Intelligent Early Warning Model for Credit Risk in Listed Companies Based on Multi-Source Data Fusion

Yuqiao Shao

School of Finance, Tianjin University of Finance and Economics, Tianjin, 300222, China

jenny20061017@stu.tjufe.edu.cn

Abstract. As global financial markets deepen their integration and accelerate digital transformation, the complexity and significance of credit risk management have become increasingly prominent. This study proposes an intelligent early warning model for listed companies' credit risk based on multi-source data fusion. By integrating financial, public sentiment, online, and compliance data across four dimensions, the model aims to enhance the accuracy and foresight of risk identification. Using non-financial A-share listed companies from 2017 to 2022 as the sample, the study constructed a fusion model based on the XGBoost algorithm and compared it with a benchmark logistic regression model using only financial data. Empirical results show that the fusion model achieved an Area Under Curve (AUC) value of 0.941 and an F1-Score of 0.742 on the test set, significantly outperforming the benchmark model. Feature importance analysis further reveals that alternative data features such as negative sentiment indices and litigation amounts possess predictive power comparable to traditional financial indicators, confirming the incremental informational value of multi-source data in credit risk assessment. This study provides theoretical support and practical reference for intelligent risk control, suggesting future integration with explainable AI technologies to further optimize model transparency and application scope.

Keywords: Credit Risk Early Warning, Multi-source Data Fusion, XGBoost, Intelligent Risk Control.

1. Introduction

With the deepening integration of global financial markets and the acceleration of digital transformation, the complexity and importance of credit risk management have become increasingly prominent. Traditional credit risk assessment models primarily rely on corporate financial indicators, such as Altman's Z-score model and Ohlson's logistic regression model. These models, constructed based on historical financial data, played a significant role in specific historical periods. However, their limitations have become increasingly apparent in the era of big data: First, financial data suffers from lag, making it difficult to capture sudden risk events; Second, they cannot effectively quantify "soft information" such as market reputation, public sentiment pressure, and supply chain relationships. Third, corporate financial statements may contain embellishments, leading to distorted model input data (Duan et al., 2022).

In recent years, with the rapid advancement of natural language processing, graph computing, and machine learning technologies, the application value of alternative data in financial risk management has gradually been recognized. Unstructured, multi-source, heterogeneous data—such as public sentiment data, supply chain network data, and corporate compliance/judicial data—provides richer, more real-time information dimensions for credit risk modeling (Huang et al., 2021; Cao et al., 2023). How to systematically integrate these data to construct more accurate and forward-looking credit risk early warning models has become a shared focus of both academia and industry (Chen et al., 2020).

Against this backdrop, this study constructs a multi-source credit risk early warning model integrating four-dimensional data—financial, public sentiment, network, and compliance—using non-financial A-share listed companies from 2017 to 2022 as samples. Employing the XGBoost algorithm for training and prediction, it compares results with traditional logistic regression models. The aim is to validate the incremental informational value of multi-source data in credit risk assessment, providing theoretical support and practical reference for intelligent risk control.

2. Literature Review and Theoretical Foundations

Credit risk assessment stands as a central concern in financial risk management. Traditional models primarily rely on corporate financial data. The Z-score model pioneered the use of multivariate discriminant analysis to quantitatively predict corporate bankruptcy risk, marking the inception of modern credit risk modeling (Mai et al., 2021). Subsequently, Ohlson employed a logistic regression model, overcoming the stringent data distribution assumptions of Multivariate Discriminant Analysis (MDA) and further enhancing model applicability and accuracy. The core explanatory variables in these classical models all derive from corporate balance sheets and income statements. Despite their historically significant success, inherent limitations have become increasingly apparent: First, financial data exhibits pronounced lag, typically released quarterly or annually, rendering it incapable of capturing sudden corporate risk events; Second, they struggle to quantify "soft information" such as market reputation, management capability, and supply chain stability. Finally, companies may engage in financial statement window dressing, leading to distorted model inputs (Duan et al., 2022).

To overcome these limitations, academia and industry have begun to explore broader data dimensions, driven by the deepening development of information asymmetry theory and behavioral finance. Traditional models rely on "hard information," yet in reality, vast amounts of "soft information" concerning a firm's future solvency permeate the market (Chen and Guestrin, 2016). Tetlock's pioneering research demonstrates that media tone serves as an effective predictor of market volatility, revealing from a behavioral perspective the significance of market sentiment as an information carrier. Subsequent studies have confirmed that sentiment data from news and social media capture market participants' real-time emotions and perceptions toward companies, serving as crucial forward-looking indicators for credit risk. Through natural language processing techniques such as sentiment analysis and topic modeling, unstructured text data can be quantified into usable feature variables (Huang et al., 2021).

Based on social network theory, corporate behavior is not isolated; the characteristics of the embedded relational networks also contain substantial risk information. Cao et al. (2023) found that features such as corporate network centrality extracted using graph algorithms significantly contribute to predicting financial distress (Cao et al., 2023). A company's supply chain relationship network determines both the robustness of its operations and the vulnerability to risk propagation. A firm occupying a critical node in the supply chain or exhibiting excessive dependence on a few key customers/suppliers will inevitably display distinct credit risk characteristics. Furthermore, signaling theory indicates that a firm's compliance and litigation behaviors serve as potent signals to the market regarding its governance quality and operational robustness (Bardos et al., 2020). Administrative penalties and major judicial litigation often foreshadow potential operational crises and financial losses, acting as leading indicators of deteriorating creditworthiness.

These studies collectively form the theoretical foundation for this research's multi-source data fusion approach, addressing different dimensions—market sentiment, network connectivity, and behavioral signals. They demonstrate that a model capable of comprehensively and proactively assessing credit risk must transcend singular financial dimensions. It must systematically integrate diverse information from public sentiment, supply chain networks, and corporate compliance behaviors to more accurately capture a firm's true risk profile.

A core theoretical bridge between traditional financial models and emerging alternative data research lies in re-examining the nature of "information." Classic financial theory assumes market participants can access and comprehend all relevant information at zero cost, yet the real world is rife with severe information asymmetry. An information gap exists between corporate insiders (e.g., management, major shareholders) and outsiders (e.g., investors, creditors). Traditional financial reporting systems serve only as one channel to bridge this gap, limited by their inherent lag and susceptibility to manipulation. The multi-source data integrated in this research fundamentally seeks to capture the "soft information" or "digital footprints" (Shimpi, 2017) that are overlooked by financial statements yet equally, if not more, forward-looking. Public sentiment data reflects market

consensus and emotional fluctuations, embodying collective intelligence; supply chain network data reveals a firm's embedded position and vulnerabilities within real economic activities, representing a direct application of social network theory at the microeconomic level; while compliance and litigation data serve as direct signals of corporate governance effectiveness and operational compliance (Lundberg and Lee, 2017). Therefore, constructing a multi-source fusion model is not merely a simple aggregation of features, but rather grounded in a more comprehensive and realistic theoretical framework: corporate credit risk is the ultimate manifestation of the combined effects of four dimensions—financial health, market sentiment expectations, network structural resilience, and behavioral compliance (Zhang et al., 2023). Only by systematically integrating these signals scattered across different information domains can people construct a "holographic portrait" that more closely approximates a company's true risk profile. This enables sharper and more reliable early warnings in areas where traditional models have proven ineffective—such as predicting sudden "black swan" risk events or identifying financial statement window-dressing (Dou and Xin, 2022).

3. Research Design

3.1. Data Sources and Variable Definitions

The three-dimensional dependent variable (Y) in this study refers to credit risk events. Credit risk events are defined using a binary variable. If a sample company experiences any of the following events within 12 months after the observation period (period t), it is marked as 1 (high risk); otherwise, it is marked as 0 (normal).

Credit rating downgrade refers to a reduction in the issuer's long-term credit rating issued by a domestic authoritative rating agency (e.g., China Chengxin, United Credit Rating). Special treatment (ST/*ST) indicates the imposition of special handling or delisting risk warnings by the stock exchange due to financial anomalies (e.g., consecutive annual losses). This definition integrates third-party professional assessments and regulatory determinations to capture credit deterioration more comprehensively and objectively.

The independent variables (X) employed in this study comprise multi-source risk characteristics. This paper constructed a four-dimensional independent variable set, with all variable data sourced from period t. The financial dimension (X1) incorporates 12 core financial metrics including current ratio, debt-to-equity ratio, return on assets, and revenue growth rate. Data originates from China Stock Market & Accounting Research (CSMAR) and Wind databases. The sentiment dimension (X2) employs web crawling technology to extract headlines and body text related to sample companies from mainstream financial news sites like Sina Finance and East Money, as well as stock forums. An LSTM-based sentiment analysis model calculates daily sentiment scores, aggregated into monthly negative sentiment indices. The Network Dimension (X3) is calculated by constructing a weighted directed supply chain network based on the top five customers and suppliers disclosed in corporate annual reports. The NetworkX library is used to compute each company node's out-degree (number of customers), in-degree (number of suppliers), and PageRank value, thereby measuring its importance within the network and dependency risk. The Compliance Dimension (X4) is calculated by crawling administrative penalty decisions and major litigation announcements involving the sample companies and their subsidiaries from the official websites of the China Securities Regulatory Commission and local courts. It quantifies the number of administrative penalties and the amount involved in major litigation (normalized by total assets) over the t-period.

As shown in Table 1, this table presents a statistical analysis of factors influencing credit risk for A-share listed companies from 2017 to 2022. The dependent variable is credit risk events (Y), while the independent variables encompass four dimensions: financial dimension (debt-to-asset ratio X1-lev, return on assets X1-roa), public sentiment dimension (negative sentiment index X2), online dimension (PageRank value X3), and compliance dimension (amount involved in litigation X4). Each variable reports mean and standard deviation—e.g., the mean debt-to-asset ratio is 43.2%, and the

mean negative sentiment index is 31.8%—reflecting the average levels and volatility across dimensions for the sample companies.

Table 1. Definitions and descriptive statistics of key variables.

Variable Type	Variable Symbol	Variable Name	Mean	Standard Deviation
Dependent	Y	Credit Risk Event	0.057	0.232
Independent				
Financial	X1-lev	Asset-Liability Ratio	0.432	0.211
	X1-roa	Return on Total Assets	0.041	0.056
Sentiment	X2	Negative Sentiment Index	0.318	0.152
Network	X3	Page Rank Value	0.005	0.012
Compliance	X4	Litigation Involved Amount	0.008	0.031

Statistical analysis results for a sample of A-share listed companies from 2017 to 2022

3.2. Model Construction and Research Methodology

To rigorously validate the incremental informational value of multi-source data, this paper constructs and compares the following two models:

The baseline model (Model_Baseline) utilizes only financial dimension (X1) features and employs the Logistic Regression (LR) algorithm.

The Fusion Model (Model_Fusion) utilizes features from all four dimensions (X1, X2, X3, X4) and employs the XGBoost (eXtreme Gradient Boosting) algorithm.

XGBoost is an advanced gradient boosting decision tree algorithm. It efficiently handles mixed-type features, automatically captures complex nonlinear relationships and interaction effects between features, and incorporates built-in regularization to prevent overfitting. This makes it highly suitable for processing the high-dimensional, heterogeneous dataset in this study. Furthermore, XGBoost's built-in L1/L2 regularization and complexity controls for individual trees (e.g., maximum depth, minimum leaf node size) provide robust anti-overfitting capabilities. This is crucial for handling alternative data types prone to noise, such as fluctuations in public sentiment. This comparative design between “traditional linear models” and “modern complex ensemble models” allows us to attribute model performance improvements more purely to the introduction of multi-source feature information itself, rather than the selection of any specific powerful algorithm. This enhances the persuasiveness and robustness of the core conclusion that “multi-source data possesses incremental value.”

The research process is as follows. First, the initial sample of non-financial A-share listed companies from 2017 to 2022 was cleaned (excluding ST stocks and samples with severe data missingness), yielding 15,238 company-year observations. Subsequently, the sample was randomly split into a training set and a test set in a 7:3 ratio. All continuous independent variables were standardized on the training set before being applied to the test set. Model performance was evaluated on the independent test set.

4. Empirical Analysis Results

This paper employs four metrics—precision, recall, F1-score, and AUC—to comprehensively evaluate model performance. The results are presented in Table 2.

Table 2. Model performance comparison on the test set.

Model	Arithmetic	Precision Ratio	Recall Rate	F1-Score	AUC
Model-Baseline	FR	0.632	0.518	0.570	0.872
Model-Fusion	XGBoost	0.781	0.706	0.742	0.941

As shown in Table 2, the results demonstrate that the fusion model significantly outperforms the baseline model across all metrics. The F1-Score (the harmonic mean of precision and recall) increased substantially from 0.570 to 0.742, indicating that the introduction of multi-source data not only enhances accuracy in identifying high-risk enterprises (improved precision) but also broadens coverage (enhanced recall). The AUC value improved from 0.872 to 0.941, approaching the level of a perfect classifier, demonstrating the fusion model's exceptionally strong overall discrimination capability.

The XGBoost model provides feature importance rankings based on gains (Table 2). Analysis reveals that the debt-to-asset ratio (X1_lev) remains the most significant predictor variable. However, the negative sentiment index (X2_nsi) and litigation amount involved (X4_law) both rank among the top five most important features, matching the significance of core financial metrics such as return on assets (X1_roa). This finding provides robust data-driven evidence that alternative data sources—such as public sentiment and compliance metrics—deliver highly predictive incremental insights independent of traditional financial data.

5. Conclusion

This study constructed a credit risk early warning model integrating multi-source data and reached the following key conclusions:

First, the XGBoost model incorporating multi-source data significantly outperformed the baseline logistic regression model using only financial data across all performance metrics. The integrated model achieved an AUC of 0.941 and an F1-Score of 0.742, demonstrating superior accuracy and recall in identifying credit risk events.

Second, feature importance analysis reveals that beyond traditional financial indicators (e.g., debt-to-asset ratio), alternative data features such as the “negative sentiment index” from public sentiment and “litigation amount involved” from compliance dimensions also exhibit strong predictive power. This demonstrates that multi-source data provides incremental risk signals independent of financial information.

In summary, this study confirms the crucial role of multi-source data in credit risk assessment, providing empirical evidence and methodological support for big data-driven intelligent risk control.

This study has certain limitations: First, in terms of data, it primarily relies on public data and does not cover more unstructured data sources (such as corporate recruitment information, supply chain logistics data, etc.). Second, regarding model interpretability, XGBoost, as a “black box” model, still requires further transparency in its decision-making mechanism.

Future research may explore the following directions: First, incorporate more diverse data sources, such as novel alternative data like corporate energy consumption and satellite imagery. Second, integrate explainable artificial intelligence (XAI) techniques to enhance model interpretability and credibility. Third, extend the research framework to non-listed companies or bond markets to validate its universality and transferability.

References

[1] Bardos, K. S., Zhu, J., & Zhang, M. (2020). The value of alternative data in credit scoring: Evidence from a large field experiment. Journal of Banking & Finance, 121, 105961.

[2] Cao, Y., Jiang, J., & Li, X. (2023). The role of supply chain networks in predicting corporate financial distress: A machine learning approach. Journal of Corporate Finance, 78, 102345.

[3] Chen, M., Liu, T., Chen, M., et al. (2020). Early warning model of enterprise default risk based on multi-source big data. Journal of Management Sciences in China, 23(5), 1–25.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(pp. 785–794).

[5] Dou, E. X., & Xin, L. (2022). Application of RegTech in risk early warning: Theoretical framework and China practice. International Finance Research, (4), 12–15.

[6] Duan, T., Feng, J., & Li, Y. (2022). Beyond financials: The value of alternative data in corporate credit risk assessment. Journal of Financial Economics, 145(2), 507–527.

[7] Huang, Y., Li, X., & Zhang, H. (2021). Social media and corporate credit risk: Evidence from China. Pacific-Basin Finance Journal, 68, 101596.

[8] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).

[9] Mai, F., Tian, S., Lee, C., & Ma, L. (2021). Deep learning models for corporate credit risk assessment with textual disclosures. Journal of Financial and Quantitative Analysis, 56(6), 2012–2045.

[10] Shimpi, P. A. (2017). The rise of alternative data: Data protection, privacy and other legal issues. Journal of Investment Compliance, 18(4), 42–47.

[11] Zhang, Y., Li, J., & He, J. (2023). Explainable AI in credit risk modeling: An application of SHAP to corporate loan defaults. Expert Systems with Applications, 213, 118887.