Integrated RFE-XGBoost Credit Risk Prediction for SMEs Using Multi-Source Heterogeneous Big Data

Yuwen Zeng[1], Juan He[2*], Jun Ren[3], Xingyu Liu[4]

zyw94666@my.swjtu.edu.cn [1], hejunlin93@163.com [2*], 24486553@qq.com [3],
liuxingyu@my.swjtu.edu.cn [4]

School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, Sichuan Province, China

Abstract. This paper addresses the challenge of credit risk prediction in the financing of SMEs, focusing on 716 listed manufacturing SMEs from 2018 to 2022. A predictive model based on RFE and XGBoost was developed by integrating multi-source heterogeneous data, including key indicators such as ESG scores, public sentiment, and litigation records. The results indicate that after applying RFE, the average F1 score of XGBoost improved from 0.8822 to 0.9147, an enhancement of approximately 3.25%. This improvement is about 4.75% higher than the next best model (Random Forest, with an average F1 score of 0.8732). These findings underscore the significant role of multi-source heterogeneous data in credit risk management and provide financial institutions with a more advanced tool for risk assessment.

Keywords: Credit Risk Prediction; Multi-Source Heterogeneous, Big Data; Text Mining; RFE-XGBoost;

1 Introduction

Small and medium-sized enterprises(SMEs) are instrumental in driving innovation, increasing employment, and maintaining economic stability[1]. However, factors such as information asymmetry, low financial transparency, and intense market competition result in higher credit risk for SMEs, making financing difficult and expensive[2][3]. Supply chain finance, an innovative financial service that integrates logistics, information, and capital flows, effectively mitigates these financing challenges for SMEs but also increases the risk management demands of commercial banks [4]. The spread of credit risk can affect the entire supply chain and may even trigger systemic risks, undermining the stability and sustainable development of the supply chain. Therefore, the precise prediction and early warning of credit risk in SMEs through scientific methods are crucial.

In the field of credit risk prediction, technologies are primarily divided into two main categories: traditional methods and machine learning methods. Traditional logistic regression models and their variants have been widely applied in credit risk assessments.For instance, Sohn and Kim [5] utilized such a model in conjunction with financial and non-financial factors to effectively improve predictive accuracy, Zhou L. et al.[6] applied the model to construct a credit risk scorecard within the vehicle manufacturing industry. However, these traditional methods may face performance limitations with limited or unevenly distributed samples and rely on the

BDEDM 2024, January 12-14, Ningbo, People's Republic of China

DOI 10.4108/eai.12-1-2024.2347291

assumption of linear relationships between variables, which are not always applicable. To overcome these limitations, researchers have begun to explore machine learning methods, particularly deep learning models. Kuang H. B. et al. [9] used recurrent neural networks to select risk indicators in supply chain finance, while Lu Z. and Zhang J.[10] employed a multilayer perceptron (MLP) to address sample imbalance issues. Wang X. and Wang Y.[11] proposed a method that combines long short-term memory networks (LSTM) with convolutional neural networks (CNN) to further enhance prediction effectiveness. Additionally, Zhu et al[7] and Zhang et al.[8] have explored ensemble learning and firefly algorithm-optimized support vector machines, respectively. Thus, exploring more effective prediction methods continues to be a primary direction for researchers.

SME credit risk prediction is reliant on evaluation indicators, which currently depend mainly on static indicators like traditional financial data. However, these indicators have their limitations. ESG factors play an important role in risk management; for example, RWE[12] has demonstrated the value of ESG in adapting to globalization and regulatory changes. Research [13][14][15] indicates that ESG can improve corporate financial performance and reduce risk. Nevertheless, the application of ESG in credit risk prediction research is limited, partly due to the multidimensionality and quantification challenges of ESG factors, as well as incomplete information disclosure by SMEs[16][17]. Therefore, developing methods that can reflect ESG factors and are applicable to SME credit risk prediction has become urgent. To address these challenges, Zhu Y. et al[18]have proposed the collection and integration of multi-source heterogeneous big data. Furthermore, the use of multi-source heterogeneous data, such as the combination of legal judgment documents and financial information by Yin c. et al[19]. and the integration of financing behavior and demographic data by Zhang et al.[20], has shown potential for a more comprehensive assessment of SME credit risk.

To achieve more precise credit risk prediction and mitigate credit risk issues caused by information asymmetry, this study will integrate multi-source heterogeneous big data. By constructing multi-dimensional indicators from massive textual data, this paper aims to further enrich and refine the methods of predicting SME credit risk. The main contributions of this paper are reflected in the following aspects: First, by employing the XGBoost model and comparing it with classic models such as Logistics, MLP, KNN, SVM, LSTM, and Random Forest, we provide an in-depth and precise methodological reference for SME credit risk prediction and underscore the importance of model selection. Secondly, we analyze the interpretability of the model through feature importance and SHAP values, offering a clear understanding of the decision-making factors, thereby providing support for decision-makers. Finally, our empirical research highlights the effectiveness of multi-source heterogeneous indicators, particularly ESG indicators, in credit risk prediction, validating the core value of ESG indicators in the credit risk management of SMEs and emphasizing the critical role of integrated multi-source data analysis in risk assessment.

2 Feature Engineering and Model Construction

2.1 Construction of credit risk evaluation index system for SMEs

The evaluation indicator system constructed in this paper focuses on multi-source heterogeneous indicators, integrating data from financial reports, news coverage, and litigation records to form a rich heterogeneous dataset that includes ESG scores, public sentiment analysis, and litigation histories. Table 1 provides a detailed display of these multi-source heterogeneous indicators, revealing their role in enhancing the precision of credit risk assessment.

Table 1. Multi-source heterogeneous partial indicator system.

Primary indicator	Secondary indicator
ESG Performance	Environmental Score, Social Score, Governance Score, ESG Composite Score
News Sentiment	Sentiment Attention, Average Sentiment Score
Negative Impact	Interval Litigation Count, Interval Litigation Amount, Interval Regulatory Penalties Count, Interval Regulatory Penalties Amount

In terms of ESG performance, relying on the Wind ESG rating system, this study synthesizes a company's environmental, social responsibility, and governance conditions, covering more than 400 underlying indicators to accurately assess a company's sustainable development capacity.

In terms of news and public sentiment, these serve as vital markers for gauging a company's reputation and market response. This study has amassed over 590,000 pieces of sentiment data related to companies from the Wind and CSMAR databases, covering sentiment risk levels and emotions. The news sources are diversified, including thousands of media outlets, mainly professional financial ones. Two indicators were established: news attention and sentiment score. News attention refers to the total number of company reports within a quarter, while the sentiment score is computed based on the news importance and sentiment emotion, as outlined in Formula 1.

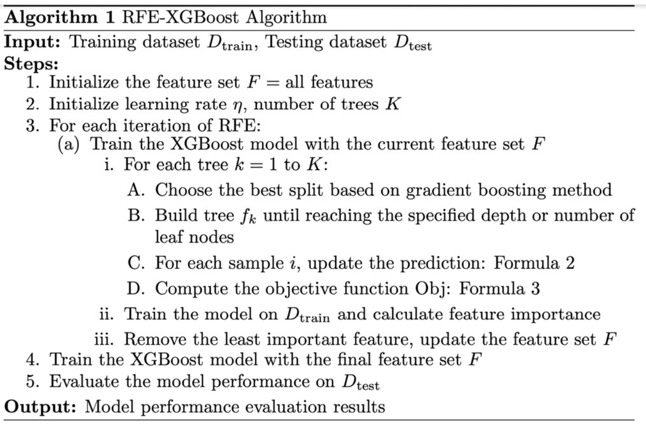
$$S_t = \frac{\sum_{i=1}^{N}(I_i \times E_i)}{N} \tag{1}$$

Within the context of this formula:S_t represents the sentiment score of the news for period t;$N_t$denotes the total number of news articles in period t;I_t signifies the importance of the i news article;E_i is the sentiment value of the i news article. The weighting of news importance is as follows: low importance($I = 1$)、medium importance ($I = 2$)、high importance ($I = 4$). The values for news sentiment are: neutral ($E = 1$)、positive ($E = 2$)、negative ($E = -2$).

In terms of negative impact indicators, a comprehensive consideration was given to various indicators related to legal litigations and securities market compliance, such as the number of lawsuits, the amount of money involved in the litigation, and instances of regulatory penalties. These indicators are compiled on an annual basis to quantify a company's credit status in terms of compliance and legal risk.

2.2 RFE-XGBoost model

Extreme Gradient Boosting (XGBoost) is an efficient machine learning algorithm based on the gradient boosting framework, particularly well-suited for classification problems such as credit risk prediction. XGBoost enhances prediction accuracy by iteratively adding new weak learners, typically decision trees, which is ideal for dealing with the nonlinear and complex patterns often found in credit risk prediction. In applications of credit risk prediction, XGBoost excels at accurately identifying SMEs with high and low risk. Among them, the pseudo code using Feature recursive elimination(RFE) combined with the XGBoost model is shown in Fig 1:



```
Algorithm 1 RFE-XGBoost Algorithm
Input: Training dataset $D_{\text{train}}$, Testing dataset $D_{\text{test}}$
Steps:
  1. Initialize the feature set $F$ = all features
  2. Initialize learning rate $\eta$, number of trees $K$
  3. For each iteration of RFE:
     (a) Train the XGBoost model with the current feature set $F$
         i. For each tree $k = 1$ to $K$:
            A. Choose the best split based on gradient boosting method
            B. Build tree $f_k$ until reaching the specified depth or number of
               leaf nodes
            C. For each sample $i$, update the prediction: Formula 2
            D. Compute the objective function Obj: Formula 3
         ii. Train the model on $D_{\text{train}}$ and calculate feature importance
         iii. Remove the least important feature, update the feature set $F$
  4. Train the XGBoost model with the final feature set $F$
  5. Evaluate the model performance on $D_{\text{test}}$
Output: Model performance evaluation results
```

Fig 1. RFE-XGBoost pseudocode.

Formula 2 and Formula 3 in Fig 1 are as follows:

$$\widehat{y_i^{(t)}} = \widehat{y_i^{(t-1)}} + \eta \cdot f_t(x_i) \tag{2}$$

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(fk) \tag{3}$$

The symbol $\widehat{y_i^{(t)}}$ represents the predicted value for the sample i after the t-th iteration， $f_t(x_i)$ is the output of the newly added weak learner for sample i and η is the learning rate. This iterative process helps to progressively reduce prediction errors and improve model accuracy. $l(y_i, \hat{y}_i)$ is the loss function, measuring the error between the actual values and the predicted values; $\Omega(fk)$ represents the regularization term, which is used to control the complexity of the model. By calculating the gain of each decision tree node, XGBoost can identify which features are most critical for predicting credit risk.

2.3 Project structure

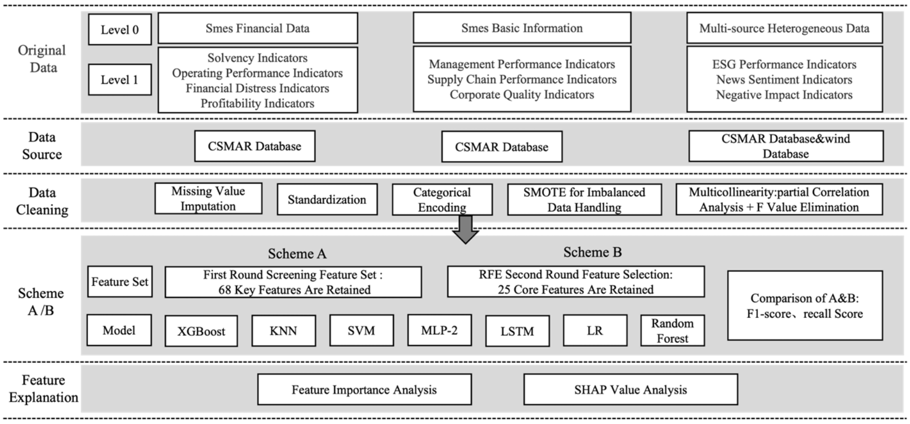

Original Data
Level 0
Smes Financial Data
Smes Basic Information
Multi-source Heterogeneous Data
Level 1
Solvency Indicators
Operating Performance Indicators
Financial Distress Indicators
Profitability Indicators
Management Performance Indicators
Supply Chain Performance Indicators
Corporate Quality Indicators
ESG Performance Indicators
News Sentiment Indicators
Negative Impact Indicators
Data Source
CSMAR Database
CSMAR Database
CSMAR Database&wind Database
Data Cleaning
Missing Value Imputation
Standardization
Categorical Encoding
SMOTE for Imbalanced Data Handling
Multicollinearity:partial Correlation Analysis + F Value Elimination
Scheme A /B
Scheme A
Scheme B
Feature Set
First Round Screening Feature Set : 68 Key Features Are Retained
RFE Second Round Feature Selection: 25 Core Features Are Retained
Comparison of A&B: F1-score、recall Score
Model
XGBoost
KNN
SVM
MLP-2
LSTM
LR
Random Forest
Feature Explanation
Feature Importance Analysis
SHAP Value Analysis

Fig 2. Project structure flowchart.

The flowchart(Fig 2) begins with 'Original Data', which primarily consists of SMEs' financial data and basic information sourced from the CSMAR database. It also includes multi-source heterogeneous data derived from the WIND database, demonstrating the comprehensive and diverse nature of the data sources used in this study.

The next step, 'Data Cleaning', involves several processes to ensure the quality and reliability of the data. Missing values were filled using mean imputation, ensuring that no data point is discarded due to incomplete information. Some data were standardized to maintain consistency and comparability across different scales. Categorical data were transformed using one-hot encoding, facilitating their use in the subsequent modeling processes. Given the issue of credit sample imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed, ensuring a balanced dataset for more robust and reliable results.

The ' Project Design' stage involves model comparison to verify the necessity of RFE and the excellent performance of the combined XGBoost model. Lastly, 'Feature Explanation' provides an understanding of the selected features, emphasizing their relevance and contribution to the model. This step highlights the transparency and interpretability of the model, critical aspects in the field of credit risk prediction.

3 Empirical analysis

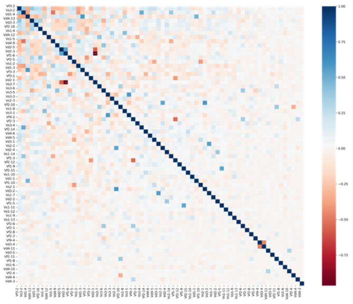
3.1 Sample selection and sources

The research sample selected 716 representative SMEs from the manufacturing industry between 2018 and 2022, all of which are listed on either the Shanghai Stock Exchange or the Shenzhen Stock Exchange. In defining credit risk, two main indicators were used: whether the company has been marked as ST or *ST (indicating consecutive losses for two and three years, respectively, which are special treatments and delisting warnings on the Chinese securities market); and whether the company's interest-bearing debt ratio exceeds the lower value in the "Enterprise Performance Evaluation Standard Values". These criteria are used to judge whether a company has credit risk, with those at risk marked as 1, and those with low credit risk marked as 0. Out of the 9,284 samples, 853 were marked as having credit risk, accounting for 9.19% of the total sample. All data were sourced from the CSMAR and WIND databases.

3.2 Optimal indicator screening

In building a credit risk assessment model for SMEs, a key step is to filter out the most predictive indicators from the vast array of available metrics. For this purpose, this study utilized correlation elimination and RFE methods to optimize the indicator set.

Initially, all indicators underwent a correlation analysis, and features within pairs that had a correlation coefficient higher than 0.6 were scrutinized, with the less important features of the pairs being eliminated. As shown in the correlation heatmap in Fig 3, the selected features exhibit lower inter-correlations after this screening, successfully mitigating the effects of multicollinearity and retaining 68 core indicators.



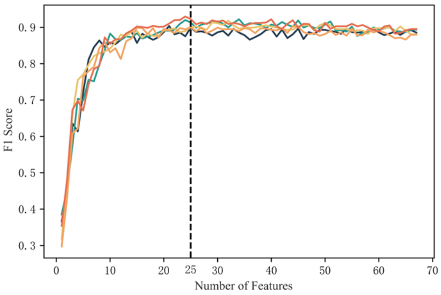


Fig 3. Feature correlation heat map.

Fig 4. Feature selection results.

Subsequently, the feature selection was further refined using the RFE method. RFE progressively removes the least important features and evaluates the predictive power of the remaining features using an XGBoost-based classifier. Throughout the elimination process, the model parameters were kept at their default settings to ensure the robustness of the selection process. As illustrated in Fig 4 after RFE processing, the model achieved the highest average F1 score during five-fold cross-validation with 25 features, indicating that this feature set can provide optimal credit risk prediction performance. The selection of these features took into account both the statistical significance of each indicator and ensured the model's generalizability.

3.3 Grid search tuning

To enhance the precision and stability of the XGBoost model, the Grid Search method was employed to tune key parameters of XGBoost. Table 2 displays the search ranges for three main parameters of XGBoost:

Table 2. Grid search tuning parameters.

Model variable	search scope
learning_rate	[0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3]
max_depth	[3, 5, 7]
n_estimators	[50, 100, 150, 200, 250, 300]

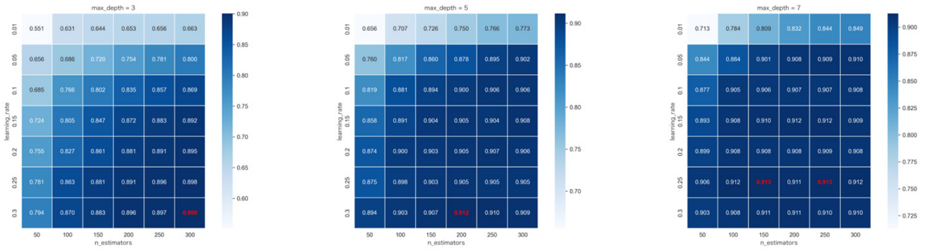


Fig 5. Grid search result graph.

The Grid Search method was utilized to adjust key parameters such as learning rate, maximum depth of trees, and the number of base learners. As illustrated in Fig 5, after the grid search, it was determined that the XGBoost model achieved optimal predictive performance with 250 base learners, a learning rate of 0.5, and a maximum tree depth of 7. The XGBoost model used this set of parameters in subsequent comparative modeling.

3.4 Model comparison

In the comparative study of credit risk prediction models, this paper designed two schemes to assess the impact of model performance and feature selection:

In Scheme A(Non- RFE Method for Feature Selection), the main model and all the comparative models (MLP, KNN, LSTM, SVM, Logistic Regression, and Random Forest) did not adopt RFE. After excluding multicollinearity, the selected feature set contains 68 key features. The primary purpose of this method is to assess the performance of the models without further optimization, serving as a benchmark for evaluating the performance of the RFE-XGBoost model.

Scheme B (RFE-Based Feature Selection Method) adds the RFE optimization process based on Scheme A. After eliminating multicollinearity, the original 68 features were streamlined to 25 core features through the RFE method. The objective of this scheme is not only to simplify the feature set but more importantly, to enhance the model's predictive performance and its generalization ability by selecting the features that contribute the most to the prediction. This method focuses on assessing the effectiveness of RFE in selecting the most impactful features for credit risk prediction and exploring the performance of these carefully selected features in the XGBoost model.

The results indicate that after applying RFE, the average F1 score of XGBoost increased from 0.8822 to 0.9147, an improvement of approximately 3.25%, which is about 4.75% higher than the next best model (Random Forest, with an average F1 score of 0.8732). In terms of recall, XGBoost improved from 0.8922 to 0.9109 after RFE, an increase of approximately 2.09%, which is about 5.28% higher than the next best model (SVM, with an average recall of 0.8652).

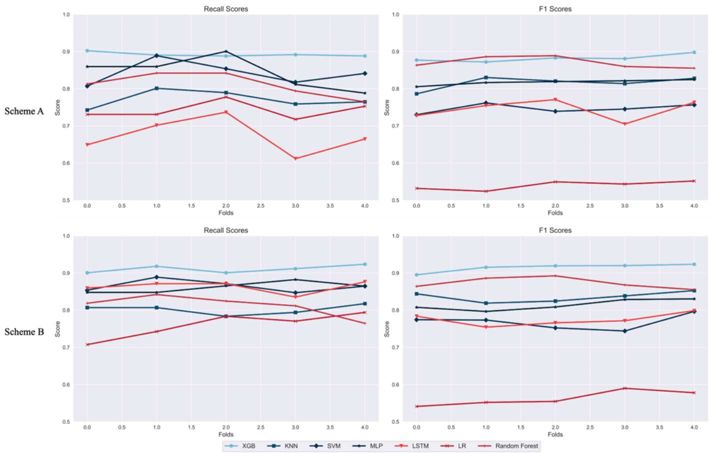


Fig 6. Comparison of results.

Fig 6 clearly illustrates the F1 score performance and recall of the RFE-XGBoost model compared to the benchmark models under Schemes A and B. Based on the remaining features after RFE elimination, the interference from irrelevant variables was removed, simultaneously reducing the complexity of the model and the risk of overfitting. The results attest to the superior performance of the RFE-XGBoost model in predicting credit risk.

3.5 Feature impact analysis

In the feature impact analysis of credit risk prediction models, the importance scores of features and Hapley Additive explanation(SHAP) values are two critical assessment metrics. Positive SHAP values indicate that the feature has an increasing effect on the prediction outcome, while negative values suggest a decreasing effect.

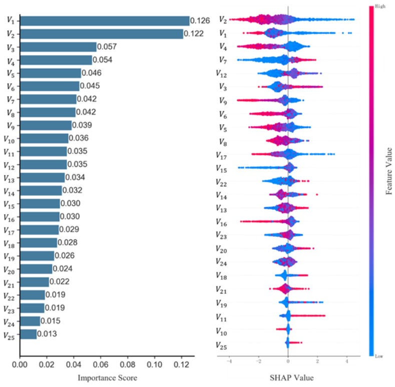


Fig 7: Feature importance and SHAP analysis.

As shown in Fig 7, features such as Z-score warning (V_1), ESG comprehensive score (V_2), and listing year (V_3) have a significant impact on credit risk prediction. Of particular note is that the ESG comprehensive score (V_2) and public sentiment score (V_6) are mainly concentrated on the positive side in the SHAP value distribution, revealing a positive correlation between the decrease in these indicator values and the increase in credit risk level. That is, the lower the ESG comprehensive score and public sentiment score, the higher the credit risk level of SMEs.

Conversely, an increase in feature values such as the environmental dimension score (V_{20}), the number of legal cases during the period (V_{11}), the number of legal cases involved during the period (V_{18}), and the number of regulatory penalties (V_{25}) indicates an increase in corporate credit risk level. These findings not only align with the actual meanings of the features but also enhance understanding of corporate litigation disputes and securities market punishments.In terms of supply chain indicators, the heights of accounts payable, advances from customers, and long-term debt are positively correlated with the credit risk level predicted by the model. These indicators reflect the financial condition and operational risk of the enterprise, so they should be given special attention when warning of SME risks.

In-depth analysis of feature importance and SHAP values reveals the extent of each feature's impact in the credit risk prediction model. Moreover, the integration of multi-source heterogeneous data based on machine learning makes the prediction of SME credit risk more accurate, providing strong support for the management and decision-making of SME credit risk.

4 Conclusion

This study delves into the significance of the Environmental, Social, and Governance (ESG) framework in the credit risk assessment of small and medium-sized manufacturing enterprises. Investigating the effectiveness of credit risk prediction models under the ESG framework from a multi-source heterogeneous perspective, the following conclusions are drawn:

1.The combined RFE-XGBoost model, in comparison to existing models, achieves more accurate predictions with a selected group of optimal indicators through feature selection, highlighting the model's generalization ability and precision.

2.Multi-source heterogeneous data influences the accuracy of SME credit risk prediction. Effective indicators for credit risk prediction are mined from multi-source data such as semi-structured and unstructured news sentiment, litigation, etc., contributing more to the model than traditional financial indicators.

3. Specifically, indicators such as public sentiment score, ESG comprehensive score, the amount involved in legal cases in the interval, and the number of lawsuits in the interval, have a significant impact on prediction outcomes.

The limitations involved in this article are considered opportunities to extend this research. One such limitation is the restriction on the size of the existing dataset, which could affect the performance of complex models requiring large amounts of data to improve expressiveness, such as neural networks and LSTM models. Therefore, future research could contribute to this study by expanding the dataset size.

Acknowledgments. The authors gratefully acknowledge the financial support from the following grants: "The Design and Risk Management of Supply Chain Finance Mechanisms Based on New Financial Technologies" (Project Number: 71873111); "Integrated Development and Benefit Assessment of Global Tourism and Rural Revitalization through Railway Tourism Industry"(Project Number: N2022Z022)

References

[1] Li W., Liu K., Belitski M., et al. E-Leadership through Strategic Alignment: An Empirical Study of Small- and Medium-sized Enterprises in the Digital Age[J/OL]. Journal of Information Technology, 2016, 31(2): 185-206.

[2] Lv J. On Financing Constraints of Small and Medium Enterprises [J]. Journal of Financial Research, 2015(11): 115-123.

[3] Sun J. Analysis on Financing Strategy of Small and Micro Enterprises in Qingdao[C/OL]//2016 5th International Conference on Social Science, Education and Humanities Research (SSEHR 2016). Atlantis Press, 2016: 1280-1284[2023-03-19].

[4] ZhengY,Zhang K X. Research on the Risk Management of the Supply Chain Finance
——From the Perspective of SME Financing [J/OL]. Journal of Financial Development Research, 2020(10): 45-51.

[5] Sohn S. Y., Kim H. S.. Random effects logistic regression model for default prediction of technology credit guarantee fund[J/OL]. European Journal of Operational Research, 2007, 183(1): 472-478.

[6] Zhou L.,Qiu X.,Zhu Y.,et al. An Empirical Study on Credit Risk Assessment in Supply Chain Finance Based on Big Data —— Taking the Vehicle Manufacturing Industry as the Example [J/OL]. ournal of Financial Development Research, 2022(5): 64-70.

[7] Zhu Y, Zhou L, Xie C, et al. Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach[J/OL]. International Journal of Production Economics, 2019, 211: 22-33

[8] Zhang H, Shi Y, Yang X, et al. A firefly algorithm modified support vector machine for the credit risk assessment of supply chain finance[J/OL]. Research in International Business and Finance, 2021, 58: 101482.

[9] Kuang H. B.，Du H.，Feng H. Y.. Construction of the credit risk indicator evaluation system of small and medium - sized enterprises under supply chain finance [J/OL]. Science Research Management, 2020, 41(04): 209-219.

[10] Lu Z,Zhang J. Credit risk Prediction of the Listed Companies Based on SMOTETomek——RFE-MLP Algorithm[J]. Journal of Systems Science and Mathematical Sciences, 2022, 42(10): 2712-2726.

[11] Wang X, Wang Y. Credit Risk Prediction of Small and Medium-Sized Enterprises Based on LSTM-CNN [J]. Journal of Systems Science and Mathematical Sciences, 2022, 42(10): 2698-2711.

[12] ZIOŁO M., BĄK I., SPOZ A.. Incorporating ESG Risk in Companies' Business Models: State of Research and Energy Sector Case Studies[J/OL]. Energies, 2023, 16(4): 1809.

[13] Wang L L, Lian Y H, Dong J. Research on the impact mechanism of ESG performance on corporate value [J]. Securities Market Herald, 2022(5): 23-34.

[14] Qiu M Y, Yin H. An Analysis of Enterprises'Financing Cost with ESG Performance under the Background of Ecological Civilization Construction [J/OL]. Journal of Quantitative & Technological Econmics, 2019, 36(3): 108-123.

[15] Li H, Zhang X., Zhao Y.. ESG and Firm's Default Risk[J/OL]. Finance Research Letters, 2022, 47: 102713.

[16] Cao Q, Xu J. Research on the construction of financial "environmental, social and governance" (ESG) system [J/OL]. Financial Regulation Research, 2019(4): 95-111.

[17] Fang X. M.,Hu D.. Corporate ESG Performance and Innovation:Empirical Evidence from A-share Listed Companies [J]. Economic Research Journal, 2023, 58(2): 91-106.

[18] Zhu Y., Jia R., Wang G. J., et al. A review of supply chain finance risk assessment research: Based on knowledge graph technology[J]. Systems Engineering — Theory & Practice, 2023, 43(3): 795–812.

[19] Yin C., Jiang C., Jain H. K., et al. Evaluating the credit risk of SMEs using legal judgments[J/OL]. Decision Support Systems, 2020, 136: 113364.

[20] Zhang W., Yan S., Li J., et al. Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data[J/OL]. Transportation Research Part E: Logistics and Transportation Review, 2022, 158: 102611.