Vol. 19 (2026): Proceedings of 2026 International Conference on Financial Technology, Mathematical Statistics, and Business Analytics (ICFMB 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/icfmb.202619440447

Enhancing SME Credit Risk Prediction with Soft Information: Evidence from Machine Learning Models

Zihan Zhou*

SILC Business School, Shanghai University, Shanghai, China

**Corresponding author: Zihan Zhou.*

Abstract

This study examines whether integrating soft information with machine learning models can improve credit risk prediction for small and medium-sized enterprises (SMEs). Both financial (hard) and non-financial (soft) variables are incorporated into the analysis by using firm-level data from A-share listed companies. The study compares the performance of logistic regression, random forest, and XGBoost models under different data frameworks. The results show that soft information significantly enhances model performance when combined with hard information and nonlinear models. Feature importance analysis further indicates that although financial variables remain the primary predictors, soft information plays a meaningful role in credit risk prediction. Overall, the findings suggest that integrating soft information with machine learning provides a more comprehensive and effective approach to SME credit risk assessment. This study highlights the value of soft information and emphasizes the importance of combining data diversity with advanced modeling techniques.

Keywords

SME credit risk, soft information, machine learning, random forest

1. Introduction

Small and medium-sized enterprises (SMEs) play a vital role in economic development, contributing significantly to employment, innovation, and overall economic growth in countries. However, SMEs often face a financing dilemma due to limited support through traditional credit channels [1]. Conventional banking systems originally rely on hard financial information to assess borrowers' repayment capacity, but some SMEs lack sufficient financial records or collateral, making it difficult for financial institutions to evaluate their repayment capacity. As a result, improving credit risk assessment for SMEs has become an important issue in both academia and practice.

In credit markets, information asymmetry commonly exists between SMEs and financial institutions. A recent study [2] believes that incorporating soft information can provide additional insights into borrower quality, which can help alleviate the challenges associated with SME credit assessment. Besides, because of superior predictive performance and the ability to model nonlinear patterns, machine learning techniques have been increasingly applied to credit risk assessment [1]. However, limited attempt has been made to study the combination of soft information and machine learning models in the context of SME credit risk prediction.

Many studies have focused on machine learning or the use of soft information to predict the credit risk of SMEs, but there are still some significant gaps in the research. First, rarely existing studies examine the combined effect of machine learning and soft information in a systematic framework. Second, most empirical study is derived from consumer lending or P2P platforms, with limited attention paid to SME credit risk [3]. Therefore, the situation remains unclear whether incorporating soft information with machine learning models can improve credit risk prediction in the SME context. This study aims to address the gap by systematically evaluating the impact of soft information on the predictive performance of different models.

To address the discovered research gap, this study investigates whether the merging of soft information and machine learning can improve SME credit risk prediction. Specifically, the study constructs and compares three models—logistic regression, random forest, and XGBoost by using datasets that include both traditional hard financial information and extended soft information variables. By evaluating the models' predictive performance, the research aims to assess the incremental value of soft information in reducing prediction errors.

2. Literature Review

Historically, credit risk assessment has relied on “hard information”, which is standardized, quantitative factors from financial statements. Early foundational models, such as Altman’s Z-score [4] and Ohlson’s O-score [5], established the dominance of multivariate analysis in predicting default risk. In that case, Logistic Regression (LR) became the main method due to the computational efficiency and interpretability. However, traditional linear models, which are based on the assumption of linear relationships, often fail to capture the complex, multidimensional characteristics of modern credit markets.

The merging of “soft information”, such as firm-specific information and industry factors, is theoretically based on the Information Asymmetry Theory [6]. The theory argues that the mismatch in credit occurs because lenders lack comprehensive data on borrower quality; thus, Soft data has become the key to bridging this gap. Furthermore, Petersen points out that soft information contains forward-looking insights, but financial indicators lack the ability [7]. Grunert et al. empirically prove that combining non-financial factors with financial ratios significantly increases the accuracy of credit rating systems [8]. To process high-dimensional and unstructured soft data, Machine Learning (ML) offers a superior theoretical framework compared to linear regression. Based on Statistical Learning Theory [9], ML methods like Extreme Gradient Boosting (XGBoost) and Random Forests are designed to minimize empirical risk without strict distributional assumptions. Unlike LR, ML models can automatically detect non-linear and hidden relationships between soft information and hard financial data [10]. Because most of the qualitative factors on credit risk are nonlinear, the capability of ML is crucial for verifying the incremental value of soft information.

Despite the importance of soft information being recognized, a significant gap remains. Few studies have employed ML frameworks to systematically quantify the incremental predictive power of soft information relative to traditional hard data. Most existing research focuses solely on a single data source or treats machine learning as a “black box,” lacking collaboration.

Based on the above considerations, this study adopts an ML-based approach to combine traditional financial indicators with multi-dimensional soft information. By comparing the performance of ML models (XGBoost and Random Forest) with traditional benchmarks, this research aims to evaluate the effectiveness of soft information enhancing the accuracy of credit prediction models under the SME scenario.

3. Research Objective

This research investigates whether the inclusion of soft information enhances credit risk assessment for SMEs. In this case, the study estimates two objectives. First, by comparing a dataset containing only hard information with a dataset integrating both hard and soft information, the study evaluates the incremental predictive value of soft information. The reliability of the results is validated through statistical significance tests and area under the curve (AUC) analysis. Second, the study examines the key determinants of credit risk via feature importance analysis, aiming to quantify the relative contributions of financial and non-financial indicators in default risk.

4. Methodology

This study employs an empirical approach to examine whether incorporating soft information improves SME credit risk prediction. Using a dataset of small and medium-sized enterprises, the study constructs a series of models to evaluate the predictive performance of different methodologies. Specifically, the study compares logistic regression with two machine learning models, which are random forest and XGBoost. Furthermore, to assess the incremental value of soft information, the tests design two sets of experiments: one using only hard financial data and the other incorporating both hard and soft information. The overall research framework is explained in the following sections.

4.1 Data and Variables

The study utilizes firm-level data from the A-share market, defining SMEs as firms with total assets below the sample median. The dependent variable is a binary indicator of Special Treatment (ST/*ST), where $Y = 1$ signifies a financial distress event and $Y = 0$ otherwise.

The independent variables are categorized into two feature sets:

- Hard Information (X_H): Standard financial ratios including leverage, profitability, and liquidity.
- Soft Information (X_S): Non-financial indicators such as company age, patent counts, ownership structure, and CEO background.

4.2 Model Specification

To establish a baseline, the study employs Logistic Regression (LR) to model the linear relationship between predictors and default probability. For non-linear patterns, Random Forest (RF) and XGBoost are implemented.

As RF shows a superior performance in preliminary tests, the theoretical framework of RF will be explained further. RF is an ensemble learning method that constructs a multitude of decision trees $\{T_1(x), T_2(x), \ldots, T_B(x)\}$. For a given input x, the final prediction $\hat{f}(x)$ is the result from aggregating the results of B individual trees:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{1}$$

The mechanism reduces variance and prevents overfitting, making RF particularly effective for high-dimensional soft information.

4.3 Empirical Strategy and Hypothesis Testing

The core of the study is quantifying the incremental value of soft information. The study designs two experimental settings:

- Baseline Model: $P(Y = 1) = f(X_H)$
- Augmented Model: $P(Y = 1) = f(X_H, X_S)$

The study uses the Area Under the Curve (AUC) as the primary evaluation standard due to the imbalanced nature of credit default data. To verify whether the improvement in predictive power is statistically significant, the study conducts a Bootstrap test on the difference in AUC $\Delta AUC = AUC_{Augmented} - AUC_{Baseline}$). The hypotheses are formulated as follows:

$$H_0: \Delta AUC \leq 0$$
$$H_1: \Delta AUC > 0 \tag{2}$$

By resampling the test set 1,000 times, the study calculates the p-value to determine if the inclusion of soft information provides a significant improvement to model accuracy. Finally, the study ranks the relative contribution of each variable through Feature Importance analysis.

5. Result

5.1 Overall Model Performance

Table 1 summarizes the predictive performance of different models under various data configurations by using AUC as the primary evaluation metric. The baseline LR model, applied only to hard financial information, achieves an AUC of 0.833. The result establishes a benchmark for credit risk assessment. However, when trained exclusively on soft information, the AUC decreases to 0.813, suggesting that qualitative data alone lack sufficient discriminatory power for actual default prediction. Besides, the LR model's performance remains at 0.833 after integrating both hard and soft information. The lack of improvement indicates that traditional linear models have limitations in capturing the incremental information and nonlinear relationships in soft variables.

In contrast, machine learning models show a markedly different trend. The RF model attains an AUC of 0.813 using only hard information, but the predictive accuracy increases substantially to 0.919 because of the inclusion of soft information. Similarly, XGBoost achieves a high AUC of 0.905 on the combined dataset. The above results suggest that machine learning models effectively explore the synergistic effects between quantitative (hard information) and qualitative (soft information) features, thereby showing superior predictive performance compared to traditional econometric methods.

Table 1: Model Performance Comparison (AUC)

Model	AUC
Logistic (Hard)	0.833
Logistic (Soft)	0.813
Logistic (Hard + Soft)	0.833
Random Forest (Hard only)	0.813
Random Forest (Hard + Soft)	0.919
XGBoost (Hard + Soft)	0.905

To further examine whether the improvement in predictive performance is statistically significant, a bootstrap-based hypothesis test is estimated. According to Table 2, statistical validation via a bootstrap procedure (N=5000) confirms the significance of the incremental predictive power provided by soft information. The analysis discloses a mean AUC increase of 0.1073 (SD=0.0115). With a p-value of less than 0.01, the one-sided hypothesis test confirms that the Full model statistically outperforms the Hard-only baseline. The result strictly validates the research hypothesis, proving that soft information contributes additional information beyond traditional financial indicators, thereby enhancing overall model performance.

Table 2: Result of Bootstrap Hypothesis

Metric	Hard Model	Full Model	Difference
AUC	0.813	0.919	0.1073
p-value	-	-	0.0000

Overall, the above results establish that machine learning models outperform the traditional logistic regression model. Furthermore, soft information can significantly improve predictive performance when combining nonlinear models.

5.2 Feature Importance Analysis

To explain the underlying drivers of model performance, a feature importance analysis ranked by their Gini importance is conducted within the Random Forest. The procedure aims to identify the primary determinants of credit risk prediction. The relative importance of features is explained in Table 2 and Figure 1. The results indicate that quantitative financial metrics remain the most influential predictors of credit risk. Specifically, Earnings Per Share (EPS), profitability, revenue, and total equity exhibit the highest importance scores.

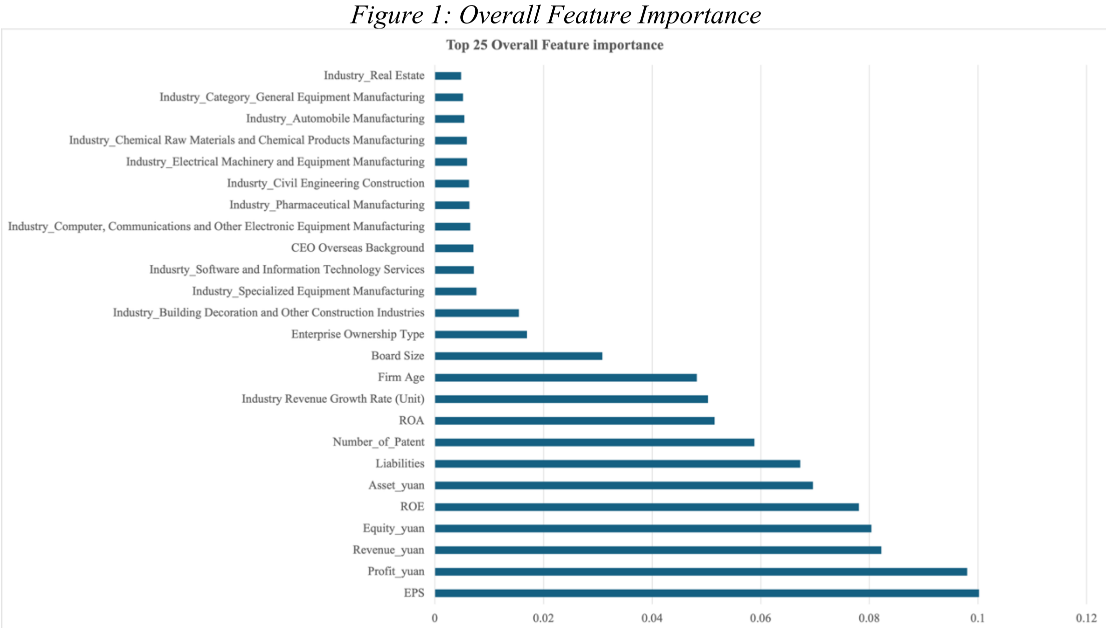
Meanwhile, several non-financial variables demonstrate notable predictive power. Industry-specific indicators and firm-level characteristics rank among the top-contributing features, which reinforces the incremental value of soft information in credit assessment.

In summary, while 'hard' financial data maintains a dominant influence, soft information performs a significant role in the model’s decision-making process, and provides critical insights beyond traditional accounting metrics.

Table 3: Overall Feature Importance in Table

Feature	Gini Importance
EPS	0.100152
Profit_yuan	0.097992
Revenue_yuan	0.082192
Equity_yuan	0.080369
ROE	0.078047
Asset_yuan	0.069612
Liabilities	0.067237
Number_of_Patent	0.058806
ROA	0.051498
Industry Revenue Growth Rate (Unit)	0.050267
Firm Age	0.048227
Board Size	0.030819
Enterprise Ownership Type	0.016939
Industry Building Decoration	0.015464
Industry_Specialized Equipment Manufacturing	0.007643

Figure 1: Overall Feature Importance



6. Key Findings

The empirical results of the study yield three primary findings that underscore the significance of incorporating soft information into credit risk modeling.

Superiority of Non-linear Modeling: The evaluation of predictive performance proves that machine learning—specifically Random Forest and XGBoost—consistently outperforms the traditional Logistic Regression baseline. The result suggests that machine learning models possess a superior capacity to capture the complex, non-linear relation in the SME credit situation.

Statistical Significance of Soft Information: The bootstrap-based hypothesis test provides statistical evidence of the incremental predictive power of soft information. Specifically, the significant improvement in the area under the AUC was validated at the 1% significance level ($p<0.001$), confirming that the inclusion of qualitative indicators provides a significant improvement in model accuracy compared to a hard-information-only baseline.

Determinants of Credit Risk: The feature importance analysis reveals that the data types have a complementary relationship. While fundamental financial metrics (e.g., EPS and profit) remain the dominant predictors, soft information variables—such as industry characteristics and firm-specific qualitative features—rank among the top-contributing features. The result suggests that soft information can serve as a supplement to financial information, effectively alleviating the dilemma of information asymmetry among SMEs.

7. Discussion

7.1 Practical Implications

The findings of this study have several important practical implications for financial institutions, SMEs, and policymakers.

First, for financial institutions, the results suggest that relying on financial indicators only may be insufficient and inaccurate for assessing SME credit risk. Incorporating soft information such as firm characteristics and innovation-related indicators can enhance risk evaluation. It would be more meaningful when combined with machine learning models. This implies that institutions should consider expanding data sources and adopting more flexible analytical frameworks.

Second, for SMEs, the findings highlight the importance of non-financial characteristics in credit evaluation. Factors such as innovation capability, firm age, and governance structure may influence how lenders estimate firms. This finding suggests that SMEs can enhance their financial performance, increase transparency, and highlight non-financial strengths to improve their credit ratings and financing capabilities.

Third, from a policy perspective, the results indicate that improving access to non-financial information may help reduce information asymmetry in SME lending situations. Moreover, encouraging data sharing and supporting the development of alternative data may enhance the efficiency of credit allocation and SME financing.

Overall, these implications suggest that combining soft information with machine learning techniques can provide a more effective approach to credit risk assessment in practice.

7.2 Limitations and Future Research

Despite the contributions of this study, several limitations remain that offer opportunities for further investigation.

First, the scope of the sample is primarily confined to SMEs listed on the A-share market. A-share companies provide a reliable source of data and offer greater transparency compared to unlisted companies. Consequently, the findings may not be fully generalizable to the broader population of private SMEs. Future research could extend the analytical framework to include non-listed private firms to validate the fitness.

Second, the current feature importance analysis merges industry-level indicators and firm-specific characteristics within a single evaluative framework. The approach confirms the overall value of soft information but does not fully distinguish between the macro-influence of industrial sectors and the specific risks of individual firms. Future studies could benefit from a disaggregated approach, analyzing the two categories independently to disclose the unique contributions and potential interactive effects on credit risk. Thus, the analysis would provide deeper insights into whether the predictive power of soft information is driven primarily by industry trends or by specific firm-level indicators.

Overall, addressing the above limitations will further refine the predictive stability of credit risk models. By expanding the sample diversity and enhancing feature granularity, future research can provide more practical decisions for financial institutions in assessing SME default risk.

8. Conclusion

This study investigates whether the combination of soft information and machine learning can improve SME credit risk prediction. By using firm-level data from A-share listed companies, this study compares the performance of traditional logistic regression with machine learning models, including random forest and XGBoost.

The results show that the value of soft information, including patent counts, firm characteristics, and industry-related variables, becomes evident when combined with nonlinear models. Specifically, compared with logistic regression, the performance of the random forest model improves significantly after incorporating both hard and soft information.

Furthermore, feature importance analysis discloses that soft information contributes meaningfully to prediction performance while financial variables remain dominant. This suggests that credit risk is the result of the combined influence of multiple factors, and integrating data from multiple sources can provide a more comprehensive risk assessment.

Overall, this study contributes to the literature by providing empirical evidence that combining soft information with machine learning significantly improves SME credit risk prediction. The findings highlight the importance of adopting more flexible, data-driven approaches in credit evaluation. Future research could further explore richer forms of soft information and more advanced modeling techniques to improve prediction accuracy and applicability.

References

[1] Mhlanga, D. (2021b). Financial Inclusion in Emerging Economies: The application of machine learning and artificial intelligence in credit risk assessment. International Journal of Financial Studies, 9(3), 39. https://doi.org/10.3390/ijfs9030039

[2] Djeundje, V. B., Crook, J., Calabrese, R., & Hamid, M. (2020). Enhancing credit scoring with alternative data. Expert Systems With Applications, 163, 113766. https://doi.org/10.1016/j.eswa.2020.113766

[3] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020c). Explainable machine learning in credit risk management. Computational Economics, 57(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

[4] Altman, E. I. (1968). FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY. The Journal of Finance, 23(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

[5] Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18(1), 109. https://doi.org/10.2307/2490395

[6] Stiglitz, J. E., & Weiss, A. (1981). Credit Rationing in Markets with Imperfect Information. The American Economic Review, 71(3), 393–410. http://www.jstor.org/stable/1802787

[7] Liberti, J. M., & Petersen, M. A. (2018b). Information: hard and soft. The Review of Corporate Finance Studies, 8(1), 1–41. https://doi.org/10.1093/rcfs/cfy009

[8] Grunert, J., Norden, L., & Weber, M. (2004). The role of non-financial factors in internal credit ratings. Journal of Banking & Finance, 29(2), 509–531. https://doi.org/10.1016/j.jbankfin.2004.05.017

[9] Vapnik, V. N. (2013). The nature of statistical learning theory. Springer Science & Business Media.

[10] Lessmann, S., Baesens, B., Seow, H., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.