Comparative Study of Traditional and Machine Learning Models for SME Credit Risk Assessment

Siqi Huang*

Wuhan University, Wuhan, Hubei, China
* Corresponding author

huangsiqi047@163.com

Abstract. Small and medium-sized enterprise (SME) credit risk assessment is an essential challenge for financial institutions because SME borrowers are characterized by high economic relevance, information opacity, diverse operating histories, and sensitivity to loan terms. This study presents an empirical comparison of traditional and machine learning models for SME credit risk assessment using official U.S. Small Business Administration 7(a) Freedom of Information Act (FOIA) open data. Loans approved in FY2010-FY2019 are used to develop the models, while loans approved from FY2020 through the current release are used as an out-of-time test sample. Charged-off loans are treated as default observations, and paid-in-full loans are treated as non-default observations. To minimize information leakage, variables that reveal post-origination outcomes are excluded. Logistic regression is used as the traditional benchmark, and its performance is compared with three challenger models: linear support vector machine, random forest, and XGBoost. The empirical results show that XGBoost performs best on the out-of-time test set in terms of accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC, and Brier score (accuracy 0.8981, precision 0.4416, recall 0.8199, F1-score 0.5741, ROC-AUC 0.9193, PR-AUC 0.5838, and Brier score 0.0754). Feature analysis indicates that maturity, interest-rate structure, secondary-market sale status, SBA subprogram, lender geography, revolving status, guarantee amount, and business age are important risk signals. The findings confirm that machine learning can improve SME default identification under a temporal validation design while preserving financially meaningful model interpretation. Tree-based learning provides stronger discrimination and recall and captures nonlinear interaction effects more effectively, whereas logistic regression remains a useful transparent baseline.

Keywords: SME credit risk; small business lending; logistic regression; support vector machine; random forest; XGBoost; SBA 7(a) loans; ROC-AUC; precision-recall analysis; feature importance

© The Author(s) 2026
K. Othman et al. (eds.), *Proceedings of the 2026 5th International Conference on Mathematical Statistics and Economic Analysis (MSEA 2026)*, Advances in Economics, Business and Management Research 402,
https://doi.org/10.2991/978-94-6239-774-3_34

1 Introduction

1.1 Research Background

SMEs are significant contributors to employment, innovation, local production, and regional economic strength. Recent OECD monitoring shows that SME financing conditions remain a policy concern across member and partner economies [1, 2]. Credit allocation therefore affects business formation, job security, community services, supply-chain resilience, and local tax bases. A credit-risk model is not only a statistical tool, but also a tool for screening, pricing, monitoring, and rationing limited lending capacity.

In the current SME financing landscape, accurate credit-risk assessment is especially important. OECD Scoreboard evidence highlights tight SME borrowing conditions and monitors debt finance, equity finance, asset-based finance, and financing framework conditions [1-3]. A tight credit market increases the cost of misclassification: false negatives allow high-risk borrowers to remain underpriced, while false positives can deny viable SMEs access to credit.

1.2 Motivation and Problem Statement

Financial technology has moved credit-risk modeling from linear scorecards toward model families that can capture nonlinearities, interactions, sparse categories, and threshold effects. SVM separates classes in feature space by maximizing the margin [12]. Random forests use many decorrelated decision trees to reduce variance and preserve nonlinear partitions [11]. Gradient-boosted trees, such as XGBoost, incrementally optimize weak learners and are popular because of their scalable optimization, sparsity handling, and strong predictive power [10]. Machine learning techniques have repeatedly been found to outperform simpler techniques in imbalanced and heterogeneous data environments [5, 16, 17].

The problem addressed in this study is therefore to build a repeatable SME credit-risk assessment workflow that compares traditional and machine learning-based models, validates them over time, and preserves financial interpretation. Logistic regression is chosen as the benchmark because it is widely used, interpretable, and consistent with probabilistic default modeling. The challenger models represent three common approaches: margin-based learning (linear SVM), bagged tree ensembles (random forest), and boosted tree ensembles (XGBoost). The models are evaluated using accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC, Brier score, confusion matrices, ROC curves, precision-recall curves, calibration diagnostics, and threshold analysis.

1.3 Contributions of This Study

This study makes three contributions. First, it uses an empirical design based on real-world SBA 7(a) open data rather than simulated observations or proprietary data. The dataset is relevant to SME lending because SBA 7(a) loans explicitly support small-

business lending, and the public release contains detailed loan terms, program attributes, borrower context, industry codes, geographic identifiers, and final outcomes. Second, it provides a systematic model comparison in which logistic regression, linear SVM, random forest, and XGBoost are trained using the same preprocessing pipeline and evaluated with the same out-of-time test design. Third, it adds an interpretation layer that links feature importance, default-rate segmentation, correlation analysis, threshold behavior, and calibration to lending decisions. Together, these elements explain why the best model works and how financial institutions can translate predictive evidence into operational credit-risk practice.

2 Literature Review

2.1 Traditional Credit Risk Assessment Methods

Traditional credit-risk assessment relies on financial ratios, discriminant methods, and logistic regression, while recent credit-scoring work increasingly combines transparent baselines with nonlinear machine learning models [5, 13].

Logistic regression is popular in financial institutions because of its transparency, auditability, and regulatory familiarity. A coefficient can be interpreted as a marginal change in log-odds, and standardized variables can be examined against economic expectations. Logistic scorecards can also be monitored using stability indices, calibration plots, and approval-rate thresholds. The method is therefore suitable as a benchmark in a comparative design. However, its additive linear structure is limited in SME lending. The effect of a short maturity may differ for borrowers with high interest rates, specific subprograms, or startup status. Interactions may also occur among the guarantee ratio, loan size, and collateral status. Although logistic regression can incorporate interactions, splines, and transformations, the model can become difficult to manage when many categorical variables and nonlinear relationships are present.

2.2 Machine Learning in Credit Risk Assessment

The motivation for applying machine learning to credit risk assessment is its ability to learn complex decision boundaries from past loan outcomes; recent studies also emphasize interpretability and class-imbalance handling [5-7].

Random forests are another modeling approach. They train many decision trees on bootstrap samples and random subsets of features, and then combine the predictions from those trees. Random forests are ensembles of tree predictors that rely on random vectors sampled independently from the same distribution [11]. This structure reduces variance and permits nonlinear decision rules for continuous and categorical variables. In the credit-risk context, random forests can identify interactions among maturity, interest rate, loan scale, industry, and firm age without requiring these interactions to be specified by hand. This comes at the cost of lower transparency than logistic regression, although feature-importance measures and partial dependence analysis can aid interpretation.

Gradient boosting is another major ensemble method in credit scoring. In boosting, a sequence of learners is created, and each learner focuses on the errors left by previous learners. XGBoost extends gradient tree boosting by incorporating regularization, efficient split search, sparsity awareness, and system-level optimization [10]. These properties are desirable for large public loan datasets with many one-hot encoded categorical features, a binary outcome, and missing values. Empirical support for tree-based ensembles appears in credit-scoring benchmark studies. Baesens et al. [5] compared several classification algorithms using real credit datasets and provided benchmark evidence for data mining in credit scoring. Brown and Mues [17] studied imbalanced credit-scoring datasets and found that algorithms including random forests and boosting classifiers performed well. Lessmann et al. [16] updated credit-scoring benchmarks and showed that carefully selected contemporary machine learning methods can significantly enhance predictive power.

2.3 Summary of Existing Studies

Previous studies show that machine learning techniques can outperform conventional models in several credit-scoring applications, although the size of the improvement depends on the dataset, validation strategy, outcome imbalance, features, and evaluation metric. Traditional statistical models remain useful because they are interpretable and provide strong baselines. Machine learning models offer greater representational capacity and can identify nonlinear patterns that cannot be represented by a single additive equation. Under certain conditions, ensemble methods perform especially well on datasets with many predictors and heterogeneous borrowers [16, 17]. However, a model comparison for SME lending should also consider operational implications. Accuracy can be misleading because defaults are a small minority of observations. Precision-recall analysis is more informative when the positive class is rare, while ROC-AUC is useful for evaluating ranking ability across thresholds. Calibration is also crucial because a score used for pricing or expected-loss prediction should be interpreted as a probability rather than only as an ordering statistic.

3 Data Description and Preprocessing

3.1 Dataset Description

The U.S. Small Business Administration 7(a) and 504 FOIA dataset is used in the empirical analysis. According to the SBA open data page, the 7(a) loan program files are separated by decade, a data dictionary is included, and data for the 7(a) and 504 programs are updated quarterly [4]. The analysis uses the 7(a) FY2010-FY2019 file and the 7(a) FY2020-Present file, as of March 31, 2026. This source is useful for SME credit-risk research because it is one of the most prominent channels of U.S. small-business lending and includes final loan status, approval information, borrower state, lender state, business age, NAICS sector, loan term, gross approval, SBA-guaranteed approval, interest rate, revolving status, collateral indicator, secondary-market sale indicator, and other program attributes.

The modeled label is a binary final-outcome label. Charged-off loans are coded as defaults, and paid-in-full loans are coded as non-defaults. Observations with continuing, canceled, committed, or other non-final statuses are excluded from the supervised-learning target because they do not represent clear realized outcomes. This design is conservative because incomplete outcomes, particularly for loans made near the end of the sample period, could contaminate labels. The training period is FY2010 to FY2019, and the out-of-time test period is FY2020 onward. The temporal split is more stringent than a random split because the model is estimated on earlier approvals and assessed on later approvals. The test period includes varying macroeconomic conditions, pandemic-related disruptions, and changes in post-pandemic financing conditions. Therefore, out-of-time validation provides a clearer indication of practical model performance.

The raw FY2010-FY2019 file contains 545,751 records. Among final outcomes, 391,693 loans were paid in full and 30,849 loans were charged off, producing a default rate of 7.30 percent. The raw FY2020-Present file contains 373,981 records. Among its final outcomes, 61,974 loans were paid in full and 5,663 were charged off, producing an overall final-outcome default rate of 8.37 percent. The training set used for model estimation contains 92,547 records, including all 30,849 charged-off loans and a sample of 61,698 paid-in-full loans, resulting in a training default rate of 33.33 percent. This design addresses class imbalance during estimation while preserving a naturally imbalanced out-of-time test set of 67,637 final-outcome loans. The test set is therefore closer to operational experience, where defaults are important but relatively rare. The sample composition and temporal validation design are summarized in Table 1.

Table 1. SBA 7(a) data source and temporal validation design

Sample	Records	Paid In Full	Charged Off	Default Rate Among Final Outcomes
Raw SBA 7(a) FY2010-FY2019	545,751	391,693	30,849	7.30%
Raw SBA 7(a) FY2020-Present	373,981	61,974	5,663	8.37%
Model Training Set	92,547	61,698	30,849	33.33%
Out-of-Time Test Set	67,637	61,974	5,663	8.37%

3.2 Data Preprocessing

Preprocessing aims to make the traditional and machine learning models comparable. The numeric variables include gross approval, SBA-guaranteed approval, initial interest rate, term in months, jobs supported, guarantee ratio, approval year, approval

month, log transformations of loan amounts, and loan amount per supported job. Log transformations are used to reduce skewness in loan scale because SBA approvals range from very small loans to multimillion-dollar loans. The guarantee ratio measures risk sharing and is calculated as SBA-guaranteed approval divided by gross approval. Loan scale is also measured by loan amount per job, which provides an alternative proxy for capital intensity. Recent small-business credit-risk work also supports using borrower and personal-credit signals in machine learning assessment [8].

Business type, business age, NAICS sector, project state, lender state, fixed or variable interest indicator, revolving status, collateral indicator, processing method, SBA subprogram, and secondary-market sale indicator are categorical variables. Missing categorical values are coded as an explicit missing category because missingness can convey information. For example, the missing business-age category has a much higher empirical default rate than most non-missing business-age categories. Missing numeric values are filled with the training-data median. For logistic regression and linear SVM, feature scale matters, so numeric features are standardized. One-hot encoding is used for categorical variables. Tree-based models are less sensitive to standardization than margin-based or coefficient-based models.

3.3 Feature Selection

Feature selection is kept transparent. Candidate variables are screened for economic relevance, leakage risk, missingness, and availability before loan outcome is known. The final feature set covers loan scale, guarantee structure, maturity, pricing, business age, industry, geography, lender state, collateral, revolving status, secondary-market sale status, and SBA subprogram indicators.

4 Methodology

4.1 Traditional Credit Risk Model

Logistic regression serves as the standard benchmark. The model is based on a logit transformation of the linear predictor of the probability of default. Each observation i has probability $p_i = 1 / [1 + \exp(-x_i \beta)]$ of being observed as a default, where x_i is the transformed feature vector and beta is the set of estimated coefficients. The coefficients are estimated by maximizing the likelihood of the observed binary outcomes, and regularization is applied to prevent overfitting in the high-dimensional one-hot encoded space. After calibration checks, the fitted score can be interpreted as a probability, and the signs of the coefficients can be compared with domain expectations. Probability output, simplicity, and interpretability explain why logistic regression remains an important component of credit-risk systems [13].

Logistic regression is likely to perform well when the log-odds of default are approximately linear in the selected variables. However, nonlinearities and interactions are likely to be important risk mechanisms in the SBA 7(a) data. For instance, very short-term loans have a substantially higher default rate than medium- or long-term

loans. The effect of interest rate may also vary by maturity, guarantee ratio, business age, and subprogram. These patterns can be approximated with a single linear additive model, but they may be poorly represented without explicit interaction engineering. Logistic regression is therefore treated as the standard simpler model rather than as the expected optimal model.

4.2 Machine Learning Models

The support vector machine model is implemented as a linear model using standardized numeric variables and one-hot encoded categorical variables. The SVM objective is to find a separating hyperplane with a large margin while allowing errors through slack variables. A linear SVM is justified by the dimensionality of the encoded feature space and by its role as a useful benchmark between logistic regression and nonlinear tree ensembles. SVM methods have been shown to be useful for default classification and feature discovery in credit scoring [6]. The linear SVM score is converted into a decision function and evaluated using threshold-independent ranking metrics.

The random forest model is trained as an ensemble of decision trees. Each tree is learned on a bootstrap sample and uses a random subset of predictors at each split. Aggregation stabilizes individual trees and captures nonlinear interactions without requiring them to be specified in advance. The method is well suited to mixed SME lending variables because a tree can split sequentially on loan term, interest rate, business age, and specific categories. The random forest therefore serves as a strong baseline for nonlinear tree-based learning.

The XGBoost model is an ensemble of gradient-boosted trees. The model fits trees sequentially to minimize loss, with regularization used to control complexity. XGBoost is particularly relevant for credit-risk data because it can use sparse one-hot encoded matrices and boosted trees often perform well in ranking tasks. The algorithm has been described as a scalable tree boosting system with sparsity-aware learning and efficient tree construction [9, 10]. In the current workflow, XGBoost is treated as the main challenger model because credit-scoring benchmarks have shown strong results for boosted tree methods [16].

4.3 Model Evaluation Metrics

Multiple metrics are used because no single metric captures the entire credit-risk decision problem. Accuracy is the proportion of correct predictions, but it can be misleading under class imbalance. Precision is the proportion of predicted defaults that are true defaults and is therefore connected to the number of false alarms. Recall is the proportion of actual defaults captured by the model and is associated with loss prevention. The F1-score summarizes precision and recall. ROC-AUC measures threshold-independent ranking performance by comparing true positive rates and false positive rates across thresholds; ROC analysis is a common method for assessing classifier performance [14]. Because the positive class is sparse in the out-of-time test set, precision-recall curves directly target default detection, so PR-AUC is also included.

The Brier score is the mean squared difference between predicted probabilities and actual binary outcomes, making it useful for assessing probability quality and calibration [15]. A probability or decision threshold of 0.5 is used by default when reporting confusion matrices. For XGBoost, thresholds from 0.1 to 0.9 are also examined because credit-risk systems usually do not rely on a universal 0.5 threshold without considering loss functions and operational capacity. A high-precision threshold can support manual review or capital allocation, while a high-recall threshold can support early-warning watch lists. Calibration curves plot model scores against observed default rates and show whether the scores can be interpreted as probabilities.

5 Experimental Results and Analysis

5.1 Descriptive Evidence from the SBA 7(a) Data

The retained descriptive evidence shows that default risk varies strongly by loan term, business age, industry, state, and contract structure. These patterns justify comparing linear and nonlinear models, because a single global additive equation may not capture segmented SME risk behavior across borrower and loan contexts.

5.2 Performance Comparison of Models

The model comparison shows that machine learning models outperform the traditional benchmark in this out-of-time design. Logistic regression achieves accuracy of 0.6829, precision of 0.1810, recall of 0.7906, F1-score of 0.2945, ROC-AUC of 0.7994, PR-AUC of 0.2341, and Brier score of 0.2063. This high recall indicates that the logistic model captures many defaults, but its low precision means that it also produces many false positives. Linear SVM improves accuracy to 0.7235 while maintaining high recall (0.7708) and achieving a ROC-AUC of 0.8166. Random forest has higher accuracy (0.8387) and precision (0.2831), but lower recall (0.6050), suggesting a more conservative default decision boundary.

XGBoost delivers the best overall performance. Accuracy reaches 0.8981, precision 0.4416, recall 0.8199, F1-score 0.5741, ROC-AUC 0.9193, PR-AUC 0.5838, and Brier score 0.0754. The improvement is not limited to a single metric. XGBoost has the best recall, F1-score, ROC-AUC, PR-AUC, and Brier score among all evaluated models. This combination is important because a model can achieve high recall by flagging too many loans, or high accuracy while missing too many defaults. XGBoost improves discrimination while maintaining a stronger precision-recall trade-off.

The out-of-time performance of the four evaluated models is summarized in Table 2, while the corresponding confusion-matrix components are reported in Table 3. A visual comparison of the main classification metrics is provided in Fig. 1, and the ROC-AUC and PR-AUC performance of the evaluated models is further illustrated in Fig. 2.

Table 2. Out-of-time model performance comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	Brier Score	Fit Time Seconds
XGBoost	0.8981	0.4416	0.8199	0.5741	0.9193	0.5838	0.0754	1.82
Random Forest	0.8387	0.2831	0.6050	0.3857	0.8337	0.4624	0.1605	1.63
Linear SVM	0.7235	0.2005	0.7708	0.3182	0.8166	0.2746	0.3203	1.29
Logistic Regression	0.6829	0.1810	0.7906	0.2945	0.7994	0.2341	0.2063	3.91

Table 3. Confusion-matrix components on the out-of-time test set

Model	True Negatives	False Positives	False Negatives	True Positives
Linear SVM	44,568	17,406	1,298	4,365
Logistic Regression	41,714	20,260	1,186	4,477
Random Forest	53,298	8,676	2,237	3,426
XGBoost	56,104	5,870	1,020	4,643

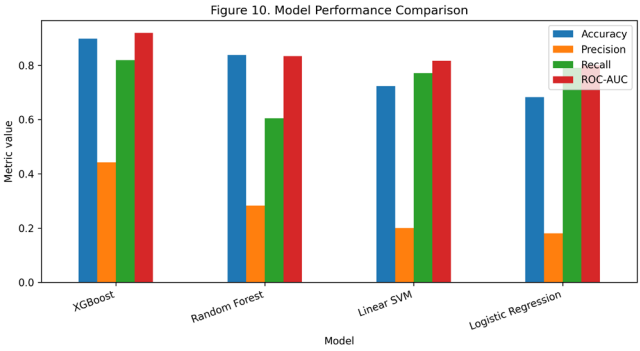

Fig. 1. Model metric comparison across evaluated algorithms.

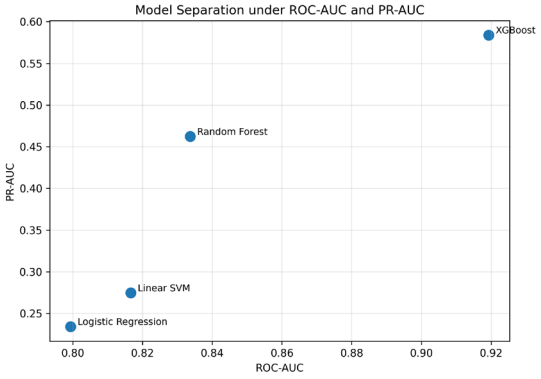

Fig. 2. Model separation under ROC-AUC and PR-AUC.

5.3 ROC Curve and Precision-Recall Analysis

Threshold analysis explains this trade-off. At the XGBoost threshold of 0.1, recall is 0.9578 and precision is 0.1585, which is suitable for a broad watch list but not for adverse underwriting action. At 0.3, recall remains 0.8571 and precision increases to 0.3485. At the default threshold of 0.5, recall is 0.8199 and precision is 0.4416. At 0.7, precision rises to 0.5575 and F1-score reaches 0.6126, while recall declines to 0.6799. At 0.9, precision reaches 0.7330, but recall declines to 0.3394. These findings indicate that model output should not be treated as a universal binary decision, but should instead be integrated into a decision policy. Low thresholds support risk monitoring, medium thresholds support underwriting risk flags, and high thresholds support manual review of the highest-risk cases.

The complete threshold-dependent performance of XGBoost is reported in Table 4. The ROC and precision-recall curves of the evaluated models are shown in Figs. 3 and 4, respectively, while the operating-threshold performance trade-off of XGBoost is illustrated in Fig. 5.

Table 4. XGBoost threshold decision table

Threshold	Accuracy	Precision	Recall	F1-Score
0.1	0.5708	0.1585	0.9578	0.2720
0.2	0.8205	0.3012	0.8661	0.4469
0.3	0.8539	0.3485	0.8571	0.4955
0.4	0.8787	0.3954	0.8492	0.5396
0.5	0.8981	0.4416	0.8199	0.5741
0.6	0.9164	0.5004	0.7770	0.6087
0.7	0.9280	0.5575	0.6799	0.6126
0.8	0.9298	0.5982	0.4927	0.5403
0.9	0.9343	0.7330	0.3394	0.4640

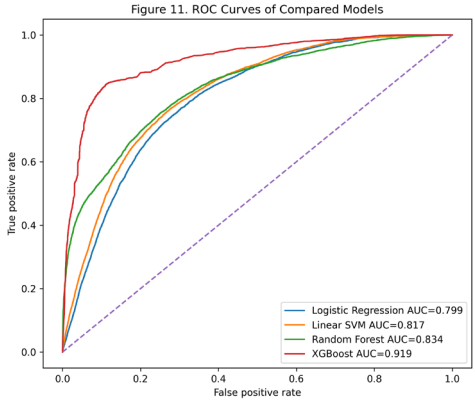

Fig. 3. ROC curves for traditional and machine learning models.

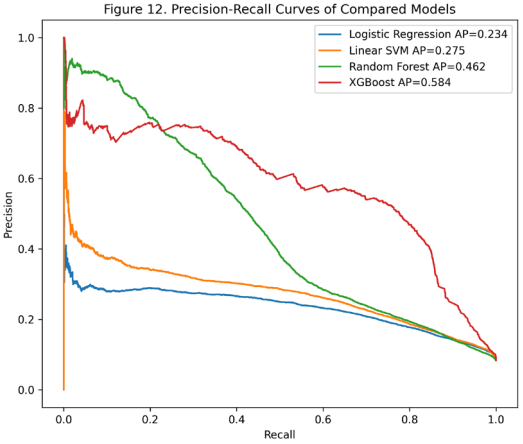

Fig. 4. Precision-recall curves for out-of-time default detection.

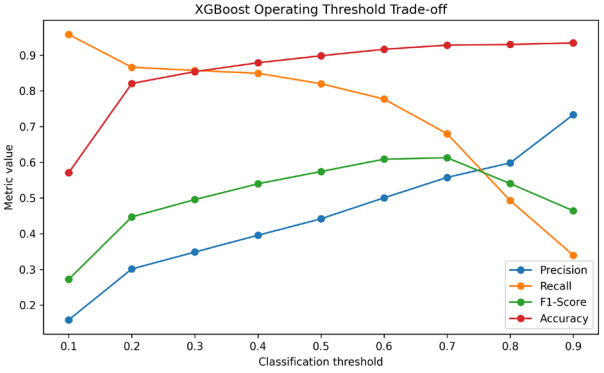

Fig. 5. XGBoost operating-threshold performance trade-off.

5.4 Feature Importance Analysis

XGBoost feature importance indicates which signals drive the performance improvement. The top grouped feature-importance values are reported in Table 5, and their relative distribution is visualized in Fig. 6. Term in months is the strongest grouped feature, with an importance value of 0.1942. This is consistent with the descriptive evidence that short-term loans have much higher default rates. The next important categories are fixed or variable interest indicator, secondary-market sale indicator, subprogram, lender state, revolving status, initial interest rate, processing method, business age, SBA-guaranteed approval, logarithmic SBA-guaranteed approval, and collateral indicator. This feature list is reasonable because it reflects the intersection of contractual maturity, price structure, marketability, guarantee-program design, lender behavior, borrower age, and collateral support.

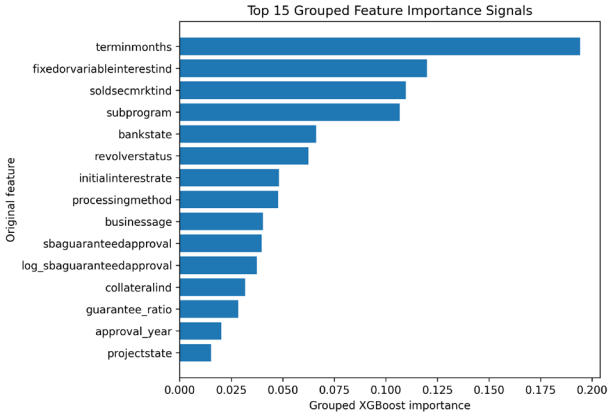

Fig. 6. Grouped feature importance by original variable.

The importance of the fixed or variable interest-rate structure suggests that contract pricing and rate flexibility contain risk information beyond the nominal interest-rate level. The interest-rate level itself is also important, and the mean initial interest rate for charged-off loans is higher than that for paid-in-full loans. This pattern may reflect risk-based pricing, loan-applicant evaluation, or changing credit-market conditions. The role of the secondary-market indicator suggests that loan saleability or secondary-market treatment may carry information about loan quality, program structure, or lender strategy. Institutional differences across the SBA program, reflected in the importance of the subprogram and processing method, also influence default risk, possibly because subprograms differ in borrower needs, risk profiles, or processing procedures.

Business age is important both descriptively and in the model. Business maturity is a key SME risk factor, as shown by the high default rates for missing business-age records and new businesses. This aligns with lending practice: firms with longer operating histories provide stronger evidence of revenue stability, supplier relationships, customer retention, and organizational learning. Lender state and project state are also relevant, but these features require careful governance because geographic signals can reflect multiple mechanisms and may vary over time. If a financial institution implements such features, it should monitor their stability, fairness, and business justification.

Table 5. Top grouped XGBoost feature-importance signals

Original Feature	Importance
terminmonths	0.1942
fixedorvariableinterestind	0.1199
soldsecmrktind	0.1097
subprogram	0.1067
bankstate	0.0662
revolverstatus	0.0624

initialinterestrate	0.0481
processingmethod	0.0478
businessage	0.0403
sbaguaranteedapproval	0.0398
log_sbaguaranteedapproval	0.0375
collateralind	0.0317

6 Discussion

The results indicate that gradient-boosted trees outperform logistic regression because SME credit risk is nonlinear and segmented. Risk varies across loan maturity, interest-rate type, subprogram, secondary-market sale status, guarantee amount, collateral, business age, and geography. XGBoost can represent these interactions more flexibly than a single additive logit model, while logistic regression remains useful as a transparent governance benchmark.

For credit policy, model thresholds should reflect review capacity and loss tolerance. Lower thresholds support broad early-warning lists, while higher thresholds support more selective manual review. Feature importance is economically interpretable, but it should be supplemented with stability monitoring, local explanations, fairness checks, and policy review before production use. Explainable AI research further supports documenting model drivers and decision support logic in credit-risk systems [7].

7 Conclusion and Future Work

7.1 Conclusion

This study compares traditional and machine learning models for SME credit-risk assessment using SBA 7(a) open data and an out-of-time validation design. XGBoost is the best-performing model, with accuracy of 0.8981, precision of 0.4416, recall of 0.8199, F1-score of 0.5741, ROC-AUC of 0.9193, PR-AUC of 0.5838, and Brier score of 0.0754. The results show that boosted trees capture nonlinear SME risk patterns more effectively than logistic regression, while logistic regression remains useful as a transparent benchmark.

References

1. OECD: Financing SMEs and Entrepreneurs 2026: An OECD Scoreboard. OECD Publishing (2026). doi:10.1787/075d8058-en
2. OECD: OECD Financing SMEs and Entrepreneurs Scoreboard: 2025 Highlights. OECD SME and Entrepreneurship Papers (2025). doi:10.1787/64c9063c-en
3. OECD: OECD Financing SMEs and Entrepreneurs Scoreboard: 2023 Highlights. OECD SME and Entrepreneurship Papers (2023). doi:10.1787/a8d13e55-en

4. U.S. Small Business Administration: 7(a) & 504 FOIA dataset, as of March 31, 2026. https://data.sba.gov/en/dataset/7-a-504-foia. Official dataset, DOI not applicable
5. Dumitrescu, E., Hué, S., Hurlin, C., Tokpavi, S.: Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. Eur. J. Oper. Res. 297(3), 1178-1192 (2022). doi:10.1016/j.ejor.2021.06.053
6. Chen, Y., Calabrese, R., Martin-Barragan, B.: Interpretable machine learning for imbalanced credit scoring datasets. Eur. J. Oper. Res. 312(1), 357-372 (2024). doi:10.1016/j.ejor.2023.06.036
7. Nallakaruppan, M.K., Chaturvedi, H., Grover, V., Balusamy, B.: Credit risk assessment and financial decision support using explainable artificial intelligence. Risks 12(10), 164 (2024). doi:10.3390/risks12100164
8. Liu, Z., Liang, H.: The role of personal credit in small business risk assessment: a machine learning approach. J. Risk Model Validation (2025). doi:10.21314/jrmv.2025.016
9. Bücker, M., Szepannek, G., Gosiewska, A., Biecek, P.: Transparency, auditability, and explainability of machine learning models in credit scoring. J. Oper. Res. Soc. 73(1), 70-90 (2022). doi:10.1080/01605682.2021.1922098
10. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. KDD, 785-794 (2016). doi:10.1145/2939672.2939785
11. Breiman, L.: Random forests. Machine Learning 45, 5-32 (2001). doi:10.1023/A:1010933404324
12. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20, 273-297 (1995). doi:10.1007/BF00994018
13. Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, 3rd edn. Wiley (2013). doi:10.1002/9781118548387
14. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. 27(8), 861-874 (2006). doi:10.1016/j.patrec.2005.10.010
15. Brier, G.W.: Verification of probability forecasts. Mon. Weather Rev. 78(1), 1-3 (1950). doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
16. Lessmann, S., Baesens, B., Seow, H.V., Thomas, L.C.: Credit scoring algorithm benchmarks. Eur. J. Oper. Res. 247(1), 124-136 (2015). doi:10.1016/j.ejor.2015.05.030
17. Brown, I., Mues, C.: Classification algorithms for imbalanced credit scoring datasets. Expert Syst. Appl. 39(3), 3446-3453 (2012). doi:10.1016/j.eswa.2011.09.033

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.