

Can Artificial Intelligence Improve Credit Decisions? Evidence from Large-Scale Lending Data

Akash Dania, Ph.D.
Professor of Finance
College of Business
Delaware State University
Email: adania@desu.edu

<https://doi.org/10.68414/YZBK5241>

Abstract

This study examines the use of machine learning for predicting default risk within the U.S. Small Business Administration (SBA) 7(a) loan program using publicly available loan-level data obtained through the Freedom of Information Act (FOIA). Using 488,772 resolved loans originated over fifteen fiscal years from FY2010 through FY2024, the study compares a traditional logistic regression benchmark with Random Forest, Gradient Boosting, and an ensemble model. Gradient Boosting produces the strongest out-of-time predictive performance, achieving an area under the receiver operating characteristic curve (AUC-ROC) of 0.9178 compared with 0.7641 for the logistic regression benchmark, representing a 20.1% relative improvement. The model also reduces the Brier Score by 44.5%, indicating substantially more accurate probability forecasts. Loan term, initial interest rate, and the SBA guarantee ratio emerge as important predictors of default. Default outcomes also vary considerably across industries, with Transportation (9.84%), Arts and Entertainment (9.70%), and Wholesale Trade (8.52%) exhibiting the highest default rates, while Agriculture (4.09%), Finance and Insurance (4.60%), and Healthcare (4.81%) exhibit the lowest. Across the study period, among the resolved-loan sample has an overall default rate of 7.47%, with annual rates ranging from 5.48% in FY2013 to 16.27% in FY2023. The findings demonstrate the potential value of machine learning for credit risk prediction in government-guaranteed small business lending while highlighting the importance of out-of-time model validation. More broadly, the study provides evidence relevant to SBA policymakers, participating lenders, and the responsible development of AI-assisted credit risk assessment tools for small business lending.

Keywords: SBA 7(a) loans, credit risk, machine learning, gradient boosting, random forest, default prediction, small business lending, FinTech

Introduction

Small business lending plays an important role in the U.S. economy by providing entrepreneurs and established firms with access to capital for business formation, operations, investment, and expansion. The U.S. Small Business Administration (SBA) defines eligibility as a small business according to industry-specific size standards, which are generally based on either the number of employees or average annual receipts. Because these thresholds vary by industry, there is no single size criterion that applies uniformly to all businesses participating in SBA programs. This industry-specific approach recognizes the considerable differences in operating scale and market structure across sectors of the U.S. economy.

The economic significance of small businesses in the United States is substantial. They account for approximately 99.9% of U.S. firms, employ nearly half of the private-sector workforce, and have contributed a considerable share of net new employment in recent decades (SBA Office of Advocacy, 2023). Despite their importance, access to credit remains a persistent challenge for many small businesses. Compared with larger firms, small businesses often have shorter credit histories, less extensive financial information, and fewer assets available as collateral. These characteristics can increase information asymmetry between borrowers and lenders, making it more difficult to assess creditworthiness and distinguish between higher- and lower-risk borrowers. Such information constraints can contribute to credit rationing or limit access to financing for otherwise viable businesses. These challenges remain relevant in government-guaranteed lending programs, where the guarantee reduces a participating lender's exposure to potential losses but does not eliminate credit risk or the need for effective borrower assessment. Improving the ability to identify and predict default risk is therefore important both for expanding access to small business credit and for maintaining the financial sustainability of government-supported lending programs.

In FY2023, the SBA 7(a) program, the agency's primary loan guarantee program for small businesses, supported approximately \$27.5 billion in lending. By guaranteeing a portion of eligible loans, the SBA reduces the credit risk borne by participating lenders and helps expand access to financing for small businesses that may have difficulty meeting conventional lending requirements. Given the scale of the program and its broader economic and public policy

significance, understanding the factors associated with loan performance and default is important for both lenders and policymakers. Yet, empirical research examining credit risk within the 7(a) program remains relatively limited, particularly studies that make use of comprehensive loan-level data released through the Freedom of Information Act (FOIA). As a result, there is considerable scope for further examination of default patterns across borrowers, industries, lenders, and loan characteristics.

At the same time, advances in artificial intelligence (AI) and machine learning (ML) have expanded the range of methods available for credit risk assessment. A growing empirical literature suggests that ML approaches can provide improved predictive performance relative to conventional statistical models, including logistic regression, across a variety of lending settings (Berg et al., 2020; Jagtiani & Lemieux, 2019; Moscatelli et al., 2020). These methods are particularly useful when default risk reflects nonlinear relationships or complex interactions among borrower, loan, and market characteristics that may be difficult to specify in advance using traditional model structures.

Despite the increasing use of machine learning in credit markets, its application to SBA 7(a) lending remains comparatively underexplored. Existing research on SBA lending has primarily examined the determinants of loan performance using traditional econometric approaches, while relatively limited evidence exists on the predictive performance of machine learning models using large-scale, longitudinal 7(a) loan data. This gap provides an opportunity to examine whether contemporary predictive methods offer additional value in identifying default risk within a major government-guaranteed small business lending program and whether their performance extends to loans originated under changing economic and credit conditions.

This study addresses this gap by developing and evaluating four credit risk models using a sample of 488,772 fully resolved SBA 7(a) loans originated between FY2010 and FY2024. The analysis draws on official loan-level disclosure data obtained from the SBA through the Freedom of Information Act (FOIA). Logistic regression is used as the traditional benchmark and is compared with three machine learning approaches: Random Forest, Gradient Boosting, and an ensemble model. To provide a more realistic assessment of predictive performance, the data are divided chronologically rather than through a random train-test split. Loans originated between

FY2010 and FY2020 are used for model training, while loans from FY2021 through FY2024 form the out-of-sample test period. This approach more closely reflects the setting in which a credit risk model developed using historical lending experience would be applied to subsequent loan originations. It also reduces the potential for information leakage between the training and testing samples. Beyond comparing model performance, the study examines differences in default risk across industries, loan sizes, business structures, and SBA loan processing methods, providing additional insight into the factors associated with credit outcomes within the 7(a) program.

Our study makes several contributions to the literature. First, it provides a comprehensive empirical examination of default risk within the SBA 7(a) loan program using population-level microdata obtained through the Freedom of Information Act (FOIA). The breadth of the dataset allows default patterns to be examined across a large and diverse set of government-guaranteed small business loans. Second, the study compares the predictive performance of traditional logistic regression with several machine learning methods, providing evidence on whether more advanced modeling approaches offer meaningful improvements in identifying default risk in government-backed small business lending. Third, the findings identify industries, loan characteristics, and program features that are associated with a greater likelihood of default. These results have practical relevance for SBA policymakers and participating lenders, as well as for financial technology firms seeking to incorporate AI-assisted tools into small business credit assessment and underwriting. Across the 488,772 resolved loans included in the analysis, the Gradient Boosting model achieved an out-of-time AUC-ROC of 0.9178, compared with 0.7641 for logistic regression, representing a 20.1% improvement in discriminatory power. It also reduced the Brier score by 44.5%, indicating substantially more accurate probability forecasts.

Literature Review

Credit Risk Modeling in Small Business Lending

The literature on credit risk and small business lending has developed considerably over the past several decades, although relatively few studies have examined default risk using large-scale data from government-guaranteed lending programs. Early contributions by Altman (1968) and Ohlson (1980) helped establish discriminant analysis and logistic regression as

important quantitative approaches for assessing financial distress and default risk. These methods provided the foundation for much of the subsequent empirical work in credit risk modeling. Small business lending, however, presents challenges that are not always fully captured by conventional quantitative models. Unlike larger firms, small businesses often have less extensive financial records, shorter credit histories, and greater variation in the quality and availability of borrower information. Research has therefore emphasized the importance of relationship lending and soft information in evaluating small business borrowers. Frame et al. (2001) and Berger and Udell (2006), for example, show that information developed through interactions between lenders and borrowers can play an important role in small business credit decisions, particularly among community and relationship-oriented banks. Such information may capture aspects of borrower quality that are difficult to observe through financial statements and conventional credit measures alone. This distinction is important because it highlights both the value and the limitations of traditional quantitative approaches when applied to small business credit risk.

Within the SBA lending literature, Glennon and Nigro (2005) used microdata from the SBA 7(a) program to examine factors associated with loan default. Their findings indicate that loan maturity, lender characteristics, and the geographic location of the borrower were among the factors associated with default outcomes. Kalnins and Williams (2014) also used SBA data, focusing on the survival of small businesses and differences associated with ownership structure. While these studies provide important insights into SBA lending and borrower outcomes, they rely on data from earlier periods and predominantly employ traditional econometric techniques. As a result, there remains an opportunity to revisit default risk within the SBA 7(a) program using more extensive loan-level data and contemporary machine learning methods that can capture nonlinear relationships and interactions that may be difficult to identify using conventional statistical models.

Machine Learning in Credit Scoring

The application of machine learning (ML) to credit risk assessment has expanded considerably over the past decade as financial institutions have sought more flexible and accurate methods for evaluating borrowers. A growing empirical literature suggests that ML models can improve default prediction relative to traditional statistical approaches, particularly when large and diverse datasets are available. Berg et al. (2020), for example, demonstrate that models

incorporating digital footprint information can provide substantial predictive value in consumer credit assessment. Jagtiani and Lemieux (2019) find that alternative data and proprietary modeling approaches used by fintech lenders can contribute information beyond traditional credit bureau measures, particularly for borrowers with limited credit histories. Similarly, Moscatelli et al. (2020) provide evidence that machine learning techniques can improve the prediction of corporate credit risk relative to conventional statistical models.

An important advantage of machine learning is its ability to identify relationships that may be difficult to capture through traditional credit scoring approaches. Conventional credit assessment generally relies on a defined set of borrower and financial characteristics, including repayment history, credit utilization, outstanding obligations, length of credit history, and recent credit activity. Models such as logistic regression remain widely used because they are relatively transparent, computationally efficient, and straightforward to interpret. Their performance, however, can depend on how relationships among explanatory variables are specified. Unless nonlinearities and interaction effects are explicitly incorporated into the model, potentially important patterns in the data may remain unidentified (Lessmann et al., 2015).

Machine learning methods provide a different approach by allowing relationships to be learned more flexibly from the data. Tree-based methods such as Random Forest and Gradient Boosting can accommodate nonlinearities, interactions, and threshold effects without requiring each relationship to be specified in advance. Random Forest generates predictions by combining results across multiple decision trees, while Gradient Boosting builds trees sequentially, with each new model attempting to improve upon errors made by the preceding models. These characteristics can be particularly useful in credit risk applications where default may reflect complex interactions among borrower characteristics, loan terms, economic conditions, and other risk factors (Butaru et al., 2016; Moscatelli et al., 2020).

The distinction between traditional statistical modeling and machine learning also reflects differences in their primary objectives. Econometric models are frequently used for both prediction and statistical inference, allowing researchers to estimate and interpret relationships between individual explanatory variables and an outcome of interest. Machine learning methods place greater emphasis on predictive performance, particularly the ability of a model to generalize to observations that were not used during estimation. Techniques such as

cross-validation, regularization, and out-of-sample testing are commonly employed to limit overfitting and assess how well a model performs on new data. Comparative studies generally find that logistic regression remains a useful and competitive benchmark in credit risk applications, while tree-based and ensemble methods can provide additional predictive gains when relationships in the data are nonlinear or involve complex interactions (Lessmann et al., 2015; Butaru et al., 2016).

Taken together, the existing literature provides substantial evidence that machine learning can complement traditional approaches to credit risk assessment (Lessmann et al., 2015; Butaru et al., 2016; Berg et al., 2020; Jagtiani & Lemieux, 2019; Moscatelli et al., 2020). Much of this evidence, however, comes from consumer, fintech, and commercial credit markets. Comparatively less attention has been given to government-guaranteed small business lending, where borrower characteristics, lender incentives, and program requirements may produce different patterns of credit risk. The SBA 7(a) program therefore provides a useful setting in which to examine whether the predictive advantages documented elsewhere in the credit literature extend to government-guaranteed small business loans. The present study contributes to this literature by comparing logistic regression with Random Forest, Gradient Boosting, and an ensemble approach using a large population of SBA 7(a) loans.

SBA 7(a) Program and Default Risk

Research on the SBA 7(a) loan program has generally focused on how lender characteristics, borrower attributes, loan terms, and program design influence loan performance. Craig et al. (2008), for example, find that lenders with greater experience in SBA lending tend to have lower default rates, suggesting that repeated participation may contribute to stronger borrower screening and underwriting practices. Their findings point to the importance of specialized lending expertise and institutional learning in determining loan outcomes. Cole and Dietrich (2012) similarly document variation in loan performance across industries and geographic areas, indicating that local economic conditions, industry characteristics, and regional business environments can influence repayment outcomes.

The broader literature identifies several factors associated with small business loan performance. Borrower characteristics, including business age, financial condition, credit history,

and managerial experience, can influence default risk. Loan-specific features such as size, maturity, collateral, guarantee percentage, and interest rate may also affect repayment outcomes. At the same time, broader economic conditions, including employment trends, economic growth, and regional business activity, can influence the financial capacity of small firms to meet their debt obligations. Taken together, this literature suggests that small business loan performance is shaped by a combination of borrower, lender, loan, industry, and economic characteristics rather than by any single factor.

Credit risk assessment within the SBA 7(a) program is further influenced by the program's institutional structure. Unlike conventional commercial lending, eligible 7(a) loans receive a partial federal guarantee, which reduces the lender's exposure to losses while supporting access to financing for small businesses that may otherwise face difficulty obtaining credit on reasonable terms. The presence of a guarantee, however, does not eliminate credit risk or the need for effective borrower screening and monitoring. Accurate assessment of repayment capacity therefore remains important to participating lenders and to policymakers responsible for maintaining the effectiveness and financial sustainability of the program.

A related body of SBA research has examined the Paycheck Protection Program (PPP), which was established as a temporary emergency assistance program during a period of significant economic disruption. Although administered by the SBA, PPP differed substantially from the traditional 7(a) program in both its objectives and design. The program was intended primarily to provide rapid financial support to eligible businesses and incorporated streamlined eligibility requirements, expedited processing, and provisions for loan forgiveness. Consequently, the credit assessment and repayment considerations associated with PPP loans differ from those underlying traditional 7(a) lending. Findings from the PPP literature therefore provide useful evidence on emergency small business financing but are less directly applicable to the study of long-term default risk within the 7(a) program.

Despite considerable research on the determinants of SBA loan performance, relatively few studies have applied contemporary machine learning methods to default prediction using large, longitudinal datasets from the 7(a) program. Much of the existing empirical literature relies on traditional econometric approaches, including logistic regression and survival analysis. These

methods remain valuable for estimation and inference but may require researchers to specify nonlinear relationships and interactions among explanatory variables explicitly. In a lending environment as heterogeneous as the SBA 7(a) program, relationships among borrower characteristics, lender behavior, loan terms, industry conditions, and default outcomes may be more complex. Machine learning methods, particularly tree-based ensemble approaches such as Random Forest and Gradient Boosting, provide a flexible means of identifying such relationships without requiring them to be specified in advance (Lessmann et al., 2015; Butaru et al., 2016).

This study extends the existing literature by examining whether machine learning methods provide additional predictive value in assessing default risk within the SBA 7(a) program. Using 488,772 fully resolved loans originated over fifteen fiscal years, the analysis compares a traditional logistic regression benchmark with Random Forest, Gradient Boosting, and an ensemble model. The study therefore brings contemporary predictive methods to a large government-guaranteed small business lending dataset while preserving a direct comparison with a widely used traditional credit risk model. In doing so, it provides additional evidence on the determinants and prediction of default risk within the SBA 7(a) program and contributes to the broader literature on the application of machine learning to small business credit assessment.

Data

Source and Sample Construction

Data for this study were obtained from the U.S. Small Business Administration (SBA) Freedom of Information Act (FOIA) public loan disclosure files available through the SBA web portal. The SBA publishes loan-level information on loans originated under the 7(a) program and periodically updates these disclosure files. For this study, data covering fiscal years (FY) 2010 through FY2024 were compiled by combining the SBA's FY2010–FY2019 historical disclosure file with the subsequent active disclosure files available as of March 31, 2026.

The study period begins in FY2010, the first full fiscal year following the 2008–2009 Global Financial Crisis and a period of significant changes in federal small business lending policy under the American Recovery and Reinvestment Act (ARRA). Beginning in FY2010 provides a clearer post-crisis starting point and reduces the potential influence of the exceptional

lending conditions that characterized the financial crisis. Extending the sample through FY2024 provides a fifteen-year observation window encompassing different economic and credit environments, including the post-recession recovery, an extended period of economic expansion, a period of substantial economic disruption, subsequent increases in interest rates and tighter financial conditions, and the adjustment of lending markets that followed. This extended time horizon provides an opportunity to evaluate SBA loan performance across substantially different economic conditions.

The combined raw dataset contains 815,187 loan observations covering FY2010 through FY2024. To construct a clearly defined binary outcome for supervised classification, the analytic sample is restricted to loans with resolved repayment outcomes: Paid in Full (PIF) or Charged Off (CHGOFF). Loans reported with unresolved or intermediate statuses, including CURR, EXEMPT, COMMIT, LIQUID, DELINQ, PSTDUE, and DEFERD, are excluded because their ultimate repayment outcomes cannot yet be determined from the available data. Including such observations would introduce uncertainty into the dependent variable and, for loans that remain active, raise concerns related to right censoring. Restricting the modeling sample to resolved loans therefore ensures that each observation used for model estimation and evaluation has an observed repayment outcome.

Cancelled loans (CANCLD) are also excluded because these loans do not represent completed credit outcomes and therefore provide limited information about subsequent borrower repayment performance. After applying these criteria, the final analytic sample consists of 488,772 resolved loans, representing approximately 59.9% of the 815,187 observations in the original dataset. The dependent variable is consequently defined as a binary measure distinguishing loans that were paid in full from those that resulted in a charge-off.

The resulting dataset provides a large longitudinal sample of observed SBA 7(a) loan outcomes across fifteen fiscal years. Its size and time coverage make it particularly suitable for evaluating whether machine learning methods provide incremental predictive value relative to traditional credit risk models. At the same time, restricting the analysis to loans with resolved outcomes is an important sample consideration, particularly for loans originated in the later years of the study period, which have had less time to reach a final repayment status. This issue is addressed in the empirical design and should be considered when interpreting the findings.

A Note on Sample Resolution and Right-Censoring

Restricting the sample to loans with fully resolved outcomes (PIF or CHGOFF) introduces a right-censoring consideration for the most recent vintage years. SBA 7(a) loans carry terms of up to 300 months, and a charge-off typically follows a period of delinquency and workout; loans originated in FY2022–FY2024 have therefore had comparatively little time to reach a resolved status relative to earlier vintages. Consequently, the FY2021–FY2024 test-period default rate (9.82%) is more likely to reflect loans that defaulted quickly than the eventual full default experience of these vintages, which may understate the long-run default rate for the most recent cohorts and could affect the calibration, though less the rank-ordering, of out-of-time performance. We view the reported AUC-ROC and KS statistics, both threshold-independent measures of relative risk ranking, as more robust to this form of censoring than the absolute default-rate and Brier Score figures, and we flag this as an important caveat for readers applying these results to loss forecasting rather than risk ranking.

Outcome Variable

The binary outcome variable Default is coded 1 for loans reported as CHGOFF (Charged Off), and 0 for loans with status PIF (Paid in Full). This definition is consistent with standard practice in SBA credit risk research (Glennon & Nigro, 2005; Cole & Dietrich, 2012). The overall default rate in our sample is 7.47% ($n = 36,494$ defaults out of 488,772 resolved loans).

Features and Predictors

We construct predictor variables from the raw SBA FOIA fields, organized into four categories:

- **Loan Structure:** Log of gross loan amount, SBA guarantee ratio (SBA guaranteed amount / gross approval), loan term in months, initial interest rate, fixed vs. variable rate indicator, revolver vs. term loan indicator, and collateral indicator.
- **Lender/Program Characteristics:** Processing method classification (grouped as Preferred Lenders, Express Programs, Community Advantage, and Standard 7(a)), and secondary market sale indicator.

- Borrower Characteristics: Business type (Corporation, Individual, Partnership), and business age classification (New Business / Startup vs. Established, available from FY2018 forward).
- Sector and Geography: NAICS two-digit sector dummies (18 sectors) capturing industry-level systematic risk. State-level variation is captured implicitly through sector composition.

Variable Availability, Missing Data, and Model Tuning

All loan characteristics, lender and program variables, and industry classifications included in the analysis are obtained directly from the SBA FOIA disclosure files and reflect information recorded at the time of loan approval or origination. Two variables, however, require additional explanation. First, the business age classification, which distinguishes New Business/Startup from Established firms, is available only for loans originated in FY2018 and later. Rather than imputing values for earlier loans, observations from FY2010–FY2017 are retained as a separate Not Reported category. As a result, the estimated effect of business age is driven primarily by loans originated between FY2018 and FY2024. Second, the indicator for secondary market sale reflects the status reported in the SBA disclosure files. Because some loans may have been sold after the original underwriting decision, this variable should be interpreted as being associated with realized loan performance rather than as a characteristic available to lenders at the time of loan approval.

Missing data are minimal for the principal loan characteristics used in the analysis, including loan amount, interest rate, loan term, and guarantee ratio. Observations with missing values for these variables are removed using listwise deletion because the number of affected loans is very small. The only variable with a meaningful level of missing data is the NAICS industry classification, which is unavailable for approximately 1.6% of loans. These observations are excluded only from the industry-level analyses and summary tables.

The Random Forest and Gradient Boosting models were estimated using hyperparameter values commonly reported in the credit risk literature for datasets of similar size and class distribution. Rather than conducting an extensive grid search, we selected well-established parameter settings and evaluated their stability using five-fold cross-validation in addition to the

holdout test sample. The cross-validation results provide evidence that the selected model specifications perform consistently across different data partitions.

Predicted default probabilities are reported directly from each model without additional probability calibration. Consequently, the reported Brier Scores reflect the models' raw probability estimates rather than values adjusted through post-estimation techniques such as Platt scaling or isotonic regression. Future research could explore systematic hyperparameter optimization and probability recalibration, particularly in applications where accurately calibrated default probabilities are required for capital allocation, reserve estimation, or regulatory risk management, rather than solely for ranking borrowers by credit risk.

Table 1: Descriptive statistics for the key continuous variables in the analytic sample for key variable, SBA 7(a) Resolved Loans FY 2010-FY2024.

Variable	N	Mean	Std Dev	Min	Median	Max
Gross Loan Amount (\$)	488,772	\$366,599	\$680,270	\$1,000	\$109,000	\$5,000,000
SBA Guarantee Ratio	488,772	0.65	0.15	0.08	0.50	1.00
Initial Interest Rate (%)	488,772	6.60%	1.72%	0.00%	6.00%	16.50%
Loan Term (Months)	488,772	119.6	76.8	1	120	300
Jobs Supported	488,772	10.65	20.54	0	4	2,150
Default (0/1)	488,772	0.0747	0.263	0	0	1

Methodology

Temporal Train-Test Split

To simulate realistic model deployment and avoid data leakage, a critical concern in credit risk model evaluation (Japkowicz & Shah, 2011), we employ a temporal holdout design rather than random cross-validation. We train all models on FY2010–FY2020 observations ($n = 444,716$) and evaluate performance on the out-of-time test set comprising FY2021–FY2024 ($n = 44,056$). This design tests whether models estimated on pre-pandemic data generalize to the post-pandemic credit environment, which saw elevated default rates in FY2022–FY2023. The training default rate is 7.23%; the test set default rate is 9.82%, reflecting the challenging post-pandemic economic environment.

Model Specifications

To evaluate the predictive performance of traditional statistical methods relative to modern machine learning algorithms, we estimate four competing models that represent increasing levels of model complexity and predictive capability. The selected models are among the most widely used approaches in credit risk analytics and provide a comprehensive comparison between interpretable benchmark methods, ensemble learning algorithms, and hybrid predictive frameworks.

- **Model 1 - Logistic Regression (Benchmark):** Estimated using the five traditional loan structure variables (log loan amount, interest rate, term, collateral, and guarantee ratio) with standardized inputs. Logistic regression serves as the benchmark because it remains the dominant statistical approach used in consumer and commercial credit risk modeling due to its simplicity, interpretability, and regulatory acceptance. The model represents a parsimonious traditional underwriting scorecard analogous to those employed by many community banks and provides an appropriate baseline against which the performance gains of machine learning methods can be evaluated.
- **Model 2 - Random Forest:** An ensemble of 200 decision trees estimated using the full 19-variable feature set plus NAICS sector indicator variables. Model parameters include a maximum tree depth of 10, a minimum of 50 observations per terminal node, balanced class weighting to address the 92.5%/7.5% class imbalance, and bootstrap aggregation with $\sqrt{p}$ randomly selected predictors at each split. Random Forest was selected because it effectively captures nonlinear relationships and high-order interactions while reducing overfitting through bootstrap aggregation. It also provides robust performance in high-dimensional credit datasets and is relatively insensitive to multicollinearity and outliers.
- **Model 3 - Gradient Boosting:** Estimated using 200 sequential boosting iterations on the same full feature set with a maximum tree depth of 4, learning rate of 0.05, and subsample ratio of 0.80. Gradient Boosting was selected because it has consistently demonstrated superior predictive performance in the credit risk literature (Lessmann et al., 2015; Butaru et al., 2016). Unlike Random Forest, Gradient Boosting builds trees sequentially, with each successive tree correcting the residual errors of the previous iteration. This iterative learning process enables the model to capture complex nonlinear relationships, subtle variable

interactions, and heterogeneous borrower behavior that traditional statistical models cannot adequately represent.

- **Model 4 - Ensemble:** Estimated as the simple average of the predicted probabilities generated by the Random Forest and Gradient Boosting models. Ensemble learning was selected because combining predictions from complementary algorithms often improves generalization performance by reducing model variance while preserving predictive accuracy. Random Forest contributes stability through bagging, whereas Gradient Boosting contributes lower bias through sequential optimization. Averaging their predictions exploits this diversity and has been shown to produce more robust and reliable predictions than individual models in many classification applications.

Together, these four models represent progressively more sophisticated approaches to credit risk prediction, allowing us to assess whether the additional complexity of machine learning algorithms yields statistically and economically meaningful improvements over conventional credit scoring methodologies.

Evaluation Metrics

We report three complementary performance metrics. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measures rank-order discrimination between defaulted and non-defaulted loans across all classification thresholds, a threshold-independent measure widely used in credit scoring (Hand, 2009). The Kolmogorov-Smirnov (KS) statistic measures the maximum separation between the cumulative distributions of defaulted and non-defaulted loan predicted scores, the standard metric in banking regulatory practice. The Brier Score measures the mean squared error of predicted default probabilities, rewarding both discrimination and calibration. A threshold-optimal classification at the predicted probability maximizing the F1 score is also reported for the best-performing model.

Model Training Process and Performance Advantage Over Traditional Methods

Overview of the Training Pipeline

This section explains how each model is developed, the main steps involved in the training process, and the additional capabilities offered by Gradient Boosting relative to

traditional logistic regression. All models use the same time-based sample design. The training sample includes 444,716 loans originated between FY2010 and FY2020, of which 32,169 defaulted, producing a default rate of 7.23%. Model performance is then evaluated using a separate holdout sample of 44,056 loans originated between FY2021 and FY2024. This test sample contains 4,325 defaults and has a default rate of 9.82%.

The temporal split is intended to reflect a realistic lending environment in which a model is estimated using historical information and then applied to loans originated in later periods. Training the models on pre-2021 observations and evaluating them on subsequent loan vintages provides a more demanding test of out-of-sample performance, particularly because the holdout period includes the economic disruptions and changing credit conditions associated with and following the 2020 pandemic. This approach more closely resembles how a model developed at the end of FY2020 would have performed after deployment in FY2021. It also avoids the overly optimistic performance estimates that can arise from random train-test splits, which may place loans from the same economic period in both the training and test samples.

Step 1: Logistic Regression (The Traditional Benchmark)

Logistic regression remains the workhorse of conventional credit scoring. Model estimation proceeds in three stages. First, five predictor variables are selected through expert judgment: the natural logarithm of loan amount, initial interest rate, loan term in months, a binary collateral indicator, and the SBA guarantee ratio. Second, all predictors are standardized to zero mean and unit variance prior to estimation. Third, a single coefficient per predictor is estimated by maximizing the log-likelihood function across all 444,716 training observations.

The resulting scoring equation takes the canonical form:

$$P(\text{default}) = \frac{1}{1+e^{-z}}, z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

where $x_1 = \log(\text{loan amount})$, $x_2 = \text{initial interest rate}$, $x_3 = \text{loan term}$, $x_4 = \text{collateral indicator}$, and $x_5 = \text{SBA guarantee ratio}$.

This specification offers two well-known advantages: full interpretability of the estimated coefficients and straightforward compliance auditability. It imposes, however, two binding

structural constraints. All predictors enter the index z linearly and additively, precluding non-linear threshold effects and interaction terms. Additionally, predictive signal is confined to the five pre-selected variables; any information contained in the remaining loan characteristics is excluded by design.

On the held-out test set, the logistic regression baseline achieves an AUC-ROC of 0.7641 and a Kolmogorov–Smirnov statistic of 0.4241. While respectable for a parsimonious five-variable scorecard, these figures suggest that meaningful predictive information remains uncaptured, information that more flexible, data-driven approaches may be positioned to recover.

Step 2: Random Forest — Parallel Ensemble Learning

A Random Forest constructs an ensemble of decision trees, each of which partitions the data into progressively narrower subgroups using a sequence of binary if-then rules. Where logistic regression fits a single global equation, a Random Forest lets each tree independently discover local patterns, including non-linear thresholds and variable interactions, that a linear model cannot represent. Standard library, scikit-learn in Python is used for this analysis.

The model trains 200 trees simultaneously. Each tree is built on a *bootstrap resample* of the training data (a random sample drawn with replacement, equal in size to the original), ensuring that no two trees see the same observations. At every node split, each tree considers only a random subset of available features rather than all predictors, a deliberate injection of randomness that forces the trees to be meaningfully different from one another. Diversity among the trees is the mechanism that makes the ensemble more accurate than any individual tree.

As an illustrative rule, a single node in one tree might encode: *"If loan term < 36 months AND interest rate > 8.5% AND industry sector = Transportation, classify as high default risk."* Across 200 trees, thousands of such rules are learned, collectively capturing complex, context-dependent patterns in the data. Three parameter choices guard against overfitting. Maximum tree depth is capped at 10 splits, preventing any tree from memorizing training-set noise by growing arbitrarily deep. Each terminal leaf must contain at least 50 observations, ensuring that probability estimates at the endpoints rest on meaningful sample sizes rather than isolated cases. Class weights are set inversely proportional to class frequency to address the

pronounced imbalance in the outcome variable (7.2% defaults vs. 92.8% non-defaults); without this adjustment, a model that simply predicts "no default" for every loan would achieve 92.8% accuracy while providing no useful signal.

At prediction time, all 200 trees independently estimate the probability of default for each loan. The final prediction is the simple average of these 200 estimates. This averaging is the core statistical benefit of ensemble methods: because each tree's errors are at least partially uncorrelated with the others, they tend to cancel out in the aggregate, reducing prediction variance substantially relative to any single tree.

Five-fold cross-validation on the training data confirms that the ensemble is stable and not overfit: $\text{AUC-ROC} = 0.939 \pm 0.006$ across folds. On the temporal holdout, the more demanding and policy-relevant test, because the model must generalize to loans originated after the training window, the model achieves $\text{AUC-ROC} = 0.8864$ and $\text{KS} = 0.6750$, representing a meaningful improvement over the logistic regression baseline ($\text{AUC} = 0.7641$, $\text{KS} = 0.4241$).

Step 3: Gradient Boosting (Sequential Error Correction)

Gradient Boosting achieves the highest performance ($\text{AUC} = 0.9178$, $\text{KS} = 0.7326$) through a fundamentally different learning philosophy. Where Random Forest builds 200 trees in parallel and averages them, Gradient Boosting builds 200 trees sequentially, each new tree explicitly trained to correct the prediction errors of all prior trees combined.

The process unfolds as follows. Iteration 1 fits a single shallow decision tree (maximum depth = 4, meaning at most four sequential if-then splits) to the full training set. This first tree already achieves a test AUC of 0.8639 by capturing the dominant signals in loan term and interest rate. Iteration 2 computes the residual prediction errors from iteration 1, which loans are currently misclassified or miscalibrated? and, fits a new tree exclusively to those residuals rather than to the original outcome variable. Each subsequent tree is estimated to reduce the residual prediction error of the existing ensemble. The learning rate of 0.05 controls how aggressively each new tree's corrections are incorporated: low enough to prevent any single tree from dominating the ensemble, high enough to make steady progress across 200 iterations. The subsampling ratio of 0.80 means each tree is fit on a random 80% draw of training observations,

introducing diversity and further reducing overfitting beyond what tree depth constraints alone provide.

Table 2 tracks the learning curve at checkpoints throughout training. The largest single-window gain occurs in iterations 1–21 (+0.0364 AUC), as the model rapidly identifies non-linear threshold effects around loan term, loans under 24 months face dramatically elevated default rates that a linear model can only approximate. Gains diminish but remain consistent through iteration 100 as the model incorporates sector interactions, processing method risk differentials, and guarantee ratio effects. Convergence is reached near iteration 160–180, with no evidence of test-set overfitting throughout.

Table 2: Gradient Boosting Learning Curve — Test AUC-ROC at Staged Iteration Checkpoints.

Boosting Iteration	Test AUC-ROC	Gain vs. Prior	Economic Interpretation
1 (first tree)	0.8639	Baseline	Term and interest rate thresholds captured immediately
21	0.9003	+0.0364	Rapid: non-linear term/rate interactions refined
41	0.9044	+0.0041	Sector risk premiums begin contributing
61	0.9079	+0.0035	Processing method and revolver effects incorporated
81	0.9115	+0.0036	Additional predictive contribution of the guarantee ratio
101	0.9143	+0.0028	Fine-grained residual corrections on edge cases
141	0.9153	+0.0010	Marginal gains; model approaching stable solution
181	0.9161	+0.0008	Near-convergence; additional iterations yield minimal gain
200 (final)	0.9178	+0.0017	Complete, all iterations used

Side-by-Side Process and Performance Comparison

Table 3 summarizes the process differences and performance outcomes across the traditional and ML approaches, enabling direct comparison across all key dimensions.

Table 3. Traditional vs. Machine Learning Credit Scoring — Process and Performance Comparison

Dimension	Logistic Regression (Traditional)	Gradient Boosting (Machine Learning)
Input variables	5 manually selected structured fields	loan structure + sector dummies + program type
Feature interactions	Linear additive only; none captured	All non-linear interactions learned automatically from data
Class imbalance	Not addressed	Balanced class weighting + 80% stochastic subsampling
Model passes over data	Single estimation pass	200 sequential error-correcting iterations
Learning rate	N/A (one-shot estimation)	0.05 shrinkage (prevents any single tree from dominating)
Model interpretability	High (single equation, 5 coefficients)	Moderate (feature importances + partial dependence plots)
5-fold CV AUC (training)	0.848 ± 0.009	0.960 ± 0.003 (subsample estimate)
Test AUC-ROC	0.7641	0.9178 (+20.1%)
Test KS Statistic	0.4241	0.7326 (+72.7%)
Brier Score	0.0891	0.0495 (−44.5%)
Defaults caught at 10% flag rate	1,192 of 4,325 (27.6%)	2,982 of 4,325 (68.9%) — 2.5x higher yield
Est. additional losses avoided	Baseline	~\$273M over test window (~\$68M/year)

Risk Stratification: The Practical Benefit

The most tangible expression of the ML advantage is risk stratification, the ability to rank loans so that real defaults concentrate at the top of the predicted-risk distribution. Table 4 presents the Gradient Boosting decile table on the FY2021–FY2024 test set: 44,056 loans divided into ten equal groups ranked from lowest to highest predicted default probability.

Table 4. Gradient Boosting Risk Decile Analysis — SBA 7(a) Test Set FY2021–FY2024 (n = 44,056). Highest-risk decile highlighted.

Decile	Loans	Mean Pred. Score	Defaults	Default Rate	Cumul. Default Capture
1 (Lowest)	4,421	0.35%	16	0.36%	0.4%
2	4,398	0.45%	52	1.18%	1.6%
3	4,398	0.61%	56	1.27%	2.9%

Decile	Loans	Mean Pred. Score	Defaults	Default Rate	Cumul. Default Capture
4	4,406	0.74%	50	1.13%	4.1%
5	4,405	0.86%	42	0.95%	5.0%
6	4,406	1.07%	121	2.75%	7.8%
7	4,406	1.66%	171	3.88%	11.7%
8	4,405	2.44%	211	4.79%	16.6%
9	4,405	8.33%	624	14.17%	31.0%
10 (Highest)	4,406	52.50%	2,982	67.68%	100.0%

The highest-risk decile (Decile 10) contains only 10% of all loans but captures 2,982 of 4,325 actual defaults, 68.9% of all defaults, at a realized default rate of 67.68%. The lowest five deciles combined (50% of all loans) contain only 216 defaults (5.0% of total), with individual default rates between 0.36% and 0.95%. By flagging only the top decile for enhanced underwriting scrutiny or denial, a lender intercepts 68.9% of eventual defaults while disturbing only 10% of the loan population. The equivalent logistic regression screen at the same 10% flag rate catches only 1,192 defaults (27.6%), less than half the ML model's yield. The difference in practical terms is 1,790 additional defaults identified per test window: at an average SBA guaranteed loss of \$152,666 per charged-off loan, this represents in avoidable federal guaranteed losses over four years, or \$68 million per year on this cohort alone.

This figure should be read as an illustrative upper bound rather than a precise forecast. It assumes flagged loans would have been declined or restructured without offsetting increases in false-positive declines of otherwise creditworthy borrowers, holds lender and borrower behavior fixed in response to the model, and does not net out guarantee fee revenue or the cost to small-business capital access of tightening approvals in the flagged decile. Realized savings from any operational deployment would depend on the specific underwriting action taken on flagged loans and would warrant a separate cost-benefit analysis.

Why Machine Learning Outperforms: Five Structural Reasons

- Non-linearity: Default risk is non-linear in loan term; there is a sharp threshold around 24–36 months below which default rates spike dramatically. Gradient Boosting captures this

through its tree splits; logistic regression forces a linear approximation that systematically underestimates short-term loan risk.

- **Interaction effects:** The joint occurrence of short loan term, high interest rate, and high-risk sector is far more dangerous than any factor individually would predict. Gradient Boosting learns these interactions automatically; logistic regression requires manual construction of interaction terms, which introduces researcher discretion and is practically infeasible across 18 sectors.
- **Class imbalance handling:** With 7.47% defaults, standard logistic regression systematically underestimates default probabilities for the minority class. Balanced class weighting and stochastic subsampling in Gradient Boosting correct for this, yielding better calibrated probability estimates and higher sensitivity.
- **High-dimensional categorical predictors:** Eighteen NAICS sector dummies and multiple processing method categories provide rich differentiated risk signals. ML models exploit these through data-driven node splitting; logistic regression requires the analyst to decide which sectors to parameterize separately—decisions that invariably introduce specification error.
- **Temporal generalization:** Trained on FY2010–FY2020 and tested on FY2021–FY2024, a period with a materially higher base default rate (9.82% vs. 7.23%), Gradient Boosting maintains $AUC = 0.9178$. This stability reflects its ability to learn structural relationships between loan characteristics and default outcomes, not merely period-specific correlations in the training data.

Results

Model Performance

Table 5 presents the out-of-time test set performance of all four models, alongside 5-fold cross-validation AUC estimates for the Logistic Regression and Random Forest specifications.

Table 5: Model Performance Comparison, Out-of-Time Test Set (FY2021–FY2024, $n = 44,056$).

Model	Feature Set	AUC-ROC	KS Statistic	Brier Score	CV AUC (5-fold)
Logistic Regression	Traditional (5 vars)	0.7641	0.4241	0.0891	0.848 ± 0.009

Model	Feature Set	AUC-ROC	KS Statistic	Brier Score	CV AUC (5-fold)
Random Forest	Full (19 vars + sectors)	0.8864	0.6750	0.1088	0.939 ± 0.006
Gradient Boosting	Full (19 vars + sectors)	0.9178	0.7326	0.0495	—
Ensemble (RF + GB avg)	Full (19 vars + sectors)	0.9100	0.6913	0.0616	—

Gradient Boosting delivers the strongest out-of-time discrimination, with an AUC-ROC of 0.9178 and a KS statistic of 0.7326, both substantially above the thresholds considered "excellent" in credit risk practice (AUC > 0.80, KS > 0.40; Siddiqi, 2006). The traditional Logistic Regression benchmark, despite achieving reasonable discrimination (AUC = 0.7641, KS = 0.4241), is outperformed on every metric. The absolute AUC improvement from Logistic Regression to Gradient Boosting is 0.1537 (20.1% relative improvement). The Brier Score improvement is even more pronounced: Gradient Boosting reduces the mean squared probability error by 44.5% relative to Logistic Regression (0.0495 vs. 0.0891), reflecting both superior discrimination and better probability calibration.

The 5-fold cross-validation results confirm that the performance advantage of ML models is not an artifact of overfitting to the training data. Random Forest achieves a training-set CV AUC of 0.939 ± 0.006, compared to 0.848 ± 0.009 for Logistic Regression—stable across folds and consistent with the holdout results.

At the optimal classification threshold for Gradient Boosting ($p^* = 0.1729$, selected to maximize the F1 score), the model correctly identifies 72.9% of defaulted loans (sensitivity) while maintaining a specificity of 95.8%. Precision at this threshold is 65.6%, and the F1 score is 0.6905. This threshold identifies a high-risk population constituting 11.4% of the test loans that accounts for 72.9% of actual defaults—a powerful screening capability for lender risk management.

Feature Importance

Table 6 presents the Gradient Boosting feature importance scores alongside economic interpretations. Loan term dominates all other predictors with an importance score of 0.8177, capturing the substantially elevated default risk of shorter-term loans—likely because shorter terms signal emergency or distressed borrowing contexts, and because default events are

mechanically more likely to occur within the observed window for short-tenor instruments. Initial interest rate (0.0820) is the second most important feature, reflecting both systematic credit risk pricing and macroeconomic interest rate environment effects. The SBA guarantee ratio (0.0352) shows a positive risk association—higher guarantee coverage may attenuate lender screening intensity due to reduced skin in the game, consistent with moral hazard theory in guaranteed lending.

Table 6. Gradient Boosting Feature Importance Scores (FY2010–FY2020 Training Set). Note: remaining importance distributed across sector and program-type indicators

Feature	Importance	Economic Interpretation
Loan Term (months)	0.8177	Longer-term loans face greater exposure duration; strong default predictor
Initial Interest Rate	0.0820	Higher rates signal riskier borrowers or tighter market conditions
SBA Guarantee Ratio	0.0352	Higher SBA coverage may reduce lender screening intensity
Fixed/Variable Rate	0.0219	Variable-rate loans face refinancing risk; affects payment stability
Revolver Status	0.0186	Revolving credit lines exhibit different draw-down and repayment patterns
Log Loan Amount	0.0165	Larger loans correlate with lower default rates (scale effect)
Has Collateral	0.0031	Collateral provides loss mitigation but limited as default predictor
Jobs Supported	0.0022	Larger employers show marginally lower default rates

Default Risk by Industry Sector

Table 7 characterizes default rates across the 15 major NAICS sectors represented in the sample. Substantial heterogeneity is evident: the Transportation sector (9.84% default rate) and Arts & Entertainment (9.70%) carry the highest systematic default risk, while Agriculture (4.09%) and Finance & Insurance (4.60%) exhibit the lowest. The Accommodation & Food Services sector, comprising restaurants and hotels, has a default rate of 8.01% on 61,634 loans, representing the single largest contributor to absolute default volume (4,939 defaults). Construction (8.20% on 54,320 loans) and Retail Trade (8.30% on 65,705 loans) also exhibit elevated risk across large loan volumes.

Table 7. Default Rates by NAICS Sector; SBA 7(a) Resolved Loans FY2010–FY2024. Table 7 excludes 8,021 loans (1.6% of the analytic sample) with missing or unclassified NAICS sector codes. Their exclusion does not materially affect sector-level default rate estimates.

NAICS Sector	Total Loans	Defaults	Default Rate
Transportation	28,588	2,813	9.84%
Arts & Entertainment	13,955	1,354	9.70%
Wholesale Trade	26,126	2,225	8.52%
Retail Trade	65,705	5,451	8.30%
Education Services	6,596	545	8.26%
Construction	54,320	4,454	8.20%
Accommodation & Food	61,634	4,939	8.01%
Admin & Support Svcs	23,553	1,813	7.70%
Other Services	42,101	3,220	7.65%
Professional Services	48,601	3,304	6.80%
Manufacturing	38,083	2,351	6.17%
Real Estate	10,901	625	5.73%
Healthcare & Social	43,884	2,110	4.81%
Finance & Insurance	8,776	404	4.60%
Agriculture	7,928	324	4.09%
ALL SECTORS	488,772	36,494	7.47%

Healthcare & Social Assistance (4.81% default rate on 43,884 loans) and Professional Services (6.80% on 48,601 loans) represent lower-risk sectors with high loan volume, profiles attractive to lenders seeking SBA-guaranteed exposure with below-average portfolio default risk.

Default Risk by Loan Size

A pronounced inverse relationship between loan size and default probability is observed. Loans of \$50,000 or less exhibit an 8.77% default rate compared to 3.74% for loans exceeding \$1 million. This size effect likely reflects multiple simultaneous phenomena: larger loans are extended to more established, creditworthy borrowers; lender due diligence is more intensive for large exposures; and larger borrowers have greater operational buffers to weather temporary distress. The loan size gradient persists across the full study period and is robust to controlling for sector effects in the multivariate models.

Temporal Dynamics

Annual default rates exhibit meaningful cyclical and structural variation. The post-financial crisis recovery period (FY2011–FY2013) shows the lowest default rates in the sample, declining from 7.50% (FY2010) to 5.48% (FY2013) as economic conditions stabilized. Default rates then rose through the late business cycle (FY2015–FY2019), reaching 9.88% in FY2018. The pandemic shock year (FY2020) paradoxically produced a lower default rate (5.95%) in the resolved loan cohort, likely because government stimulus programs (PPP, EIDL) enabled many at-risk borrowers to remain current. Default rates for FY2022 (12.19%) and FY2023 (16.27%) reflect both post-stimulus normalization and the challenging rate environment of 2022–2023, though these cohorts have shorter resolution windows than earlier vintage years.

Discussion

The findings of this study have several implications for SBA lending policy, lender risk management, and the broader discussion surrounding the use of machine learning in credit underwriting.

First, the improvement in predictive performance associated with the machine learning models appears to be economically meaningful in addition to being statistically relevant. Gradient Boosting improves AUC by 0.1537 relative to the logistic regression benchmark, indicating a substantially greater ability to distinguish between loans that ultimately default and those that do not. In practical terms, improved discrimination can allow lenders to separate higher- and lower-risk borrowers more effectively at a given decision threshold. For participating lenders, more accurate risk estimates may support underwriting, portfolio monitoring, pricing, and loss provisioning. From the SBA's perspective, improved identification of credit risk could also contribute to more effective management of the government's guarantee exposure. The 44.5% reduction in Brier Score further indicates that the Gradient Boosting model produces probability forecasts that are substantially more accurate, on average, than those generated by the benchmark model. Such improvements are relevant when predicted probabilities are used for portfolio-level risk measurement and expected-loss estimation.

Second, loan term emerges as the most influential predictor in the Gradient Boosting model, with an importance weight of 0.8177. This result suggests that maturity contains substantial information relevant to distinguishing between higher- and lower-risk loans within

the sample. The finding should not, however, be interpreted as evidence that loan term itself causes default. Loan maturity may reflect other characteristics of the borrower, financing purpose, loan structure, or underwriting decision that are also related to credit risk. Nevertheless, the strength of this variable suggests that the relationship between loan maturity and subsequent performance deserves greater attention in both underwriting practice and future research.

Third, the analysis reveals substantial differences in default experience across industry sectors. For example, borrowers in Transportation and Arts and Entertainment exhibit default rates approximately 2.4 times those observed among Agriculture borrowers. These differences suggest that industry conditions contain information relevant to SBA credit risk and may warrant consideration in portfolio monitoring and risk assessment. They also raise a broader policy question regarding whether the structure of SBA guarantee fees adequately reflects differences in underlying portfolio risk. The current findings alone do not establish that sector-based guarantee pricing would improve program outcomes, but they provide a basis for examining whether greater risk differentiation could contribute to the long-term financial sustainability of the guarantee program.

Any consideration of greater risk differentiation must also recognize the broader mission of the SBA 7(a) program. The program is intended to expand access to capital for small businesses that may encounter difficulty obtaining financing through conventional channels. Policies that increase the cost of guarantees for sectors exhibiting historically higher default rates could therefore have unintended consequences for credit availability. Industry-level risk also does not imply that every borrower within a higher-risk sector represents a poor credit risk. Any movement toward more risk-sensitive pricing would consequently need to balance portfolio risk management with the program's access-to-capital objectives and should be evaluated using evidence on borrower-level outcomes rather than sector averages alone.

Fourth, variation in default rates over time highlights the importance of model stability and out-of-time validation. Models estimated during periods of relatively low defaults may perform differently when applied under changing economic and credit conditions. The empirical design of this study addresses this issue by training the models on loans originated during FY2010–FY2020 and evaluating them on a subsequent FY2021–FY2024 holdout sample. The

default rate increases from 7.23% in the training period to 9.82% in the test period, providing a meaningful test of model performance under a different risk environment. Despite this shift, Gradient Boosting achieves an out-of-sample AUC of 0.9178, indicating strong discriminatory performance in the later-period sample. This result provides evidence that the model retains substantial predictive ability across the temporal split used in this study, although performance under future economic conditions cannot be inferred from this test alone.

Several limitations should be considered when interpreting these findings. First, the SBA 7(a) FOIA disclosure data used in this study do not provide the demographic information necessary to conduct a direct borrower-level analysis of differences in model outcomes across demographic groups. Consequently, the study cannot evaluate whether the predictive models produce systematically different outcomes across such groups. Future research could potentially combine SBA loan information with appropriately aggregated geographic and socioeconomic data to examine whether model performance or credit outcomes vary across communities, while recognizing that area-level characteristics should not be treated as substitutes for individual borrower characteristics.

Second, the publicly available data do not contain several borrower-level financial measures that would ordinarily be available to lenders during underwriting, including detailed revenue, debt-service coverage, and credit information. The models developed here therefore operate with a more limited information set than would typically be available to a participating lender. The results should accordingly be interpreted as evidence of the predictive information contained in the SBA disclosure variables rather than as an estimate of the maximum predictive performance achievable in an operational underwriting environment.

Third, an important limitation concerns the comparability of the logistic regression benchmark with the machine learning models. The logistic regression specification uses five expert-selected explanatory variables, whereas Random Forest, Gradient Boosting, and the ensemble model use the full 31-variable feature set, including industry and program-related indicators. The observed performance difference therefore reflects two changes simultaneously: greater model flexibility and a richer information set. It cannot be attributed exclusively to the machine learning algorithms themselves. Although this specification reflects the distinction

between a relatively parsimonious traditional scorecard and a more feature-rich predictive model, a more direct methodological comparison would estimate logistic regression using the same 31-variable information set. Such an analysis would help distinguish the incremental contribution of model flexibility from that of additional predictors and represents an important extension of the present study.

Finally, the analysis is based on loans with resolved outcomes, defined as Paid in Full or Charged Off. Because loans originated in more recent fiscal years have had less time to reach a terminal status, the resolved-loan sample may differ systematically from the broader population of SBA 7(a) originations, particularly toward the end of the study period. This seasoning and potential right-censoring issue should be considered when interpreting both the observed default rates and model performance for recent originations. Future research could address this limitation through survival-based methods or by allowing a longer performance window for more recent loan cohorts.

Conclusion

This study provides new empirical evidence on the use of machine learning for credit risk prediction within the SBA 7(a) guaranteed lending program. Using 488,772 resolved loans originated between FY2010 and FY2024, the analysis compares a traditional logistic regression benchmark with Random Forest, Gradient Boosting, and an ensemble model. Among the models evaluated, Gradient Boosting produces the strongest out-of-time predictive performance, achieving an AUC-ROC of 0.9178 compared with 0.7641 for the logistic regression benchmark, representing a 20.1% relative improvement. The model also reduces the Brier Score by 44.5%, indicating substantially more accurate probability forecasts in the holdout sample. These results demonstrate the potential value of flexible, data-driven methods for identifying differences in default risk within SBA lending.

The findings also have practical implications for the administration and management of the 7(a) program. For SBA program managers, improved risk stratification could complement existing approaches to portfolio monitoring, lender oversight, and the evaluation of guarantee exposure. Participating lenders may similarly benefit from incorporating machine learning tools

into existing credit risk processes, particularly for portfolio monitoring, loss estimation, and concentration management. These models should be viewed as decision-support tools rather than substitutes for established underwriting practices, lender judgment, or regulatory requirements. Their practical value will ultimately depend not only on predictive accuracy but also on interpretability, model governance, data quality, and continued validation as lending and economic conditions change.

The results are also relevant to the growing use of AI-assisted underwriting in financial services. Previous studies have documented the predictive value of machine learning and alternative data in consumer and other credit markets (Berg et al., 2020; Bartlett et al., 2022). The present analysis extends this discussion to government-guaranteed small business lending and shows that machine learning models can retain strong predictive performance when evaluated on a later, out-of-time sample. At the same time, the observed performance difference should be interpreted carefully because the machine learning models use a broader set of predictors than the parsimonious logistic regression benchmark. The findings therefore demonstrate the predictive value of the modeling framework used in this study but do not isolate the improvement attributable solely to the choice of algorithm.

More broadly, the study illustrates both the opportunities and the limitations associated with applying machine learning to public-sector credit programs. Predictive performance is an important consideration, but it represents only one dimension of responsible model use. Transparency, fairness, model stability, access to credit, and consistency with the public mission of the SBA must also be considered when evaluating how such methods might be incorporated into lending decisions. As AI-assisted credit assessment becomes more common, continued research will be needed to determine how improvements in prediction can be translated into better lending and policy decisions without compromising these broader objectives.

References

- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4), 589–609.
<https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>

- Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 143(1), 30–56.
<https://doi.org/10.1016/j.jfineco.2021.05.047>
- Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs—Credit scoring using digital footprints. *Review of Financial Studies*, 33(7), 2845–2897.
<https://doi.org/10.1093/rfs/hhz099>
- Berger, A. N., & Udell, G. F. (2006). A more complete conceptual framework for SME finance. *Journal of Banking & Finance*, 30(11), 2945–2966.
<https://doi.org/10.1016/j.jbankfin.2006.05.008>
- Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance*, 72, 218–239.
<https://doi.org/10.1016/j.jbankfin.2016.07.015>
- Cole, R. A., & Dietrich, A. (2012). SME loan rejections in the U.S.: Evidence from the 2003 survey of small business finances. Midwest Finance Association Annual Meeting Paper.
- Craig, B. R., Jackson, W. E., & Thomson, J. B. (2008). Credit market failure intervention: Do government sponsored small business credit programs enrich poorer areas? *Small Business Economics*, 30(4), 345–360. <https://doi.org/10.1007/s11187-007-9050-4>
- Frame, W. S., Srinivasan, A., & Woosley, L. (2001). The effect of credit scoring on small-business lending. *Journal of Money, Credit and Banking*, 33(3), 813–825.
<https://doi.org/10.2307/2673896>
- Glennon, D., & Nigro, P. (2005). Measuring the default risk of small business loans: A survival analysis approach. *Journal of Money, Credit and Banking*, 37(5), 923–947.
<https://doi.org/10.1353/mcb.2005.0051>
- Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), 103–123.
<https://doi.org/10.1007/s10994-009-5119-5>
- Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management*, 48(4), 1009–1029. <https://doi.org/10.1111/fima.12295>

- Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press.
- Kalnins, A., & Williams, M. (2014). When do female-owned businesses out-survive male-owned businesses? A disaggregated approach by industry and geography. *Journal of Business Venturing*, 29(6), 822–835. <https://doi.org/10.1016/j.jbusvent.2013.12.001>
- Lessmann, S., Baesens, B., Seow, H. V., & 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>
- Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications*, 161, 113694. <https://doi.org/10.1016/j.eswa.2020.113694>
- Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109–131. <https://doi.org/10.2307/2490395>
- SBA Office of Advocacy. (2023). *Small business facts*. U.S. Small Business Administration. <https://advocacy.sba.gov>
- Siddiqi, N. (2006). *Credit risk scorecards: Developing and implementing intelligent credit scoring*. Wiley.
- U.S. Small Business Administration. (2026). FOIA 7(a) loan data [Data set]. <https://data.sba.gov>