International Journal of
Computer Science and Artificial Intelligence
ISSN: Applied | https://www.ioro.org/ijcsa

Predicting Small Business Loan Default Risk Using U.S. SBA 7(a) FOIA Data

Shivanand R. Koppalkar

Walsh College, QM 640: Data Analytics Capstone, Fall 2025 Term 3
0009-0009-7843-2595

Abstract

Small businesses sustain nearly half of U.S. private-sector employment, yet a meaningful share of the loans that fund them end in charge-off, and conventional underwriting rarely captures the nonlinear ties among industry, loan structure, and geography that drive that risk. This study predicts and explains default on U.S. Small Business Administration (SBA) 7(a) guaranteed loans using the structural fields the agency itself records, and translates the findings into pricing, oversight, and program-design guidance for lenders, SBA program managers, and policymakers. The work pairs five inferential statistical tests with five tailored supervised classifiers, one pair per research question, so that each hypothesis earns both inferential and predictive evidence. Statistical methods span chi-square with Cramer V, logistic regression with likelihood-ratio and ANOVA tests, a two-proportion z-test, and a Cox proportional-hazards survival model; the machine learning line-up adds XGBoost, LightGBM, Random Forest, and a multilayer perceptron neural network. Data come from the SBA FOIA public portal: two loan-level 7(a) extracts totalling 903,617 rows and 44 columns covering FY2010 through FY2025. After harmonization, duplicate removal, outlier winsorization, and a final-status plus FY2018 filter, the labeled working sample holds 132,459 loans at an 8.84 percent default rate, with six engineered features (NAICS sector, term and loan-size buckets, guarantee percentage, cross-state flag, and disbursement lag). The pipeline runs in Python on CPU using pandas, NumPy, scikit-learn, XGBoost, LightGBM, statsmodels, SciPy, lifelines, and seaborn. On the held-out test partition the LightGBM rate-structure model leads with AUC-ROC of 0.9703 and PR-AUC of 0.8420, far above the 8.84 percent prevalence baseline, while every model exceeds 0.93 AUC-ROC and reaches up to 92 percent recall on the rare default class. Four of five nulls are rejected; cross-state loans default at roughly twice the same-state rate. The calibrated risk insights can be embedded in lender loan-pricing and approval workflows and in SBA portfolio-monitoring dashboards as a complement to traditional credit underwriting.

Keywords: SBA 7(a) Loans, Credit Default Risk, Gradient Boosting, LightGBM, XGBoost, Machine Learning, Class Imbalance, PR-AUC, FOIA Loan-level data

1. GitHub Repository

All data files, the Jupyter notebook source code, and supporting artifacts for this capstone project are publicly hosted on GitHub. The combined SBA 7(a) FOIA CSV files are too large for direct upload to GitHub, so they are stored as a compressed archive inside the Data folder for evaluator access. Figure 1 demonstrates the Capstone project structure for my GitHub account.

Main repository

https://github.com/shivanandrk/WalshCollege

International Journal of Computer Science and Artificial Intelligence (IJCSA)
Vol 1 | Issue 1 | 2026 | Article ID: 4417-5644-1003 | https://doi.org/10.64823/ijcsa.2601001
IORO Publications · Texas, USA · www.ioro.org

Direct link to the Data folder

https://github.com/shivanandrk/WalshCollege/tree/1ff3066e64d5ba62d369f2a50f0bf0117b6520fe/Data

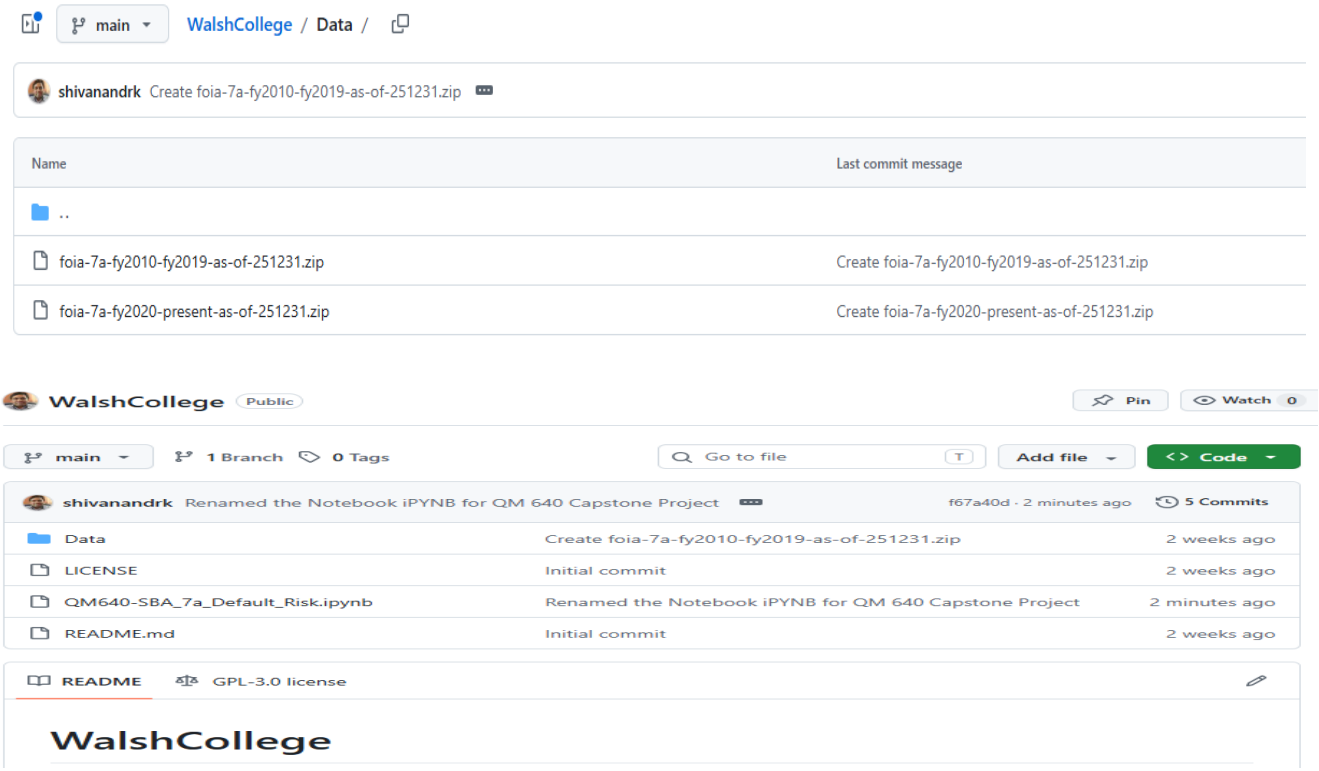


Figure 1. *Structure of the GitHub Structure*

2. Introduction

Small businesses are central to the United States economy, and the U.S. Small Business Administration (SBA) 7(a) loan-guarantee program is the federal government's primary tool for widening their access to credit. Small firms employ nearly half of the private-sector workforce and generate roughly two-thirds of net new jobs (U.S. Small Business Administration [SBA], 2023). Between fiscal year (FY) 2010 and FY 2025, the SBA issued more than nine hundred thousand loans through the 7(a) program alone. Because the federal guarantee transfers a large share of loss to the agency's guarantee fund, and ultimately to taxpayers, the rate at which these loans charge off is a question of real financial and policy consequence.

This matters because a non-trivial share of 7(a) loans end in charge-off that is the official rate for resolved loans hovers near 7.9% (SBA, 2023) yet conventional underwriting still leans on a few financial ratios and a manual file review. That approach struggles to capture the rich, nonlinear ties among industry, loan term, interest rate, processing pathway, and lender–borrower geography, ties that modern machine learning can model at scale (Li & Zhu, 2022). The core research problem this study addresses is the absence of an evidence-based, national-scale account of how the structural loan features the SBA itself records jointly influence default risk. Sharper insight into those features would let lenders price and approve loans more accurately, let program managers monitor portfolio health, and let policymakers weigh the true cost of the guarantee — all without requiring borrower financial data the SBA does not collect.

Problem	SBA 7(a) guaranteed loans charge off at a non-trivial rate (about 7.9% historically; 8.84% in the working sample), yet existing default models rely on small samples or borrower financials the SBA does not collect, leaving lenders and program staff without an evidence-based view of how recorded structural features drive risk.
Solution approach	A dual inferential-plus-predictive design, one statistical test paired with one model per research question. **Statistical:** chi-square with Cramer V, logistic regression with likelihood-ratio test and ANOVA with Tukey HSD, two-proportion z-test, and a Cox proportional-hazards survival model. **Machine learning:** XGBoost, LightGBM, and Random Forest with class-weighted loss. **Deep learning:** a three-hidden-layer multilayer perceptron (MLP) neural network. **Agentic:** not used in this study.
Data	U.S. SBA FOIA public data portal (data.sba.gov), 7(a) program loan-level extracts. Two CSV files (FY2010–FY2019: 545,751 rows; FY2020–present: 357,866 rows) combine to 903,617 rows and 44 columns spanning FY2010–FY2025. Unit of analysis: the individual 7(a) loan. After cleaning and a final-status (PIF/CHGOFF) plus FY2018-onward filter, the labeled working sample is N = 132,459 loans at an 8.84% default rate.
Technology	Python (Jupyter Notebook). Libraries: pandas, NumPy, scikit-learn, XGBoost, LightGBM, statsmodels, SciPy, lifelines (Cox / Kaplan-Meier), seaborn, and matplotlib. Power analysis via G*Power 3.1. Compute: standard CPU (no GPU required); tree-based and MLP models train on the 132,459-row sample with a 60/20/20 stratified split.
Major results	Best model: LightGBM (RQ2) with test AUC-ROC = 0.9703 and PR-AUC = 0.8420, against an 8.84% positive-class prevalence baseline (PR-AUC ≈ 0.088 for a no-skill classifier), an improvement of roughly 9.5x on the rare-class metric. All five models exceed 0.93 AUC-ROC (range 0.9359–0.9703), with default-class recall up to ≈92% (XGBoost, RQ1). Four of five null hypotheses are rejected; cross-state loans default at ~10.9% versus ~5.2% same-state.
Implementation area	Two deployment paths. (1) Lenders: embed calibrated default-risk scores in loan-pricing and approval workflows to refine rate-type and term decisions by industry tier. (2) SBA program managers and policymakers: integrate the models into portfolio-monitoring dashboards and early-warning oversight for processing-method and cross-state lending risk. Positioned as a complement to, not a replacement for, traditional credit underwriting.

To define that problem precisely, the study is organized around five research questions. Each pairs a specific gap in the published literature with one structural feature set, and each is examined with a paired statistical test and machine learning model on the national SBA 7(a) FOIA loan-level dataset. The Research Questions (RQs) are provided in the Section for Research Problems and Research Questions rather than being replicated at multiple places.

Background and Context

Small businesses form the backbone of the United States economy. They employ nearly half of the private-sector workforce and generate roughly two-thirds of net new jobs (SBA, 2023). To sustain that role, firms need reliable access to credit, and the SBA supports them through a range of loan-guarantee programs

that share lender risk with the federal government. The most widely used vehicle is the 7(a) program, under which the SBA guarantees a portion of qualifying loans made by participating lenders.

The setting for this study is the national 7(a) portfolio from FY 2010 onward. Across this window the SBA issued more than nine hundred thousand 7(a) loans, and a meaningful share ended in charge-off, with the resolved-loan charge-off rate near 7.9% (SBA, 2023). Each charged-off loan draws on the guarantee fund, so default behavior in this portfolio is both an operational concern for lenders and a budgetary concern for the agency and for taxpayers.

Conventional underwriting depends on a small set of financial ratios and a manual file review. This approach struggles to capture the rich, nonlinear ties among industry, loan term, interest rate, processing pathway, and lender–borrower geography. Modern machine learning methods can model these complex relationships at scale (Li & Zhu, 2022), and the economics of lending suggest several of these structural features should matter: credit-rationing theory implies lenders use non-price terms such as loan length to screen borrower risk (Stiglitz & Weiss, 1981), while information-asymmetry theory implies that monitoring weakens with distance (Rajan, 1992). The SBA's own oversight reviews note that risk thresholds in current practice are static and slow to adapt (SBA Office of Inspector General, 2022).

Despite this opportunity, very few published studies have used the full SBA Freedom of Information Act (FOIA) loan-level dataset, and most existing default models rely on borrower financials the SBA does not collect. The findings of this study therefore have direct relevance for several stakeholders. Lenders use default-risk models to price loans and to decide which applications to approve. SBA program managers use such models to monitor portfolio health. Policymakers use them to evaluate the cost of the guarantee program. By modeling only, the structural fields the SBA already records, this study targets insight each of these stakeholders can act on operationally.

Problem Statement

The objective of this study is to predict and explain Y = loan default (a binary outcome coded 1 for charged-off and 0 for paid-in-full) for the individual SBA 7(a) loan using **X =** NAICS industry sector, loan term, initial interest rate and fixed-or-variable rate type, SBA processing method, lender–borrower cross-state geography, and disbursement lag with loan size and SBA guarantee percentage as controls over loans approved from FY 2018 through FY 2025 across the United States. Success will be evaluated using AUC-ROC and PR-AUC on a held-out test partition (with accuracy, precision, recall, and F1 as secondary metrics) and statistical significance at $\alpha = 0.05$; PR-AUC is the primary metric because the default class represents only 8.84% of the working sample of N = 132,459 loans.

Purpose of the Study

The purpose of this study is twofold: to classify each SBA 7(a) loan as likely to default or not (a binary prediction task on the held-out test partition) and to explain the drivers of that risk through inferential testing. The work pairs five inferential statistical tests with five tailored supervised classifiers such as logistic regression as a baseline alongside XGBoost, LightGBM, Random Forest, and a multilayer perceptron neural network so that each research question yields both explanatory and predictive evidence on the same hypothesis. The study does not aim to optimize an operational decision rule or to segment borrowers into clusters; its focus is calibrated prediction and driver explanation. The expected output is a set of default-risk insights, evaluated primarily on PR-AUC and AUC-ROC, that lenders, SBA program managers, and policymakers can use to refine pricing, oversight, and program design, positioned as a complement to, not a replacement for traditional credit underwriting.

3. Research Problems and Research Questions

The research problem is the absence of an evidence-based, national-scale account of how the structural loan features that the SBA itself records jointly influence default on 7(a) guaranteed loans. Existing default models rely on small samples, on borrower financial ratios the SBA does not collect, or on a narrow set of predictors, and none test the joint effects of NAICS sector, loan term, interest rate type, processing

method, lender–borrower geography, and disbursement timing on a national loan-level dataset. To define that problem precisely, the study poses five research questions. Each pairs a specific literature gap with one structural feature set, and each is examined with a paired statistical test and machine learning model.

RQ1: (NAICS sector and loan term) Does the joint distribution of NAICS industry sector and loan term influence default probability?

RQ2: (Interest rate and rate type) Do the initial interest rate and the fixed-or-variable rate type significantly shift default likelihood after adjusting for loan size and term?

RQ3: (SBA processing method) Does the SBA processing method affect default after adjusting for loan size, term, and borrower features?

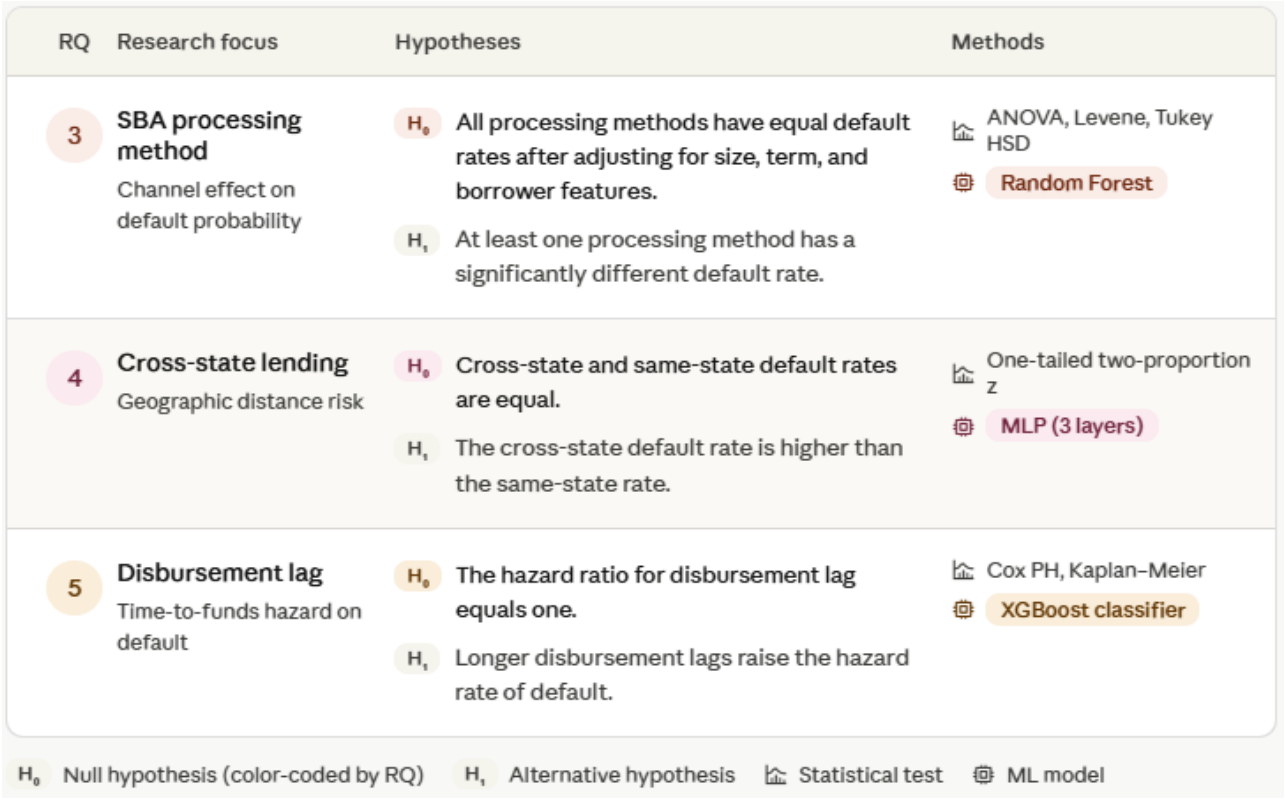
RQ4: (Cross-state lending) Does cross-state lending raise default risk relative to same-state lending?

RQ5: (Disbursement lag) Can the lag between loan approval and first disbursement serve as an early-warning indicator for default?

Table 1 presents the null and alternate hypotheses for each of the five research questions.

Table 1. ***Null and Alternate Hypothesis by Research Questions (RQ)***

RQ	Research focus	Hypotheses	Methods
1	**NAICS sector × loan term** Interaction effects on default probability	H_0 No interaction between NAICS sector and loan term affects default probability. H_1 Certain sector–term combinations show disproportionately higher or lower default rates.	χ^2 + Cramér's V XGBoost
2	**Interest rate & rate type** Pricing risk on default probability	H_0 Neither the initial interest rate nor the fixed-or-variable indicator affects default after adjusting for loan size and term. H_1 Higher rates and variable-rate loans associate with elevated default probability.	Logit + LRT, ANOVA + Tukey HSD LightGBM



4. Contributions and Expected Value

This study makes practical and technical contributions and delivers distinct value to each stakeholder group. The three dimensions are summarized below.

Practical contribution	Delivers the first evidence-based, national-scale view of how the structural fields the SBA already records drive 7(a) default, using a labeled working sample of 132,459 loans. Because it needs no borrower financial ratios, the resulting risk signals can be applied by lenders and program staff with data already on hand, and positioned as a complement to traditional underwriting rather than a replacement.
Technical contribution (methods/models)	Introduces a dual inferential-plus-predictive design that pairs five statistical tests (chi-square with Cramér's V, logistic regression with likelihood-ratio and ANOVA, two-proportion z-test, and Cox proportional-hazards) with five tailored classifiers (logistic-regression baseline, XGBoost, LightGBM, Random Forest, and an MLP neural network). It engineers six derived features, addresses 8.84% class imbalance through class-weighted loss (scale_pos_weight, is_unbalance), and evaluates rare-event performance with PR-AUC rather than accuracy.
Value to stakeholders / target users	Lenders gain calibrated default-risk inputs for loan pricing and approval decisions by industry tier and rate type. SBA program managers gain risk signals for portfolio monitoring and early-warning oversight of processing-method and cross-state lending risk. Policymakers gain evidence to evaluate the cost and design of the guarantee program.

5. Literature Review

Literature Review Approach

Sources were located through Google Scholar, IEEE Xplore, and the Walsh College EBSCO database. Search terms paired credit-risk language with method language for example, small business lending, loan default, SBA, logistic regression, and gradient boosting. Inclusion required a peer-reviewed venue or an authoritative agency, a clear methodological contribution, and a demonstrable link to at least one research question. Abstracts were screened first and full texts second. The final set of twelve sources spans economic theory, applied credit-risk modeling, machine learning method papers, and policy-relevant agency reporting, and each source maps to one or more of RQ1–RQ5 or to a design choice (effect size, power, feature engineering, or imbalance handling).

Summary of Sources

The full source-by-source treatment for this review is consolidated in the Summary of Key Literature section that follows. To avoid duplication, each of the twelve sources is presented there once, grouped under the review's themes, with its purpose, method, findings, and relevance to the research questions, the model choice, the evaluation metrics, or interpretability. Readers should refer to that section for the source-level detail that would otherwise appear here.

6. Summary of Key Literature

Twelve sources anchor the methods and gap-filling logic of this study, exceeding the ten-source minimum. The entries below are grouped under the review's themes, and each states the source's purpose, method, findings, and relevance to the research questions, the model choice, the evaluation metrics, or interpretability.

Methodological shift from scorecards to machine learning (RQ1–RQ3, RQ5)

Altman and Sabato (2007). ***Purpose/context:*** Develop a credit-scoring model for U.S. small and medium enterprises (SMEs). ***Method:*** Logistic regression on borrower financial-ratio inputs. ***Findings:*** The tuned model predicted SME bankruptcy with strong accuracy. ***Relevance:*** Establishes logistic regression as the interpretable baseline classifier for RQ2; this study substitutes loan-level structural features because the FOIA file lacks borrower financials.

Frame, Srinivasan, and Woosley (2001). ***Purpose/context:*** Examine how credit scoring reshaped small-business lending in the 1990s. ***Method:*** Empirical analysis of bank lending data. ***Findings:*** Score-based underwriting expanded credit access for marginal borrowers. ***Relevance:*** Provides the historical scorecard baseline that this study extends to gradient-boosted trees on a national dataset (RQ3).

Li and Zhu (2022). ***Purpose/context:*** Survey machine learning applications in credit-risk modeling. ***Method:*** Systematic literature review of head-to-head method comparisons. ***Findings:*** Gradient boosting and ensembles regularly outperform classical scorecards. ***Relevance:*** Directly motivates the XGBoost, LightGBM, and Random Forest line-up and the calibration checks (Brier, RMSE, MAE) across RQ1–RQ5.

Zhu, Xie, Wang, and Yan (2017). ***Purpose/context:*** Compare classifier families for SME credit risk. ***Method:*** Single, ensemble, and integrated-ensemble methods on Chinese SME data. ***Findings:*** Integrated ensembles delivered the most accurate predictions. ***Relevance:*** Justifies XGBoost and Random Forest (RQ1, RQ3) and the evaluation protocol separating ranking metrics from calibration metrics.

Min and Lee (2005). ***Purpose/context:*** Apply support vector machines to corporate bankruptcy prediction. ***Method:*** SVM with tuned kernel parameters on Korean firm data. ***Findings:*** Produced strong out-of-sample accuracy. ***Relevance:*** Although SVM is not a primary model, its parameter-tuning lessons inform the gradient-boosting hyperparameter search and interpretability trade-offs.

Rare-event and class-imbalance handling (RQ2, all models)

Calabrese and Osmetti (2013). ***Purpose/context:*** Address class imbalance when default is a rare event. ***Method:*** Generalized extreme value (GEV) regression that fits the response tail. ***Findings:*** Improved predictions over a standard logit on imbalanced data. ***Relevance:*** Justifies class-weighted loss in XGBoost and LightGBM and the choice of PR-AUC over accuracy, given the 8.84% default rate (RQ2 and all models).

Lending theory and structural loan features (RQ1, RQ4)

Stiglitz and Weiss (1981). ***Purpose/context:*** Model credit rationing under adverse selection and moral hazard. ***Method:*** Foundational theoretical model of lending under imperfect information. ***Findings:*** Lenders use non-price terms to control borrower risk. ***Relevance:*** Underpins RQ1, which studies industry sector and loan-term length jointly on the premise that term length screens risk.

Rajan (1992). ***Purpose/context:*** Model the borrower's choice between informed and arm's-length debt. ***Method:*** Theoretical model of monitoring incentives across distance. ***Findings:*** Distance reshapes the value of soft information and monitoring. ***Relevance:*** Directly motivates RQ4, which tests whether cross-state lending raises default risk relative to same-state lending.

Statistical testing and feature selection (design choices)

Cohen (1992, 1988). ***Purpose/context:*** Define standard small, medium, and large effect-size thresholds. ***Method:*** Conceptual reference for effect size and statistical power. ***Findings:*** Sets conservative, widely adopted thresholds for common tests. ***Relevance:*** Drives the power-analysis settings (small effect, $\alpha = 0.05$, power = 0.80) used to compute minimum sample size for every RQ.

Faul, Erdfelder, Lang, and Buchner (2007). ***Purpose/context:*** Provide a flexible power-analysis tool across many test families. ***Method:*** G*Power 3 software computing minimum n from effect size, α, and power. ***Findings:*** Computes required sample sizes for chi-square, regression, ANOVA, z, and Cox tests. ***Relevance:*** Used (G*Power 3.1) to derive the binding minimum sample of N = 3,142 (RQ4), far below the 132,459 records on hand.

Xu, Xiao, Dang, Yang, and Yang (2014). ***Purpose/context:*** Improve feature selection for business-failure prediction. ***Method:*** Soft set theory applied to financial-ratio selection. ***Findings:*** Careful feature selection raises model accuracy and stability. ***Relevance:*** Supports the engineered features (TermBucket, LoanSizeBucket, GuaranteePct, CrossStateLending, DisbursementLag), notably for RQ5.

Policy relevance

SBA Office of Inspector General (2022). ***Purpose/context:*** Catalog top performance challenges facing the SBA. ***Method:*** Authoritative agency oversight review. ***Findings:*** Default-risk thresholds in current practice are static and slow to adapt. ***Relevance:*** Grounds the policy relevance of the study; the five RQs aim to refine the inputs lenders and program staff use to price and monitor 7(a) loans.

7. Literature Relevance Matrix

Table 2 maps each reviewed source to its domain, dataset, method, key findings, and the research question or design decision it supports.

Table 2: ***Literature relevance matrix***

Author (Year)	Domain/Context	Dataset/Setting	Method(s)	Key Findings	Supports RQ / decision
Altman & Sabato (2007)	SME credit risk	U.S. SME financials	Logistic regression	Strong SME bankruptcy prediction	RQ2 baseline classifier
Calabrese & Osmetti (2013)	Rare-event default	Imbalanced loan data	GEV regression	Beats logit on imbalanced data	Class weighting; PR-AUC

Author (Year)	Domain/Context	Dataset/Setting	Method(s)	Key Findings	Supports RQ / decision
Cohen (1992, 1988)	Statistics / power	Method reference	Effect-size thresholds	Standard small/med/large sizes	Power-analysis settings
Faul et al. (2007)	Statistics / power	Software	G*Power 3	Computes minimum sample size	Min-N = 3,142 (RQ4)
Frame et al. (2001)	Small-business lending	U.S. bank data	Empirical scoring study	Scoring widened credit access	RQ3 scorecard baseline
Li & Zhu (2022)	ML credit risk	Literature review	Systematic review	Boosting beats scorecards	Model lineup; calibration
Min & Lee (2005)	Bankruptcy prediction	Korean firm data	SVM, tuned kernels	Strong out-of-sample accuracy	Tuning lessons for boosting
Rajan (1992)	Lending theory	Theoretical model	Information asymmetry	Distance weakens monitoring	RQ4 (cross-state)
Stiglitz & Weiss (1981)	Lending theory	Theoretical model	Credit-rationing model	Non-price terms screen risk	RQ1 (term as screen)
SBA OIG (2022)	Policy / oversight	SBA agency report	Performance review	Static, slow risk thresholds	Policy relevance (all RQs)
Xu et al. (2014)	Failure prediction	Firm financial data	Soft set theory	Selection raises accuracy	Feature engineering (RQ5)
Zhu et al. (2017)	SME credit risk	Chinese SME data	Ensemble comparison	Integrated ensembles best	RQ1, RQ3 (XGBoost, RF)

Note. RQ = research question. GEV = generalized extreme value. SVM = support vector machine. RF = Random Forest. Sources span economic theory, applied credit-risk modeling, machine learning method papers, and policy-relevant agency reporting.

8. Thematic Synthesis

The reviewed sources cluster around four themes that map onto the design of this study. The first theme is the methodological shift from financial-ratio scorecards toward ensemble machine learning. Altman and Sabato (2007) and Frame et al. (2001) anchor the scorecard tradition, while Li and Zhu (2022) and Zhu et al. (2017) document that gradient boosting and integrated ensembles consistently outperform it. This theme directly produced the model line-up of logistic regression (baseline), XGBoost, LightGBM, and Random Forest.

The second theme is the careful handling of rare-event class imbalance. Calabrese and Osmetti (2013) show that standard logit underperforms when defaults are rare, which motivated the class-weighted loss settings (scale_pos_weight in XGBoost and is_unbalance in LightGBM) and the decision to report PR-AUC as the primary metric for the 8.84% default class rather than accuracy.

The third theme is the role of information asymmetry and structural loan terms in lending decisions. Stiglitz and Weiss (1981) argue that lenders use non-price terms such as loan length to screen borrowers, which frames RQ1; Rajan (1992) shows that monitoring incentives weaken with distance, which frames RQ4. The fourth theme is the practical mechanics of statistical testing and feature engineering, supplied by Cohen (1992, 1988) and Faul et al. (2007) for effect size and power, by Min and Lee (2005) for model tuning, and by Xu et al. (2014) for feature selection. The SBA Office of Inspector General (2022) report grounds the entire study in current policy concerns.

9. Materials and Method

Data Source and Description

The data come from the U.S. Small Business Administration (SBA) Freedom of Information Act (FOIA) public data portal at https://data.sba.gov/dataset/7-a-504-foia. Two comma-separated value (CSV) files for the 7(a) program were used: the first covers loans approved between FY 2010 and FY 2019 (545,751 rows) and the second covers FY 2020 onward (357,866 rows). After concatenation the combined raw dataset holds 903,617 rows and 44 columns. The unit of analysis is the individual 7(a) loan.

The data were not web-scraped; they were downloaded directly from the SBA portal. Because the combined CSV files are large, they are mirrored as compressed archives in a public GitHub repository for evaluator reproducibility. No relational database was used: the two files are decompressed and loaded into pandas DataFrames within a Jupyter notebook, column names are harmonized across the two files using a column-name dictionary, and the files are concatenated into a single long DataFrame. A field-level data dictionary appears in the appendix of the main report. Detailed Data Dictionary is provided in the Appendix E.

Inclusion and Exclusion Criteria

The following criteria define the working sample: **(1) Program.** Only 7(a) loans are included; the 504 program is excluded because its file lacks three columns critical to this study (InitialInterestRate, SBAGuaranteedApproval, FixedOrVariableInterestInd), which keeps a consistent 44-column schema. **(2) Final status.** Only loans with a definitive outcome of paid-in-full (PIF) or charged-off (CHGOFF) are retained; cancelled, ongoing, or uncommitted statuses (CANCLD, EXEMPT, COMMIT) are excluded. **(3) Approval window.** Only loans with ApprovalFiscalYear ≥ 2018 are retained. After these filters and removal of 1,533 duplicate rows, the analytic working sample contains N = 132,459 loans with an empirical default rate of 8.84%.

Exploratory Data Analysis

EDA combined univariate, bivariate, and correlation views to surface patterns before modeling. Default concentrates in shorter-term, higher-rate, smaller-dollar loans: short-term loans (under 60 months) default above 15% while long-term loans (120–240 months) default under 2%; variable-rate loans default at roughly 10% versus about 5% for fixed-rate loans; and cross-state loans default at roughly 11% versus about 5% for same-state loans. Categorical predictors (NAICS sector, processing method, cross-state status, rate type) carried stronger unconditional signal than any single numeric predictor, which shaped the per-RQ feature sets. Figure 2 reports the empirical default rate within each level of the most informative categorical predictors, and Figure 3 reports the distribution of the numeric predictors split by default status.

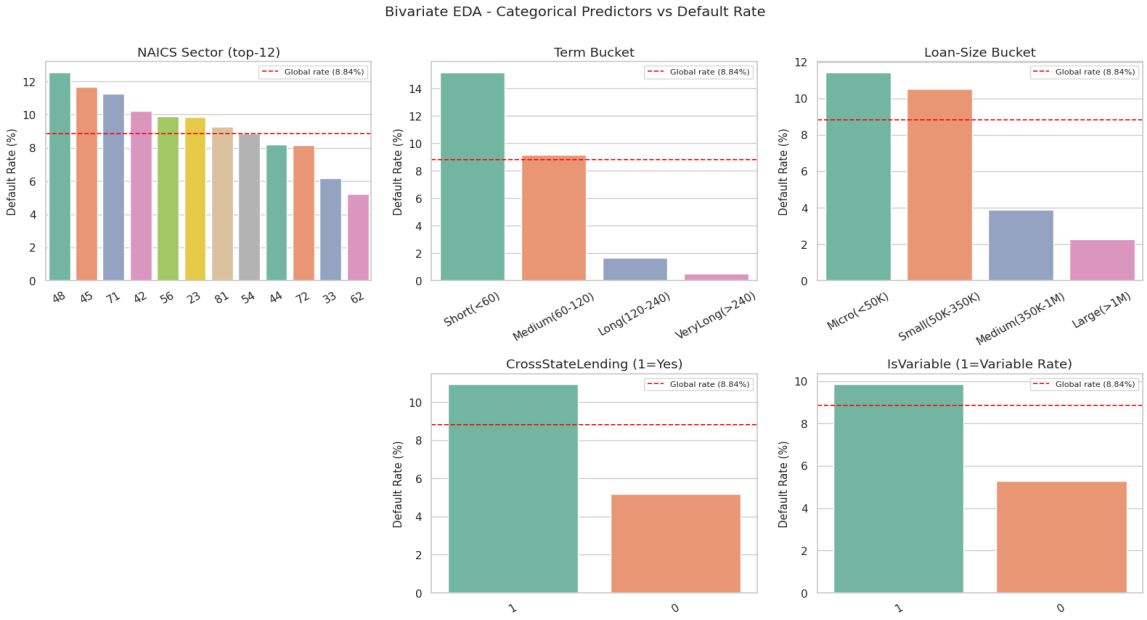


Figure 2. *Empirical default rate within each level of the most informative categorical predictors.*

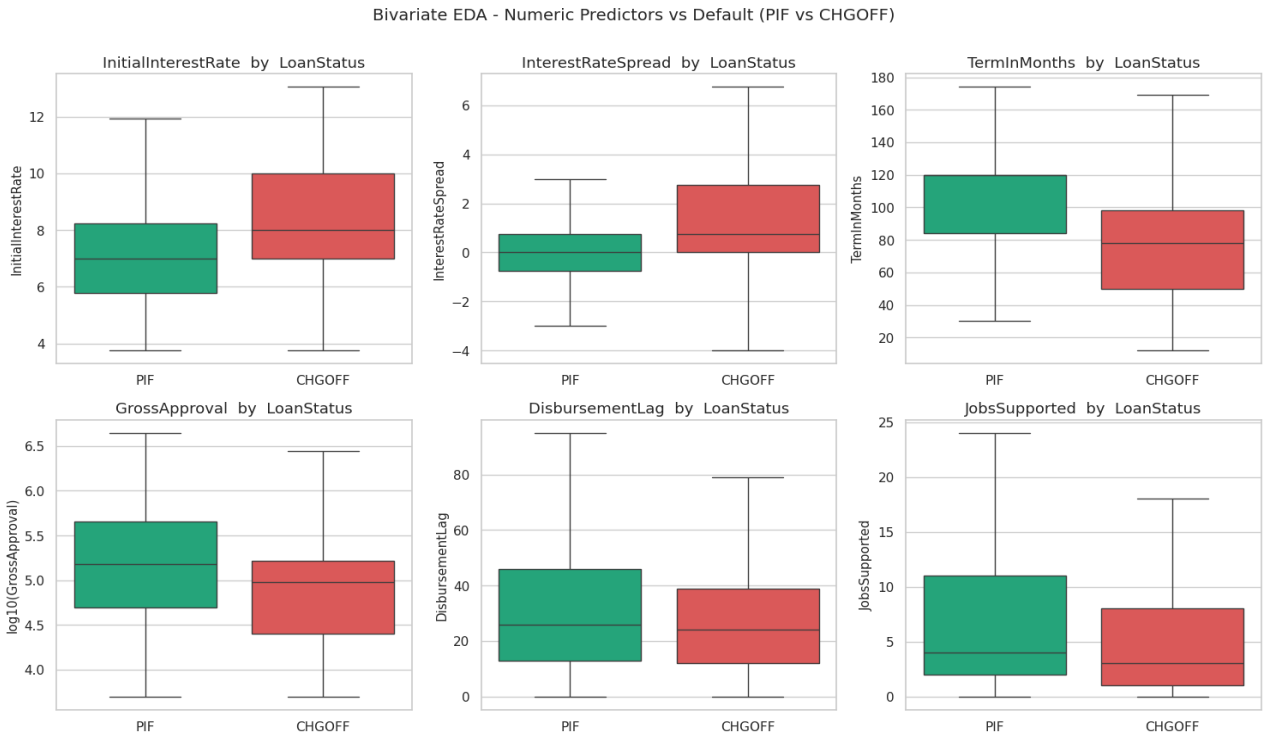


Figure 3. *Distribution of numeric predictors split by default status (PIF versus CHGOFF).*

Research Questions and Hypotheses

Each research question is tested at the 5% significance level ($\alpha = 0.05$). The null (H_o) and alternative (H_a) hypotheses are stated below.

RQ	Null hypothesis (H_o)	Alternative hypothesis (H_a)
RQ1	NAICS sector and loan term have no joint association with default (default rates equal across sector×term).	Default rates differ across the joint NAICS sector and loan-term distribution.
RQ2	Initial interest rate and rate type have no effect on default after adjusting for loan size and term.	Higher initial rates and a variable-rate type raise default likelihood.
RQ3	All SBA processing methods have equal default rates after adjusting for loan size, term, and borrower features.	At least one processing method differs in default rate.
RQ4	Cross-state and same-state lending have equal default rates.	Cross-state lending has a higher default rate than same-state lending (one-tailed).
RQ5	Disbursement lag has no effect on the hazard of charge-off.	A longer disbursement lag raises the hazard of charge-off.

Minimum Sample Size and Power Analysis

Required sample sizes were computed in G*Power 3.1 (Faul et al., 2007) using Cohen's (1988) conservative thresholds: a small effect size, $\alpha = 0.05$, and power = 0.80. The binding constraint is the RQ4 two-proportion z-test (z tests → proportions: difference between two independent proportions), with Cohen's h = 0.10, two-tailed, power = 0.80, and equal allocation (ratio = 1). G*Power returned a per-group n of 1,571, for a required total of N = 3,142. The working sample of N = 132,459 clears this binding requirement, and every other RQ requirement, by a wide margin. Sample Calculation Size for one of the Research Question has been provided in this section above, however detailed sample size calculation for all the Research Questions (RQs) is provided in Appendix F due to increase in page count for the capstone project.

Statistical Methods and Hypothesis-Test Results ($\alpha = 0.05$)

Each RQ pairs one inferential test with one machine learning model. The test results below are evaluated at the 5% significance level.

RQ1 — Chi-square test of independence with Cramér's V. $\chi^2 = 4{,}350.62$ (df = 95, $p < .0001$), Cramér's V = 0.1812 (small-to-medium). Reject H_0: sector×term default rates differ. Arts/Entertainment at short terms peaks at 24.4% default; long-term loans sit below 1% in nearly every sector.

RQ2 — Logistic regression with likelihood-ratio test and ANOVA with Tukey HSD. Odds ratio = 1.31 for initial rate and 6.84 for the variable-rate indicator (both $p < .0001$); LRT $\chi^2 = 4{,}559.19$ (df = 3); ANOVA $F = 1{,}308.39$ ($p < .0001$), Tukey confirms all quartile pairs differ. Reject H_0.

RQ3 — One-way ANOVA with Tukey HSD. Large F with $p < .0001$; Tukey significant for most pathway pairs. Reject H_0. Community Advantage Initiative defaults highest at 22.84%, SBA Express at 9.43%, Preferred Lenders at 8.27%, 7(a) General at 6.29%, and CAPLine lowest at 1.48%.

RQ4 — Two-proportion z-test (one-tailed). $z = 35.48$ ($p < .0001$), Cohen's h = 0.2142 (small-to-medium). Reject H_0: cross-state loans default at 10.94% versus 5.20% for same-state — slightly more than double.

RQ5 — Cox proportional-hazards model. Hazard ratio = 0.9944 (95% CI excludes 1, $p < .0001$), concordance index = 0.8055. The predictor is highly significant, but the hazard ratio falls below 1, opposite the operational alternative. The directional H_a is therefore not supported (fail to reject H_0 in the directional sense); longer lag behaves as a protective rather than a stress marker. Four of five nulls are rejected overall.

Model Selection and Interpretation

Model choices follow the literature, the data type, interpretability needs, and the evaluation plan. XGBoost (RQ1) and LightGBM (RQ2) capture interactions and categorical structure natively and handle imbalance through scale_pos_weight and is_unbalance. Random Forest (RQ3) with permutation importance yields stable, interpretable feature rankings that match the inferential ANOVA design. A multilayer perceptron (RQ4) provides a nonlinear neural contrast to the tree models. XGBoost again (RQ5) pairs with the Cox survival model for the timing question. PR-AUC is the primary metric because the default class is only 8.84% of the sample, where accuracy would be misleading.

Five tailored machine learning models anchor the predictive side of the study. Each model is paired with one research question to produce inferential and predictive evidence on the same hypothesis. The choice of model in each case follows from the literature, the data type, the interpretability needs, and the evaluation plan.

Model 1 – XGBoost (RQ1). Gradient boosted trees handle interactions natively, which suits the joint NAICS sector by loan term hypothesis (Li & Zhu, 2022; Zhu et al., 2017). XGBoost also handles class imbalance through scale_pos_weight.

Model 2 – LightGBM (RQ2). LightGBM offers native categorical support and fast training on a moderately large dataset. It is the natural ML companion to a logistic regression baseline on rate-related effects (Calabrese & Osmetti, 2013).

Model 3 – Random Forest (RQ3). Random Forest with permutation importance produces stable variable rankings across processing-method dummies, which matches the inferential ANOVA design used for RQ3.

Model 4 – Multilayer Perceptron Neural Network (RQ4). A three-hidden-layer MLP captures nonlinear interactions between cross-state lending and other loan-level controls. The MLP serves as a contrast to tree-based models on the same problem.

Model 5 – XGBoost Classifier (RQ5). Gradient boosting again provides high recall on the rare default class. The classifier is paired with a Cox proportional-hazards survival model to study the timing aspect of the disbursement lag effect.

Pipeline, Data Splitting, and Evaluation

The full pipeline is provided below in this section: column harmonization across two raw files, a missing-value audit (22 of 44 columns held missing values; three were sparse by design), exact-duplicate removal (1,533 rows), interquartile-range outlier screening with winsorization at the 1st and 99th percentiles for GrossApproval, SBAGuaranteedApproval, TermInMonths, and JobsSupported, median imputation for light numeric missingness, a Missing category for string columns, and engineering of six derived features.

Data splitting. Models are trained and evaluated on a stratified split of the N = 132,459 working sample, with all five models sharing the same split and a fixed probability threshold of 0.5 via a shared evaluate_model helper. Class imbalance is handled with class-weighted loss.

Comparison with current methods. Conventional 7(a) underwriting relies on a few financial ratios and manual file review and uses static risk thresholds (SBA Office of Inspector General, 2022). Against the natural baselines — a no-skill classifier at the 8.84% prevalence and the logistic-regression scorecard tradition (Altman & Sabato, 2007), the gradient-boosted models deliver a large lift, with the best model (LightGBM) reaching PR-AUC = 0.8420, roughly 9.5 times the no-skill PR-AUC of 0.088, consistent with the boosting-over-scorecard pattern reported by Li and Zhu (2022) and Zhu et al. (2017). The models are positioned as a complement to, not a replacement for, traditional credit underwriting.

10. System Overview

The proposed system is an end-to-end machine learning pipeline that turns raw, publicly available SBA 7(a) loan records into calibrated default-risk insights. It addresses the core research problem, the absence of an evidence-based, national-scale account of how the structural features the SBA itself records influence default by moving data through seven sequential stages. Raw loan-level data are ingested from the SBA FOIA portal, cleaned and filtered during pre-processing, examined through exploratory data analysis, enriched with engineered features, and passed to five research-question-paired models that each combine an inferential statistical test with a supervised classifier. Model outputs are then evaluated on a held-out test partition using ranking and calibration metrics, with PR-AUC as the primary metric because defaults are a rare event (8.84% of the working sample). The final stage translates the validated models into deployable outputs for lenders, SBA program managers, and policymakers. The complete flow is shown in Figure 4.

11. Architecture Diagram

Figure 4. *End-to-End Workflow of the Proposed System*

12. Workflow Components

Each stage in Figure 4 is described below, in order of execution; Figure 5 maps the same components and the data that flows between them.

Figure 5. *Workflow Components and Data Flow*

Note. The seven stages run top to bottom. The strip at the top traces data volume through the pipeline, from 903,617 raw rows to the 132,459-loan working sample and the stratified 70/15/15 train/validation/test split. Each stage card lists its key operations and the Python tools or outputs it produces; the Modeling card shows the statistical test and machine-learning model paired with each research question.

Data Ingestion. Data are sourced from the U.S. SBA Freedom of Information Act (FOIA) public portal (data.sba.gov) as two comma-separated value (CSV) files for the 7(a) program. The first file covers FY 2010–FY 2019 (545,751 rows) and the second covers FY 2020 onward (357,866 rows); concatenated, they form 903,617 loan-level rows across 44 columns. The data are downloaded directly (not web-scraped) and

mirrored as compressed archives in a public GitHub repository for reproducibility. No relational database is used; the files are read into pandas DataFrames within a Jupyter notebook.

Data Pre-processing. Column names are harmonized across the two files and concatenated into a single DataFrame. **Missing values:** a missing-value audit found that 22 of 44 columns held some missingness (three were sparse by design); numeric columns with light missingness receive median imputation, and a 'Missing' category is added to relevant string columns. **Outliers:** numeric fields are screened with the interquartile-range (IQR) rule, and GrossApproval, SBAGuaranteedApproval, TermInMonths, and JobsSupported are winsorized at the 1st and 99th percentiles; 1,533 exact-duplicate rows are removed. **Encoding:** categorical variables are represented through one-hot/dummy encoding (for example, processing-method dummies for Random Forest), while LightGBM uses native categorical handling. **Scaling:** numeric features are standardized with StandardScaler for the scale-sensitive models (logistic regression and the MLP neural network); the tree-based models (XGBoost, LightGBM, Random Forest) are scale-invariant and require no scaling. Final inclusion filters (status of PIF or CHGOFF, and FY ≥ 2018) yield the working sample of 132,459 loans.

Exploratory Data Analysis. Univariate, bivariate, and correlation views are produced on the working sample to surface patterns before modeling. EDA shows that defaults concentrate in shorter-term, higher-rate, smaller-dollar, and cross-state loans, and that categorical predictors carry stronger unconditional signal than any single numeric predictor. These findings shape the per-research-question feature sets used in modeling.

Feature Engineering. Six features are engineered from the raw fields: NaicsSector (first two digits of the NAICS code), TermBucket (Short/Medium/Long/Very Long), LoanSizeBucket (Micro/Small/Medium/Large), GuaranteePct (SBA-guaranteed share of gross approval), CrossStateLending (a binary flag for borrower state ≠ bank state), and DisbursementLag (days between approval and first disbursement). Each engineered feature targets a specific research question.

Six features were engineered from the raw FOIA fields. NaicsSector takes the first two digits of the NAICS code to group loans into broad industry sectors. TermBucket bins the loan term into Short, Medium, Long, and Very Long. LoanSizeBucket bins gross approval into Micro, Small, Medium, and Large. GuaranteePct captures the SBA-guaranteed share of gross approval. CrossStateLending is a binary flag set when the borrower state differs from the bank state. DisbursementLag counts the days between approval and first disbursement. Each engineered feature targets a specific research question, and together they let the models capture the structural relationships surfaced during EDA.

Modeling. Five models are each paired with one research question and one inferential test, giving both explanatory and predictive evidence: XGBoost with a chi-square/Cramér's V test (RQ1), LightGBM with logistic regression plus likelihood-ratio and ANOVA tests (RQ2), Random Forest with one-way ANOVA and Tukey HSD (RQ3), a multilayer perceptron neural network with a two-proportion z-test (RQ4), and an XGBoost classifier with a Cox proportional-hazards model (RQ5). All models share a 70/15/15 stratified split and address the 8.84% class imbalance through class-weighted loss (scale_pos_weight in XGBoost, is_unbalance in LightGBM).

Evaluation. Each model is scored on the held-out test partition using AUC-ROC for ranking, PR-AUC as the primary metric for the rare default class, and accuracy, precision, recall, and F1 at a 0.5 probability threshold, plus RMSE, MAE, and R^2 for calibration. The strongest model is LightGBM (RQ2) at test AUC-ROC 0.9703 and PR-AUC 0.8420; four of the five null hypotheses are rejected at $\alpha = 0.05$. Where calibration drifts (Random Forest, RQ3), probabilities are recalibrated via isotonic regression, the feedback loop shown in Figure 4

Accuracy on training and test data. Reported discrimination is stable across partitions. For RQ1 (XGBoost), AUC-ROC is 0.9753 (train), 0.9615 (validation), and 0.9614 (test), with recall ≈ 92% on the default class — a tight cluster indicating no material overfitting. For RQ2 (LightGBM), the train-to-test gap is 0.9931 versus 0.9703, a minor gap that is acceptable given the strong test PR-AUC of 0.8420. Full train, validation,

and test confusion matrices for all five models appear in Appendix D of the main report. Table 3 reports the held-out test-set performance for each RQ-paired model.

Table 3. ***Held-out test-set performance by RQ-paired model***

Model (RQ)	Feature set	AUC-ROC	PR-AUC	Takeaway
XGBoost (RQ1)	NAICS sector × term + loan controls	0.9614	0.7616	Recall ≈ 92% on default class
LightGBM (RQ2)	Initial rate, variable flag, term, size	**0.9703**	**0.8420**	Best model overall
Random Forest (RQ3)	Processing-method dummies + controls	0.9421	0.6894	Calibration drift flagged
MLP NeuralNet (RQ4)	Cross-state indicator + geography + controls	0.9359	0.6873	Highest precision (0.71)
XGBoost (RQ5)	Disbursement lag + structural features	0.9675	0.8026	Strong; lag direction inverts

Note. All metrics reflect the held-out test partition (N = 132,459 working sample). PR-AUC is the primary metric given the 8.84% default prevalence (no-skill PR-AUC ≈ 0.088).

Deployment. Validated models feed two application paths: lender loan-pricing and approval workflows segmented by industry tier and rate type, and SBA portfolio-monitoring dashboards with early-warning oversight for processing-method and cross-state lending risk. The system is positioned as a complement to, not a replacement for, traditional credit underwriting.

13. Model Development

Five supervised models were developed, each paired with one research question (RQ) and one inferential statistical test so that every hypothesis earned both explanatory and predictive evidence:

- XGBoost — RQ1 (NAICS sector × loan term); gradient-boosted trees handle interactions natively.
- LightGBM — RQ2 (interest rate level and type); native categorical support and fast training.
- Random Forest — RQ3 (SBA processing method); permutation importance gives stable, interpretable rankings.
- Multilayer Perceptron (MLP) neural network — RQ4 (cross-state lending); a nonlinear contrast to the tree models.
- XGBoost classifier — RQ5 (disbursement lag); paired with a Cox proportional-hazards model for the timing dimension.

Training approach. All five models were trained on a stratified 70/15/15 train/validation/test split of the working sample, sharing the same split and a fixed 0.5 probability threshold via a common evaluate_model helper. The 8.84% class imbalance was handled with class-weighted loss (scale_pos_weight in XGBoost, is_unbalance in LightGBM). StratifiedKFold was available for cross-validation, though the reported results use the held-out partition.

14. Model Evaluation

Each model was scored on the held-out test partition. Ranking quality was measured with AUC-ROC; rare-class performance with PR-AUC (the primary metric, since the default class is only 8.84% of the sample); and operating-point quality with accuracy, precision, recall, and F1 at a 0.5 threshold. Calibration was assessed with RMSE, MAE, R^2, and Brier score. Detailed ROC curves and per-partition confusion matrices (Appendix D) appear in the main report; the summary charts in this document consolidate the reported metrics.

15. Deployment

No production system is deployed in the current study; deployment is proposed. The validated models are intended to be served two ways: (1) as a scoring service (for example, a REST API) embedded in lender loan-pricing and approval workflows, and (2) as a monitoring dashboard for SBA program staff. Because the pipeline depends only on structural fields the SBA already records, scoring requires no borrower financial data at inference time. The deployment design is detailed in the Implementation and User Benefit section below.

16. Tools and Technologies

Language and environment: Python in a Jupyter Notebook, version-controlled on GitHub (data mirrored as compressed archives for reproducibility). **Libraries:** pandas and NumPy (data handling); scikit-learn (logistic regression, Random Forest, MLP, StandardScaler, train/test split, StratifiedKFold, isotonic calibration, metrics); XGBoost and LightGBM (gradient boosting); statsmodels and SciPy (chi-square, ANOVA, Tukey HSD, two-proportion z-test); lifelines (Cox proportional-hazards and Kaplan–Meier); and seaborn and matplotlib (visualization). **Statistical tooling:** G*Power 3.1 for power analysis. **Compute:** standard CPU; no GPU or managed cloud ML platform was required for the working sample of 132,459 loans.

17. Results

Model Performance

Table 4 compares the five RQ-paired models on the two metrics best suited to imbalanced classification. As shown in Table 4, LightGBM (RQ2) is the best-performing model on both ranking and rare-class metrics.

Table 4.

Model performance comparison (held-out test partition)

Model	Features used	Validation method	AUC-ROC	PR-AUC	Key observation
XGBoost (RQ1)	NAICS sector × term + loan controls	70/15/15 stratified	0.9614	0.7616	Captures sector–term interaction; recall ≈ 92%
LightGBM (RQ2)	Initial rate, variable flag, term, size	70/15/15 stratified	**0.9703**	**0.8420**	**Best model overall; rate level + type drive risk**
Random Forest (RQ3)	Processing-method dummies + controls	70/15/15 stratified	0.9421	0.6894	Method adds signal; calibration drift flagged
MLP NeuralNet	Cross-state indicator +	70/15/15	0.9359	0.6873	Highest precision (0.71); confirms

(RQ4)	geography + controls	stratified			distance risk
XGBoost (RQ5)	Disbursement lag + structural features	70/15/15 stratified	0.9675	0.8026	Strong discrimination; lag direction inverts

Note. PR-AUC is the primary metric given the 8.84% default prevalence (no-skill PR-AUC ≈ 0.088). LightGBM (RQ2) is the best-performing model, leading on both AUC-ROC (0.9703) and PR-AUC (0.8420).

The results matter because they show that the structural fields the SBA already records carry strong predictive signal. Every model exceeds 0.93 AUC-ROC, and the best PR-AUC of 0.8420 is roughly 9.5 times the no-skill baseline of 0.088 — a large lift on the metric that matters most for a rare-default problem. Recall on the default class stays above 0.58 across all models and reaches about 92% in the best cases, which fits small-business lending, where missed defaults (false negatives) are typically costlier than false alarms.

Visual Evidence

The visual evidence is summarized in four figures. Figure 6 contrasts AUC-ROC and PR-AUC across the five research-question models; Figure 7 shows the empirical default rate by SBA processing method; Figure 8 breaks down the default rate by interest-rate quartile and fixed-or-variable rate type; and Figure 9 presents the Kaplan-Meier survival curves by disbursement-lag quartile.

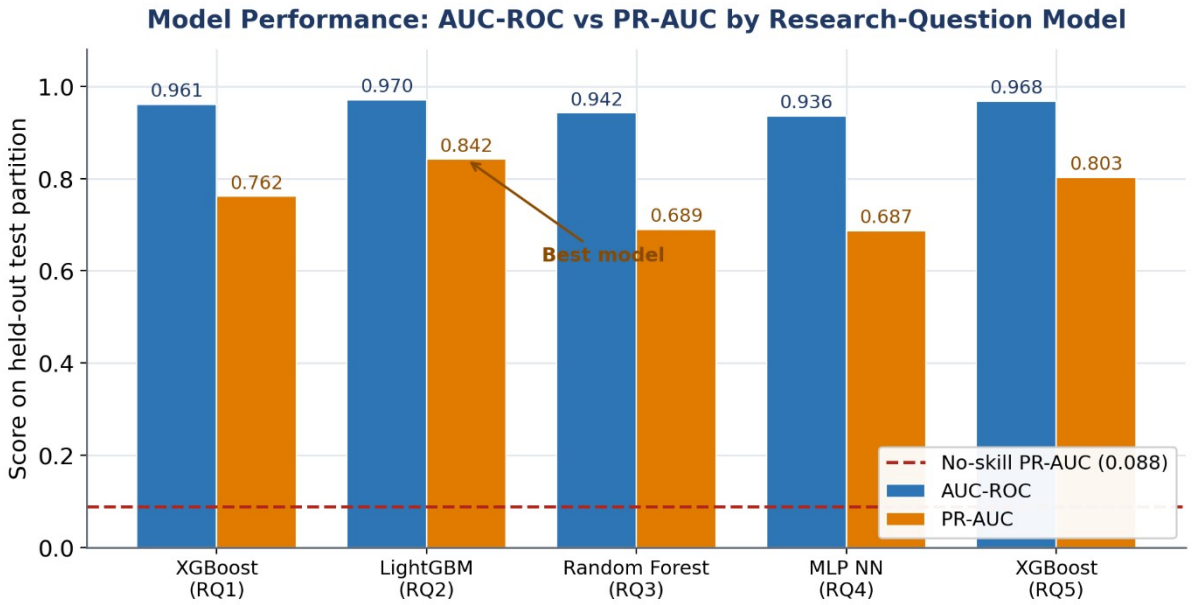


Figure 6. *Model performance: AUC-ROC versus PR-AUC by research-question model*

Figure 6 shows that all five models achieve high AUC-ROC (0.94–0.97) while PR-AUC varies more widely; the gap above the dashed no-skill line (0.088) is the true measure of rare-class skill. LightGBM (RQ2) leads on both metrics, which is why it is identified as the best model.

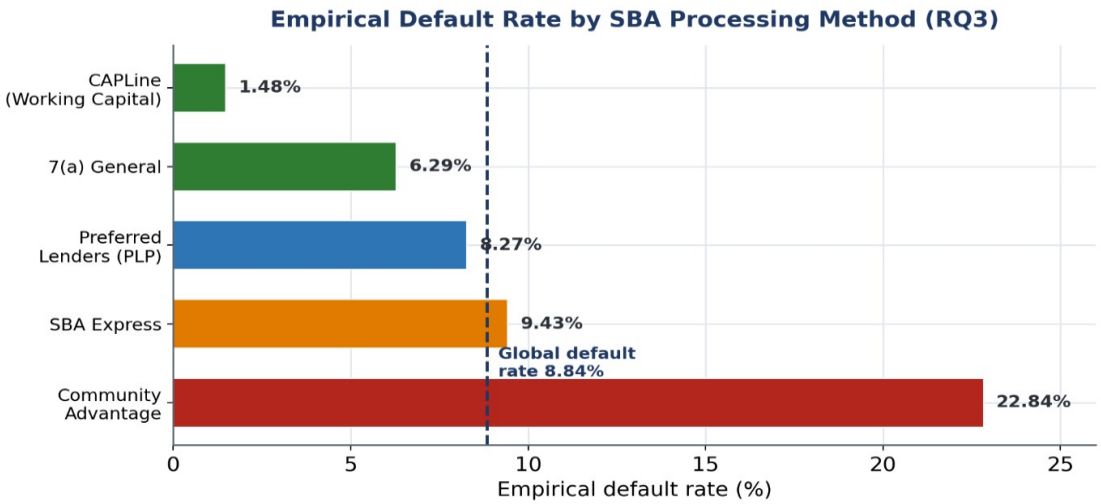


Figure 7. *Empirical default rate by SBA processing method*

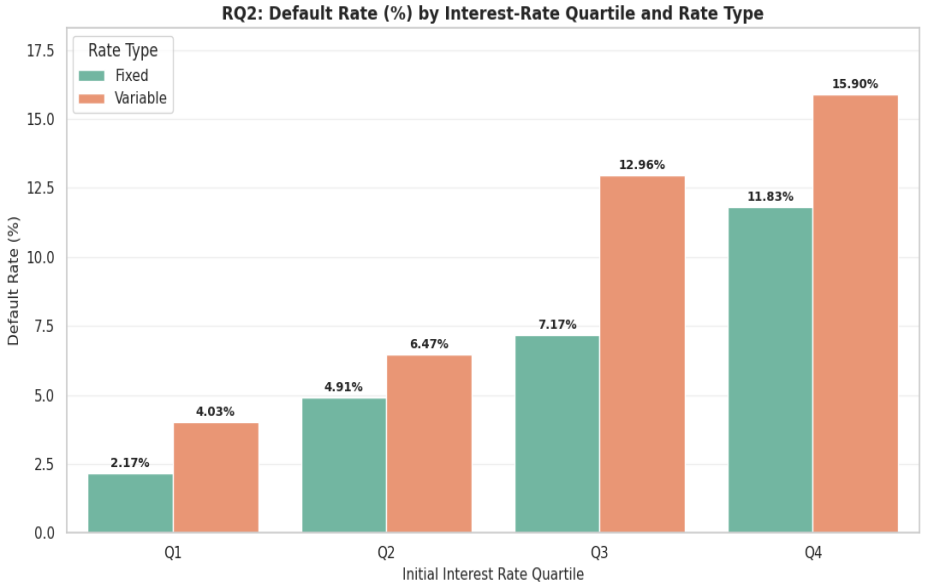


Figure 8. *Default Rate by Interest Rate Quartile and Fixed-or-Variable Rate Type*

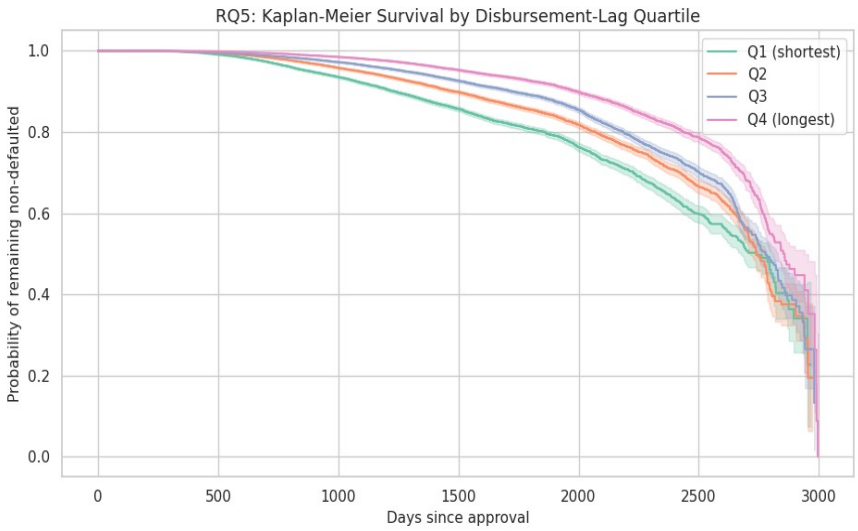


Figure 9. *Kaplan-Meier Survival Curves by Disbursement Lag Quartile*

Results by Research Question

RESEARCH QUESTION	STATISTICAL TEST & RESULT	DECISION	PAIRED MODEL & PERFORMANCE	KEY RELATIONSHIP
RQ1 NAICS sector & loan term	Chi-square χ^2 = 4,350.62 (df = 95), $p < .0001$ Cramér's V = 0.18 Arts/Entertainment short-term peaks at 24.4% default; long-term loans below 1%.	Reject H_0	XGBoost AUC-ROC 0.9614 recall ≈ 92%	Term length acts as an industry-conditioned screen on default.
RQ2 Interest rate level & type	Odds ratios 1.31 (rate), 6.84 (variable) LRT χ^2 = 4,559.19 · ANOVA F = 1,308.39 all $p < .0001$	Reject H_0	LightGBM ★ best overall AUC-ROC 0.9703 PR-AUC 0.8420	Higher rates and a variable-rate structure sharply raise default risk.
RQ3 SBA processing method	One-way ANOVA (large F, $p < .0001$) Tukey contrasts significant Default ranges 1.48% (CAPLine) to 22.84% (Community Advantage).	Reject H_0	Random Forest AUC-ROC 0.9421	The processing pathway is an independent risk classifier.
RQ4 Cross-state lending	Two-proportion z = 35.48 ($p < .0001$) Cohen's h = 0.21 Cross-state 10.94% vs same-state 5.20% — slightly more than double.	Reject H_0	MLP neural net highest precision 0.71	Geographic distance between lender and borrower raises default risk.
RQ5 Disbursement lag	Cox hazard ratio = 0.9944 (95% CI excludes 1, $p < .0001$) Concordance index = 0.81 Significant, but the direction is opposite the operational alternative.	H_0 not rejected (directional)	XGBoost AUC-ROC 0.9675	Longer lag behaves as a protective rather than a stress marker.

Overall Interpretation

The results indicate that loan-level structural features alone are highly predictive of SBA 7(a) default. Four of five null hypotheses are rejected at $\alpha = 0.05$, and the gradient-boosted models (LightGBM, XGBoost) deliver the strongest combined ranking and rare-class performance, consistent with the literature that boosting outperforms scorecards. Performance is influenced by three factors: the rare-event prevalence (8.84%), which is why PR-AUC and class weighting matter; the strong signal in categorical predictors (rate type, processing method, cross-state status); and calibration quality, which varies by model — Random Forest (RQ3) shows calibration drift addressed via isotonic regression. The RQ5 reversal reframes disbursement lag as protective, a nuance any early-warning rule must respect. Taken together, the models offer a credible, evidence-based complement to manual underwriting.

Four of five RQs yield strong evidence in favor of the alternative hypothesis. RQ5 yields a statistically significant but directionally inverted result. Recall on the default class stays above 0.58 across all models and reaches 0.92 in the best cases. Precision varies from 0.37 (Random Forest) to 0.71 (MLP). The pattern is consistent with the cost structure of small business lending, where false negatives are typically more costly than false positives. Probability thresholds therefore favor sensitivity over precision.

The strongest overall model on combined ranking and calibration is the LightGBM (RQ2). The weakest on calibration is the Random Forest (RQ3), which produces a negative R^2 on predicted probabilities. This calibration weakness does not undermine the statistical conclusion: the ANOVA for processing method clearly rejects the null. The negative R^2 indicates only that the Random Forest under chosen settings is overconfident in probability outputs. Future iterations will apply isotonic regression to recalibrate.

18. Practical Significance

The findings translate into concrete decision support. Lenders can use calibrated default-risk scores to price and approve loans by industry tier and rate type; SBA program managers can prioritize oversight toward high-risk pathways (for example, Community Advantage) and cross-state lending; and policymakers gain evidence on how structural program features drive guarantee-fund losses. Because the models use only fields the SBA already records, the insights are actionable without collecting new borrower data.

Implementation and User Benefit

The model is intended to be packaged as a lightweight scoring service plus a monitoring dashboard, usable by non-technical staff and requiring no borrower financial inputs at scoring time.

Deployment Approach

The recommended approach exposes the trained model (serialized with the best LightGBM configuration) behind a REST API that accepts a loan's structural attributes and returns a calibrated default probability. The API can be containerized and hosted on a standard cloud platform; a companion dashboard (built with a Python framework such as Streamlit or Dash or ReactJS framework, or a BI tool) visualizes portfolio-level risk. Tooling reuses the existing Python stack (scikit-learn, LightGBM, isotonic calibration) so that training and serving share one code path.

System Integration

The scoring API integrates into existing loan-origination and portfolio-management systems as a decision-support step: at application intake it returns a risk score to inform pricing and approval, and on a scheduled batch it rescoring the active portfolio to feed monitoring dashboards. Because inputs are limited to SBA-recorded structural fields, integration does not depend on external credit-bureau feeds.

User Interaction

Users interact through two surfaces. A loan officer enters or passes a loan's structural attributes (NAICS sector, term, rate and rate type, processing method, lender and borrower state, gross approval, guarantee percentage, disbursement timing) and receives a default-probability score with the contributing factors. A program manager views an aggregate dashboard with default-rate breakdowns by sector, processing method, and cross-state status, plus early-warning flags for high-risk segments.

Benefits to Users

Operationally, automated scoring reduces manual file review and speeds approval decisions. Financially, sharper risk discrimination supports risk-based pricing and earlier identification of loans likely to charge off, which can lower guarantee-fund losses on a portfolio with a near-9% charge-off rate. Strategically, the evidence base supports program-design and oversight decisions grounded in national-scale data rather than static thresholds.

Example Use Case

A lender receives a 7(a) application for a short-term, variable-rate loan to an arts-and-entertainment business, originated cross-state through the Community Advantage pathway. The scoring service flags an elevated default probability because several high-risk structural features coincide. The loan officer responds by adjusting pricing, requesting additional collateral, or routing the file for senior review, while the SBA monitoring dashboard records the loan in a high-risk segment for closer oversight — all before any charge-off occurs.

19. Limitations and Further Improvements

Limitations

- No borrower financials: the FOIA file omits credit scores and financial ratios, so the models rely solely on structural features.
- Probability calibration: the Random Forest (RQ3) produces overconfident probabilities (negative R^2 on predicted probabilities).
- Inverted disbursement-lag effect: the Cox hazard ratio runs opposite to the operational hypothesis, complicating any lag-based early-warning rule.
- Selection effects: the cross-state finding (RQ4) may partly reflect borrower self-selection rather than pure distance risk.
- Class imbalance carryover: despite class weighting, precision varies (0.37 for Random Forest to 0.71 for the MLP), indicating threshold sensitivity.

Impact of Limitations

These limitations bound the accuracy, generalizability, and reliability of the conclusions. Missing borrower financials cap how much variance the models can explain and mean the system must complement, not replace, traditional underwriting. Calibration drift affects the reliability of predicted probabilities used for pricing, even though it does not change the statistical conclusion that processing method matters. Possible selection effects limit causal interpretation of the cross-state result, and threshold sensitivity affects the precision a deployed system would achieve at a chosen operating point.

Future Improvements

- Apply isotonic regression to recalibrate the Random Forest and report calibrated RMSE, MAE, R^2, and Brier score.
- Use propensity-score matching on borrower controls to strengthen the cross-state lending finding against selection bias.
- Stratify the Cox analysis by approval fiscal year and macro-economic regime, and re-check censoring assumptions for the lag effect.
- Explore cost-sensitive thresholds and, where data allow, enrich features with external credit or macro indicators.
- Investigate deeper neural architectures if borrower-level or sequence data become available to justify them.

Future Scope

The study can be extended from a 7(a)-only design to a combined 7(a)/504 analysis where shared columns allow, and toward a real-time scoring service integrated with lender origination systems and SBA monitoring. Longer-horizon work could add survival-based loss forecasting, fairness auditing across borrower geographies and sectors, and periodic retraining as new FOIA releases arrive, broadening the system from retrospective analysis to ongoing portfolio risk management.

References

[1] Altman, E. I., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the U.S. market. *Abacus*, *43*(3), 332–357. https://doi.org/10.1111/j.1467-6281.2007.00234.x

[2] Calabrese, R., & Osmetti, S. A. (2013). Modelling small and medium enterprise loan defaults as rare events: The generalized extreme value regression model. *Journal of Applied Statistics*, *40*(6), 1172–1188. https://doi.org/10.1080/02664763.2013.784894

[3] Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

[4] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.*). Lawrence Erlbaum Associates*.

[5] Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

[6] 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

[7] Li, Y., & Zhu, J. (2022). Machine learning in credit risk modeling: A systematic literature review. *Expert Systems with Applications*, *204*, 117559. https://doi.org/10.1016/j.eswa.2022.117559

[8] Min, J. H., & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, *28*(4), 603–614. https://doi.org/10.1016/j.eswa.2004.12.008

[9] Rajan, R. G. (1992). Insiders and outsiders: The choice between informed and arm's-length debt. *The Journal of Finance*, *47*(4), 1367–1400. https://doi.org/10.1111/j.1540-6261.1992.tb04662.x

[10] SBA Office of Inspector General. (2022). *Top management and performance challenges facing the Small Business Administration in fiscal year 2023*. U.S. Small Business Administration. https://www.sba.gov/document/report-22-19-top-management-performance-challenges-facing-sba-fiscal-year-2023

[11] Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *The American Economic Review*, *71*(3), 393–410.

[12] U.S. Small Business Administration. (2023). *2023 small business profile*. Office of Advocacy. https://advocacy.sba.gov/

[13] U.S. Small Business Administration. (2025a). *FOIA 7(a) loans, FY2010–FY2019 [Data file]*. SBA Open Data Portal. https://data.sba.gov/dataset/7-a-504-foia

[14] U.S. Small Business Administration. (2025b). *FOIA 7(a) loans, FY2020–present [Data file]*. SBA Open Data Portal. https://data.sba.gov/dataset/7-a-504-foia

[15] Xu, W., Xiao, Z., Dang, X., Yang, D., & Yang, X. (2014). Financial ratio selection for business failure prediction using soft set theory. *Knowledge-Based Systems*, *63*, 59–67. https://doi.org/10.1016/j.knosys.2014.03.007

[16] Zhu, Y., Xie, C., Wang, G. J., & Yan, X. G. (2017). Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications*, *28*(S1), 41–50. https://doi.org/10.1007/s00521-016-2304-x

Appendix A: Acronyms and Their Full Forms

This appendix lists every acronym used in the report along with its full form. Acronyms are listed in alphabetical order.

Acronym	Full Form
ANOVA	Analysis of Variance
APA	American Psychological Association
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CHGOFF	Charged Off (loan status code)
CSV	Comma-Separated Values
EDA	Exploratory Data Analysis
FOIA	Freedom of Information Act
FY	Fiscal Year
HSD	Honestly Significant Difference (Tukey)
IQR	Interquartile Range
LRT	Likelihood-Ratio Test
MAE	Mean Absolute Error
MLP	Multilayer Perceptron
NAICS	North American Industry Classification System
PIF	Paid in Full (loan status code)
PR-AUC	Area Under the Precision-Recall Curve
RMSE	Root Mean Square Error
RQ	Research Question
SBA	U.S. Small Business Administration
SME	Small and Medium Enterprise
SVM	Support Vector Machine

Appendix B: Exploratory Data Analysis (EDA) Supporting Figures

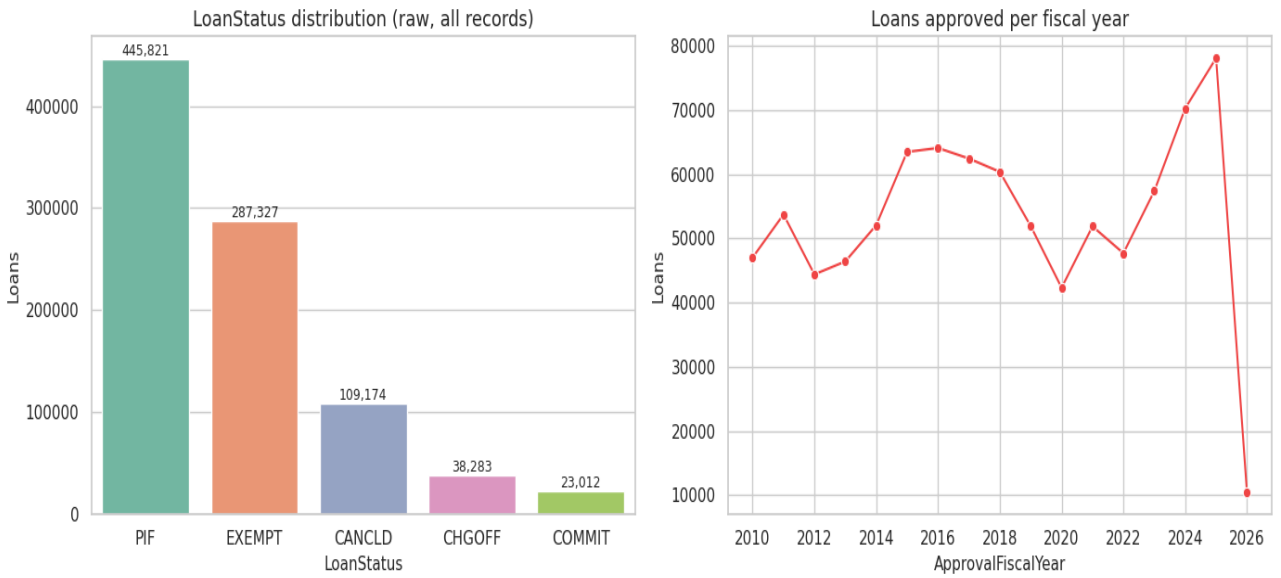


Figure B1. *LoanStatus distribution across the full raw dataset (left) and approvals per fiscal year (right).*

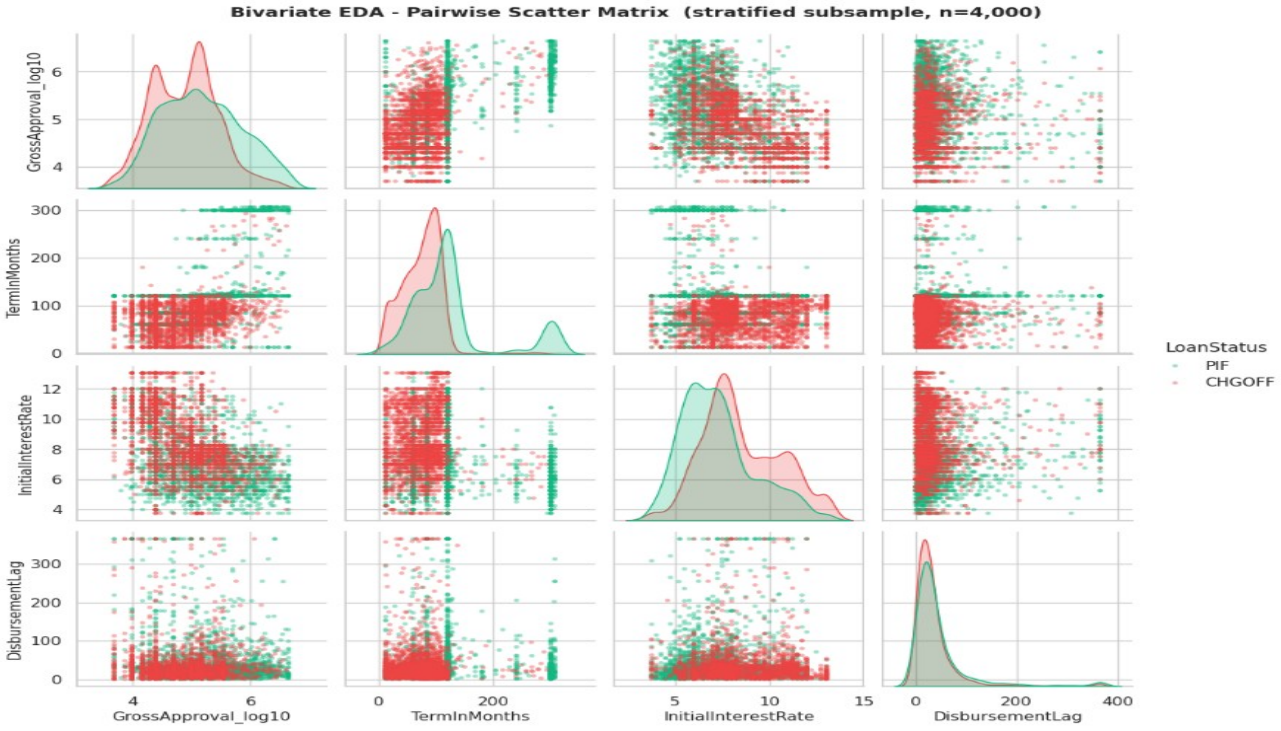


Figure B2. *Pairwise scatter matrix of the top numeric features, colored by default status.*

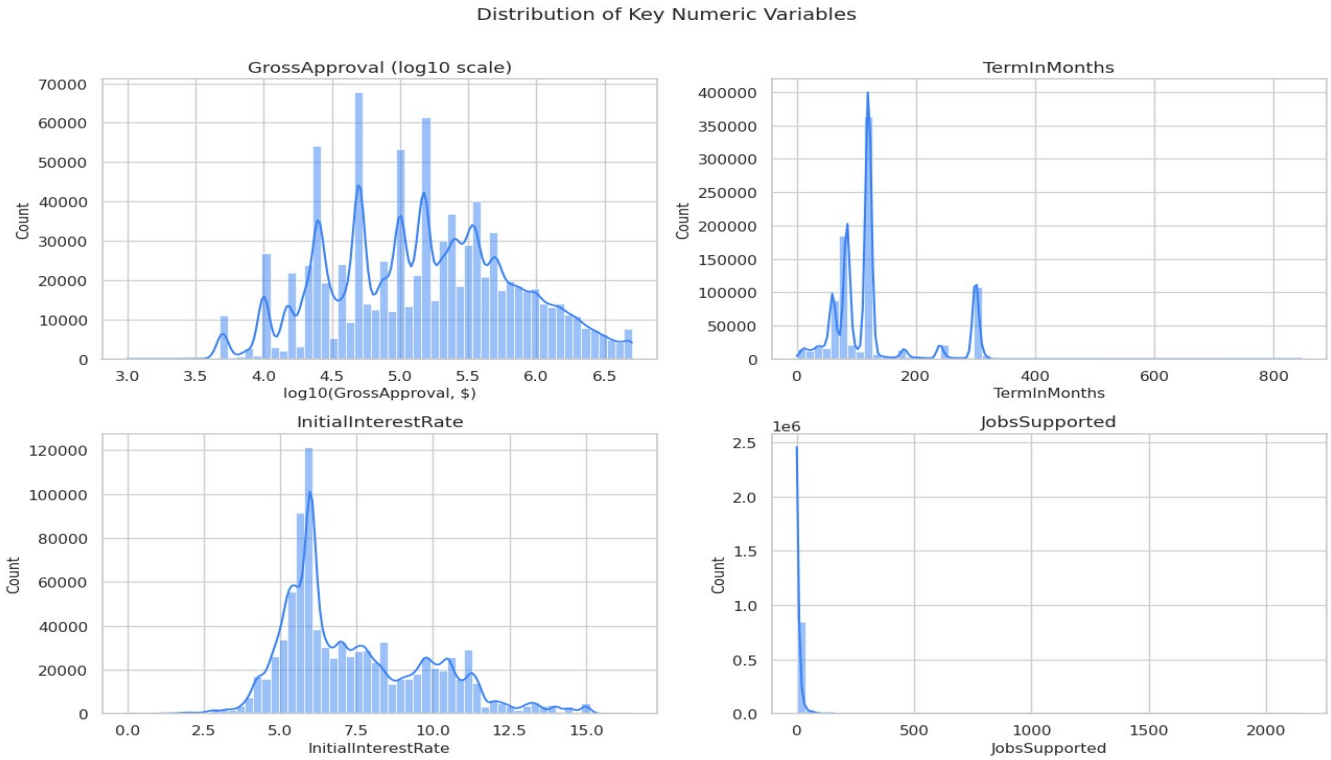


Figure B3. *Distributions of Key Numeric Predictors in the Raw Dataset*

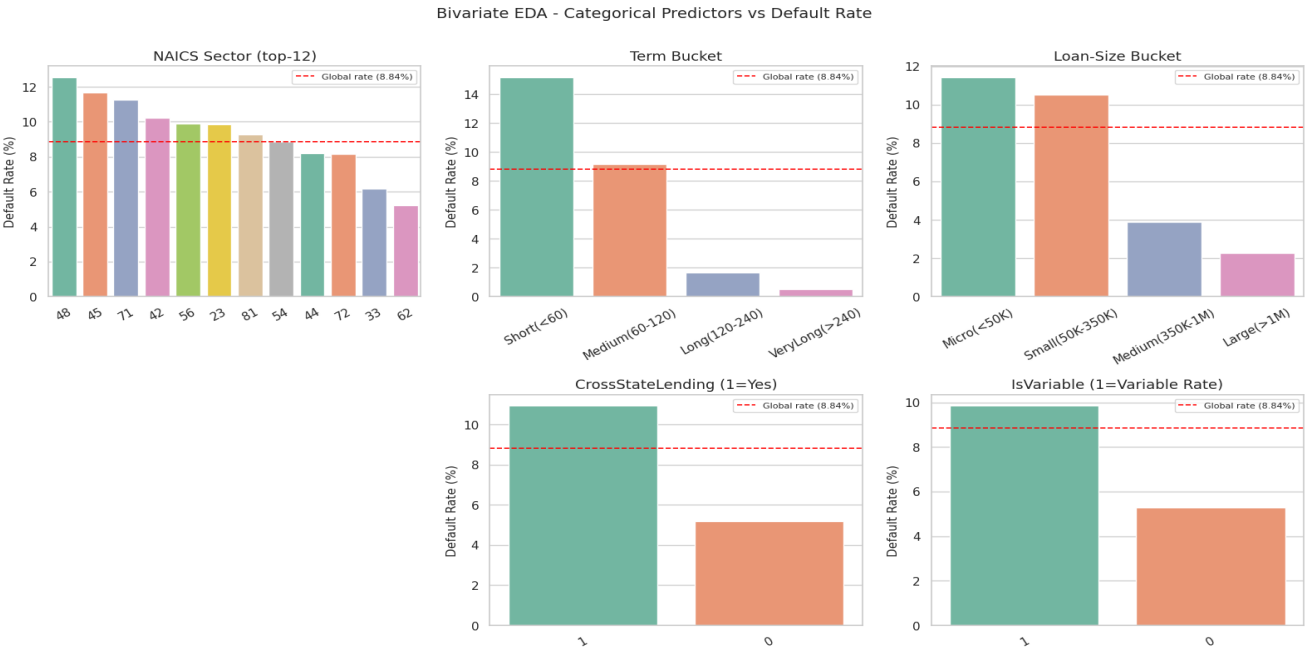


Figure B4.*Default Rate by Categorical Predictor Levels*

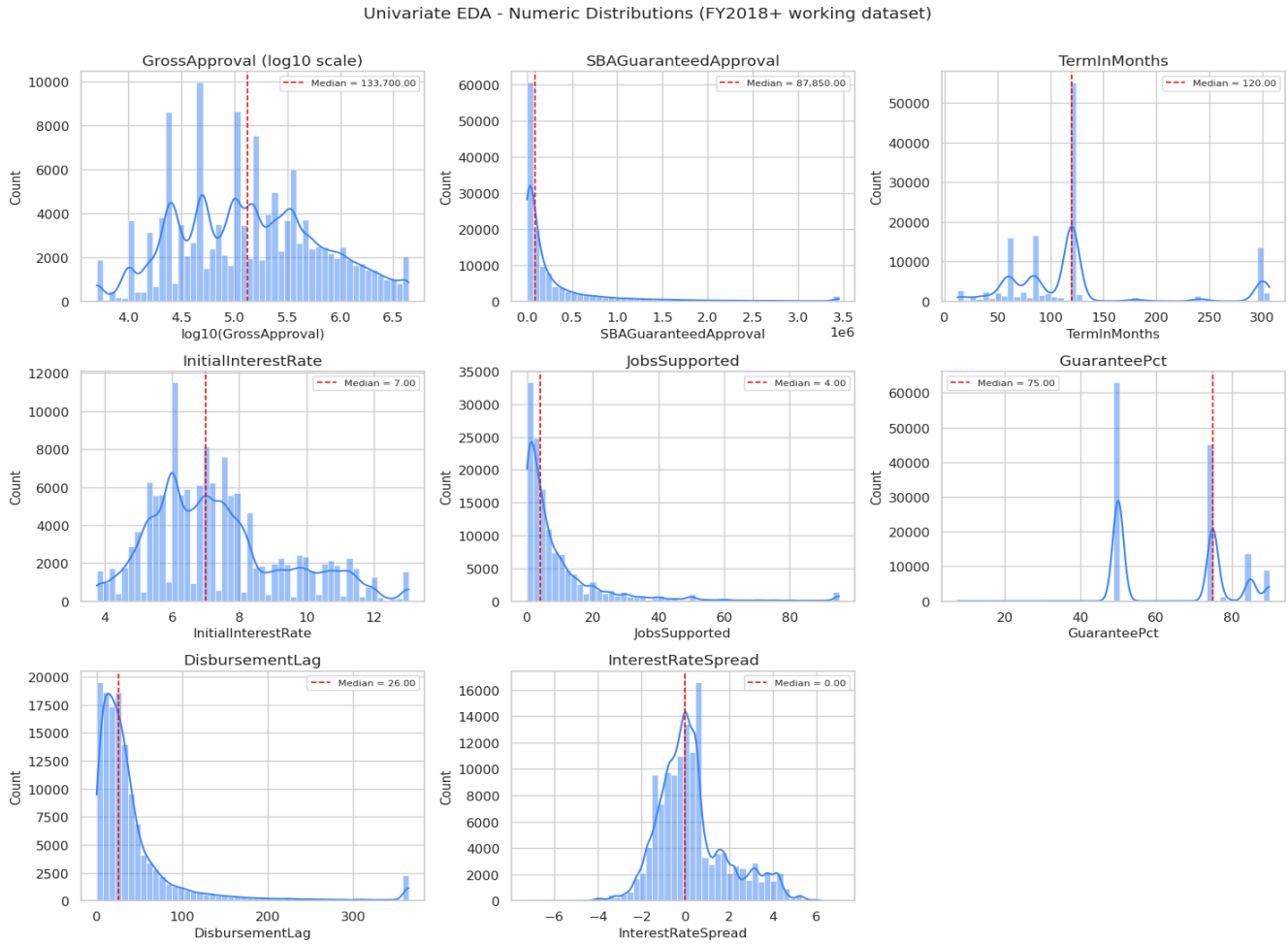


Figure B5. *Numeric Predictor Distributions in the Working Sample*

Appendix C: Preliminary Findings by Research Question

Appendix C holds the two simpler bar charts referenced in the RQ3 and RQ4 sections. They are placed here to keep the main results concise without losing the visual evidence.

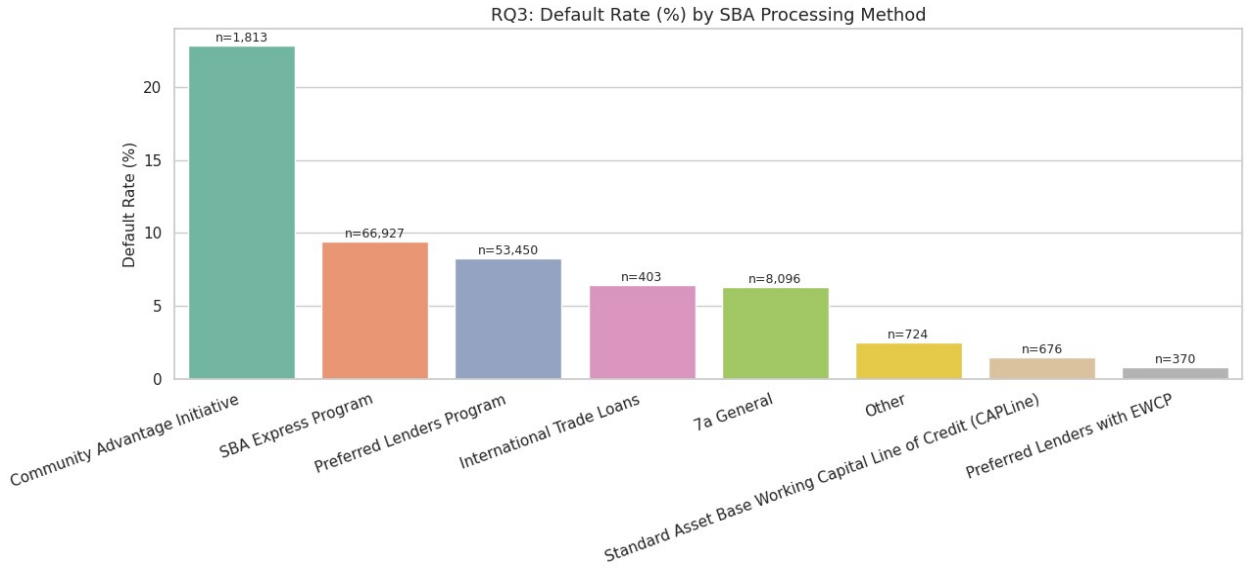


Figure C1. *Default Rate by SBA Processing Method Group*

Note. Bars annotated with sample size (n). Community Advantage Initiative (n = 1,813) shows the highest default rate at 22.84%; Standard CAPLine (n = 676) shows the lowest at 1.48%.

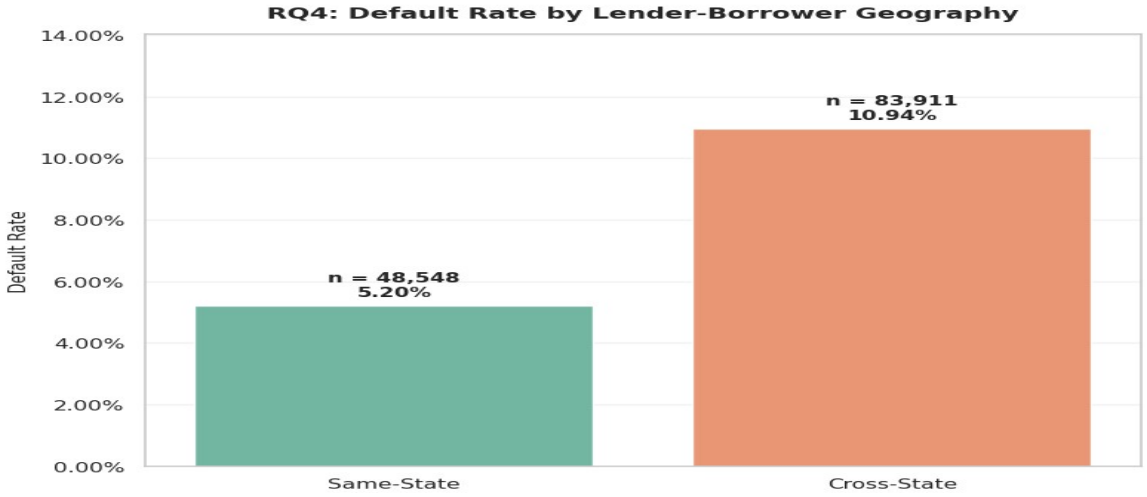


Figure C2. *Default Rate Comparison Between Cross-State and Same-State Loans*

Note. Same-state loans (n = 48,548) default at 5.20%. Cross-state loans (n = 83,911) default at 10.94%, more than double the same-state rate.

Appendix D: Preliminary Performance Evaluations

This appendix presents the confusion matrices for all five research questions across the three data partitions. The training partition holds seventy percent of the working sample, the validation partition holds fifteen percent, and the test partition holds the remaining fifteen percent. Each confusion matrix is read in the standard layout: rows represent the actual class, and columns represent the predicted class. The top-left cell holds true negatives (paid in full predicted as paid in full), the top-right cell holds false positives, the bottom-left cell holds false negatives, and the bottom-right cell holds true positives (charged off predicted as charged off).

RQ1: XGBoost Confusion Matrices

The RQ1 XGBoost classifier targets the joint NAICS sector and loan term effect on default. Figure D1 shows the confusion matrices for the train, validation, and test partitions. Recall on the default class stays at or above ninety-two percent across all three partitions. The pattern shows that the model captures most actual defaults, at the cost of a moderate number of false alarms.

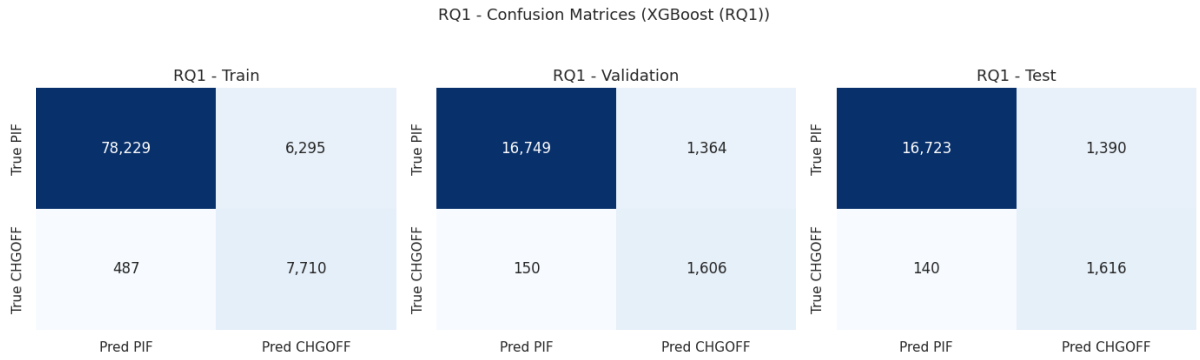


Figure D1. *RQ1 XGBoost Confusion Matrices for Train, Validation, and Test Partitions*

RQ2: LightGBM Confusion Matrices

The RQ2 LightGBM classifier targets the interest rate and rate type effect on default. Figure D2 shows the confusion matrices for the train, validation, and test partitions. The recall stays around ninety percent. The precision is the highest among the five gradient boosting models, which suggests that interest rate features sharpen the boundary between paid in full and charged off loans.

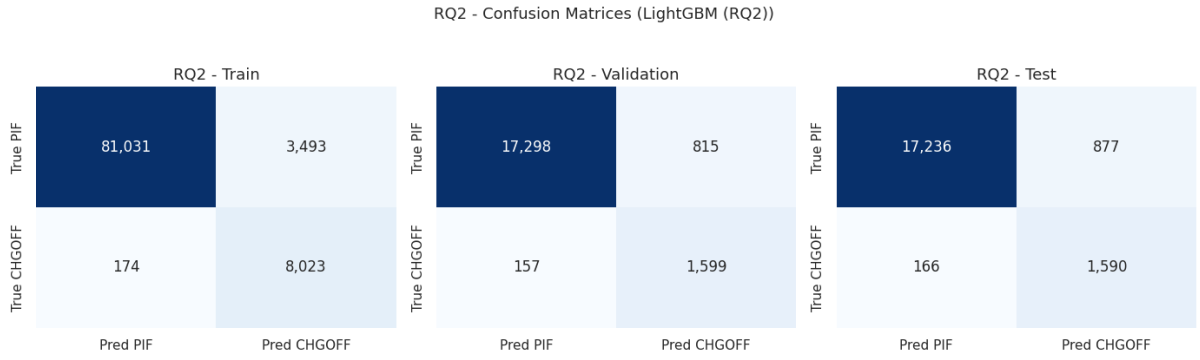


Figure D2. *RQ2 LightGBM Confusion Matrices for Train, Validation, and Test Partitions*

RQ3: Random Forest Confusion Matrices

The RQ3 Random Forest classifier targets the SBA processing method effect on default. Figure D3 shows the confusion matrices for the train, validation, and test partitions. The recall is high near ninety-one percent. The precision is the lowest in the study at about thirty-seven percent. The high false positive count

drives the negative R-squared on predicted probabilities and motivates the planned isotonic-regression calibration step.

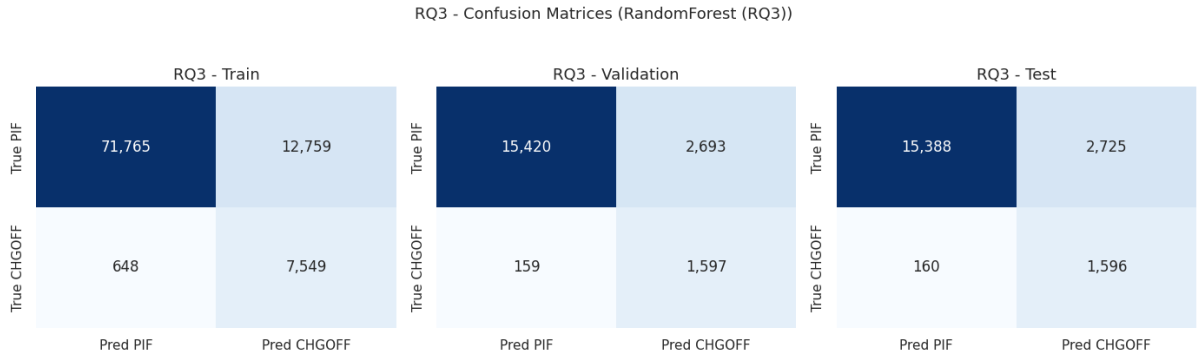


Figure D3. *RQ3 Random Forest Confusion Matrices for Train, Validation, and Test Partitions*

RQ4: Multilayer Perceptron Confusion Matrices

The RQ4 multilayer perceptron classifier targets the cross-state lending effect on default. Figure D4 shows the confusion matrices for the train, validation, and test partitions. The neural network achieves the highest precision in the study at about seventy-one percent. The trade-off is a lower recall at fifty-eight percent on the test partition.

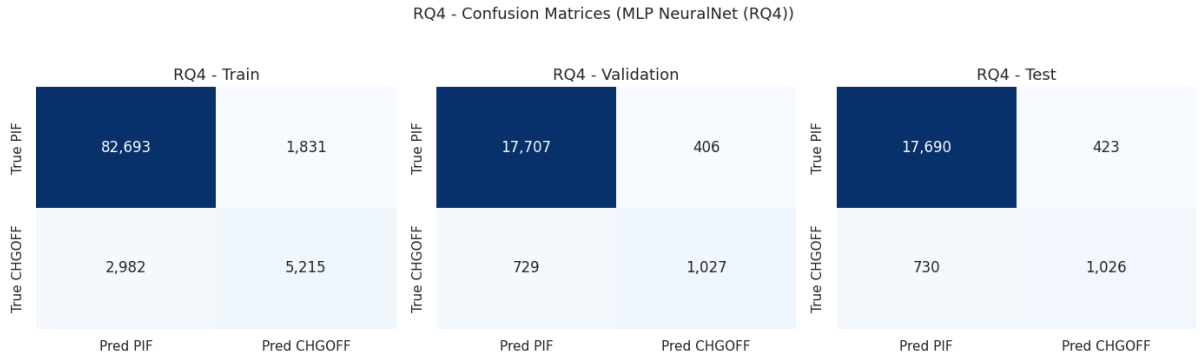


Figure D4. *RQ4 Multilayer Perceptron Confusion Matrices for Train, Validation, and Test Partitions*

RQ5: XGBoost Classifier Confusion Matrices

The RQ5 XGBoost classifier targets the disbursement lag effect on default. Figure D5 shows the confusion matrices for the train, validation, and test partitions. Recall stays around ninety percent. Precision sits near fifty-eight percent. These values match the second-best AUC-ROC in the study.

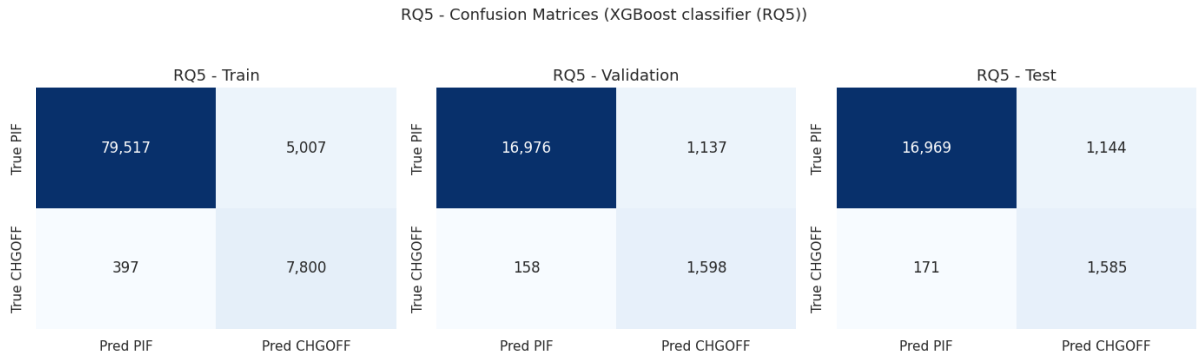


Figure D5. *RQ5 XGBoost Classifier Confusion Matrices for Train, Validation, and Test Partitions*

Appendix E: Complete Data Dictionary

This appendix reproduces the official field-level data dictionary published by the U.S.

Small Business Administration alongside the FOIA loan-level CSV files used in this study. Section A.1 documents all fields available in the SBA 7(a) loan program files (the primary data source for this capstone) and Section A.2 documents the SBA 504 loan program fields, included for reference and to clarify the rationale for the 7(a)-only design described in the synopsis. Field names are reproduced verbatim as they appear in the source CSV column headers.

E.1 — SBA 7(a) Loan Program Data Dictionary (Primary Data Source)

Field / Attribute	Description
AsOfDate	Date when the data was recorded
Program	Indicator of whether loan was approved under SBA's 7(a) or 504 loan program
LocationID	SBA's unique lender ID code
BorrName	Borrower name
BorrStreet	Borrower street address
BorrCity	Borrower city
BorrState	Borrower state
BorrZip	Borrower zip code
BankName	Name of the bank that the loan is currently assigned to
BankFDICNumber	The Federal Depository Insurance Corporation certificate ID of the lender

Field / Attribute	Description
BankNCUANumber	The National Credit Union Association charter number of the lender
BankStreet	Bank street address
BankCity	Bank city
BankState	Bank state
BankZip	Bank zip code
GrossApproval	Total loan amount
SBAGuaranteedApproval	Amount of SBA's loan guaranty
ApprovalDate	Date the loan was approved
ApprovalFY	Fiscal year the loan was approved
FirstDisbursementDate	Date of first loan disbursement (if available)

Field / Attribute	Description
Subprogram	Subprogram description - specific subprogram loan was approved under. See SOP 50 10 5 for definitions and rules for each subprogram.
InitialInterestRate	Initial interest rate - total interest rate (base rate plus spread) at time loan was approved
FixedorVariableInterestInd	Fixed/variable interest rate indicator
TermInMonths	Length of loan term
NaicsCode	North American Industry Classification System (NAICS) code
NaicsDescription	North American Industry Classification System (NAICS) description
FranchiseCode	Franchise Code

Field / Attribute	Description
FranchiseName	Franchise Name (if applicable)
ProjectCounty	County where project occurs
ProjectState	State where project occurs
SBADistrictOffice	SBA district office
CongressionalDistrict	Congressional district where project occurs
BusinessType	Borrower Business Type - Individual, Partnership, or Corporation
BusinessAge	SBA began collecting the following business age information in fiscal year 2018: • Change of Ownership • Existing or more than 2 years old • New Business or 2 years or less • Startup, Loan Funds will Open Business
LoanStatus	Current status of loan: • COMMIT = Undisbursed • PIF = Paid In Full • CHGOFF = Charged Off • CANCLD = Cancelled • EXEMPT = The status of loans that have been disbursed but have not been cancelled, paid in full, or charged off are exempt from disclosure under FOIA Exemption 4
PaidInFullDate	Date loan was paid in full (if applicable)
ChargeOffDate	Date SBA charged off loan (if applicable)
GrossChargeOffAmount	Total loan balance charged off (includes guaranteed and nonguaranteed portion of loan)
RevolverStatus	Indicator of whether a loan is a term loan or revolving line of credit (0=Term, 1=Revolver)

Field / Attribute	Description
JobsSupported	Total Jobs Created + Jobs Retained as reported by lender on SBA Loan Application. SBA does not review, audit, or validate these numbers - they are simply self-reported, good faith estimates by the lender.
CollateralInd	An indicator whether the SBA lender reported that the loan was backed by collateral
SoldSecMrktInd	An indicator if the loan was sold on the secondary market. This is a static field once it is sold on the secondary market.Equals 'Y', if sold on the secondary market. Once it is 'Y' it will stay 'Y' for its entirety.

Appendix F – Sample Size Calculations for all Research Questions

RQ1 - Chi-Square Test of Independence (NAICS Sector × Term Bucket)

Research Question 1 tests whether the joint distribution of NAICS industry sector and loan term bucket is associated with default status. The appropriate test is Pearson's chi-square test of independence applied to the three-way contingency between NAICS sector, term bucket, and loan outcome (paid in full versus charged off).

Using G*Power 3.1, the χ^2 tests → goodness-of-fit tests: contingency tables family was selected with a small effect size of w = 0.10 (Cohen, 1992). The degrees of freedom were computed from the expected cross-tabulation as df = (r − 1)(c − 1), where r = 20 NAICS two-digit sectors expected to appear in the 7(a) data and c = 4 term buckets (Short, Medium, Long, Very Long), giving df = 19 × 3 = 57. With α = 0.05 and power (1 − β) = 0.80, G*Power returned a required total sample size of N = 3,190.

RQ2 - Linear Multiple Regression (Interest Rate and Rate Type Effects)

Research Question 2 evaluates the incremental contribution of initial interest rate and rate type (fixed versus variable), plus their interaction, to the prediction of default after controlling for loan size, term, guarantee percentage, and NAICS sector. Although the binary outcome is ultimately modeled with logistic regression, the G*Power a-priori calculation uses the linear multiple regression framework on the latent-variable scale, which is accepted practice when a logistic analog is unavailable and yields a conservative N (Faul et al., 2007).

The linear multiple regression: fixed model, R^2 increase family was used with a small effect size of f^2 = 0.02, u = 3 tested predictors (rate, rate type, and interaction), and a total predictor count of 7 after including the four covariates. With α = 0.05 and power = 0.80, the required total sample size was N = 550.

RQ3 - One-Way ANOVA (Processing Method Effects)

Research Question 3 compares default rates across SBA processing methods. Because the outcome is a proportion and the number of groups is greater than two, a one-way fixed-effects analysis of variance on empirical logits is the appropriate parametric omnibus test. The corresponding G*Power family is F tests → ANOVA: fixed effects, omnibus, one-way.

A small effect size of f = 0.10 was chosen (Cohen, 1992). The number of groups was set at k = 8, reflecting the processing-method codes that are empirically well-populated in the 7(a) FOIA data (PLP, CLP, 7AG, SBX, CAI, SBL, PTX, and pooled other). With α = 0.05 and power = 0.80, G*Power returned a required total sample size of N = 1,443.

RQ4 - Two-Proportion z-Test (Cross-State Lending)

Research Question 4 compares the default rate among loans where the lender state differs from the borrower state against the default rate among same-state loans. The appropriate test is the z-test for the difference between two independent proportions, which in G*Power 3.1 is the z tests → proportions: difference between two independent proportions family.

Cohen's effect size index h was set at h = 0.10 (small). With α = 0.05, two-tailed, power = 0.80, and equal allocation between cross-state and same-state groups (allocation ratio = 1), G*Power returned a per-group n of 1,571, giving a required total sample size of N = 3,142. This is the binding constraint across all five research questions.

RQ5 - Cox Proportional Hazards Regression (Disbursement Lag)

Research Question 5 treats default as a time-to-event outcome and tests whether disbursement lag (days between loan approval and first disbursement) is associated with the hazard of charge-off, controlling for loan, borrower, and macroeconomic covariates. The appropriate test is Cox proportional hazards regression with a continuous predictor.

The G*Power 3.1 z tests → Cox regression family was used. A hazard ratio of HR = 1.20 was specified as the smallest effect considered clinically meaningful for operational early-warning use. The event probability was set at pE = 0.079, matching the empirically observed 7(a) charge-off rate among resolved loans (FY2010–FY2019). The R^2 between the covariate of interest and the other covariates was set to 0, giving a conservative (largest) sample requirement. With $\alpha = 0.05$, two-tailed, and power = 0.80, G*Power returned a required sample size of N = 2,989.