College of Business and Management, VinUniversity
BANA4020 - MACHINE LEARNING FOR BUSINESS ANALYTICS

Final Project Report

Predictive Modeling Using SBA National Dataset

Group 4
Doan Huong Thanh - V202301046
Do Quang Hai - V202200822
Hoang Nguyen Gia Huy - V202200853

Instructor, Prof. Pham Tuan Minh
Teaching Assistant, Nguyen Thi Minh Thanh
June 9th, 2025

1. Introduction:

The Small Business Administration (SBA), established in 1953, has played a crucial role in supporting small businesses and entrepreneurship in the United States. Through its loan guarantee programs, the SBA helps reduce lending risks, encouraging banks to extend credit to small enterprises that drive job creation and economic development. With a network of 10 regional offices and a workforce exceeding 8,000, the SBA facilitated over 103,000 small-business financings in fiscal year 2024, a 22% increase from the previous year, contributing a total capital impact of $56 billion.

However, not all loans are successfully repaid. Loan defaults present significant financial risks to lenders, making effective risk assessment a crucial part of the lending process. Traditionally, loan decisions were often based on credit history and qualitative assessments. Yet, as financial markets become more complex, data-driven approaches have become increasingly important. Data analysis allows lenders to move beyond intuition, leveraging historical trends and statistical models to improve the accuracy of credit risk evaluations and make more informed, objective lending decisions.

This project aims to develop a comprehensive predictive framework for small-business loan defaults using various machine learning algorithms, including Logistic Regression, Random Forest, XGBoost, Neural Network, Multilayer Perceptron. Beyond classification accuracy, we incorporate a cost-benefit analysis through a customized cost matrix to better reflect the financial consequences of loan decisions. By aligning model outputs with profit-maximization goals, the final recommendation will not only predict default risks but also support lenders in optimizing their approval strategies for maximum financial return.

2. Literature Review:

2.1. Applications of the SBA dataset

The U.S. Small Business Administration (SBA) loan dataset has become a valuable resource for research in the area of credit risk assessment. Its large scale, real-world context, and rich feature set make it particularly well-suited for modeling default risk among small businesses. Early studies, such as Li et al. (2018), introduced the dataset as part of an educational framework to teach statistical thinking and risk-based decision making through logistic regression models. Their work highlighted how real business data could support predictive modeling and informed lending decisions in a practical, accessible way.

Beyond educational applications, recent projects have leveraged the SBA dataset for more advanced machine learning implementations. For example, Bagjasatia (2021) developed a loan approval recommendation system using a web-based application, trained on historical SBA loan data. This project not only demonstrated the dataset's versatility but also demonstrated potential for real-world application, offering a scalable decision-support tool for financial institutions. By improving the accuracy and efficiency of loan approval processes, such systems can help lenders reduce default risks and better allocate capital. These outcomes highlight the necessity of translating academic research into practical solutions, particularly in high-stakes industries like finance. As a result, continuing to develop and refine such projects is crucial to bridge the gap

between theoretical models and operational decision-making, ultimately enhancing risk management and financial inclusion.

Across these studies, the SBA dataset has been commonly used for:

01. Building classification models to predict loan default (Paid-in-Full vs. Charged-Off).
02. Benchmarking machine learning algorithms under realistic credit risk scenarios.
03. Analyzing industry-level variations in loan performance based on NAICS classifications.
04. Investigating the influence of macroeconomic conditions, such as recession periods, on loan outcomes.

Key predictors repeatedly found to be influential include business industry, business age (new versus existing), loan size, real estate collateral, and the proportion of SBA guarantees.

2.2. Common modeling approaches

A range of statistical and machine learning methods have been applied to predict loan defaults based on the SBA dataset:

a. Logistic Regression: Frequently used for its simplicity and interpretability, serving as a baseline model for binary classification tasks in many studies (Li et al., 2018).
b. Decision Trees and Random Forests: Decision trees provide clear decision-making rules, while random forests aggregate multiple trees to improve predictive performance and handle high-dimensional data effectively.
c. Gradient Boosting Machines (GBM): Techniques such as XGBoost have shown superior performance, achieving approximately 91% accuracy and an AUC of 0.971 on the SBA dataset (Bagjasatia, 2021). GBM models are particularly valued for their ability to model non-linear relationships and manage class imbalance.
d. K-Nearest Neighbors (KNN) and Support Vector Machines (SVM): While less commonly used due to computational intensity, these methods have been explored in some studies for comparative purposes.

Most projects utilize an 80/20 train-test split for model validation, incorporate hyperparameter tuning methods like GridSearchCV or RandomizedSearchCV, and assess model performance using metrics such as Accuracy, Precision-Recall, ROC-AUC, and Confusion Matrices.

2.3. Insights from previous research

2.3.1. Key findings

Prior research on the SBA dataset has consistently highlighted several important insights into loan default prediction.

Default risk is not evenly distributed; it varies substantially based on borrower and loan characteristics. Businesses across industries show distinct risk profiles, with sectors like agriculture and healthcare typically exhibiting lower default rates compared to more volatile sectors such as real estate and retail (Li et al., 2018).

Other significant predictors include the age of the business, where new ventures often face higher risks, the loan amount, the presence of real estate collateral, and exposure to recession periods (Li et al., 2018).

In terms of modeling approaches, gradient boosting methods, particularly XGBoost, have demonstrated the strongest predictive performance, outperforming traditional models like logistic regression in terms of both accuracy and area under the curve (AUC) (Bagjasatia, 2021).

Gradient oosting's ability to capture complex, non-linear relationships and handle class imbalance makes it especially effective for credit risk modeling in the context of SBA loan data.

2.3.1. Limitations and areas for improvement

Despite the progress made, several limitations persist in studies using the SBA dataset. First, the dataset includes only approved loans, excluding rejected applications. This selection bias may distort the model by overrepresentation loans already considered acceptable by lenders (Li et al., 2018). Second, key financial variables such as borrower credit scores, annual revenues, and debt levels are not available, limiting the models' ability to fully replicate the loan approval decision-making process (Li et al., 2018). Third, many existing models focus primarily on maximizing classification metrics like accuracy and AUC, often neglecting the asymmetric costs of misclassifications – where a false approval can be far more costly than a false rejection (Bagjasatia, 2021). Finally, although complex models like gradient boosting provide better predictive performance, they also reduce model interpretability, creating challenges for implementation in highly regulated financial environments where transparency is critical (Bagjasatia, 2021).

To overcome these limitations, recent research advocates for the incorporation of cost-sensitive learning frameworks and the application of model explainability techniques such as SHAP (SHapley Additive exPlanations) and LIME (local interpretable model-agnostic explanations) to balance accuracy, financial impact, and transparency (Bagjasatia, 2021).

Following this direction, in our project, we have integrated a customized cost matrix to optimize profit and augmented the dataset with macroeconomic features to enhance predictive performance.

3. Data Description:

3.1. Data source and collection

The dataset used in this project is the publicly available SBA 7(a) loan dataset, covering the period from 1987 to 2014. It was sourced from the Small Business Administration's (SBA) National Loan Database. The data is aggregated from participating lenders, including banks, credit unions, and community lenders, and contains detailed information on borrowers, loan terms, and final repayment statuses as reported by the lenders.

3.2. Structure and variables

The dataset consists of approximately 899,000 individual loan records, with each row representing a unique SBA-guaranteed loan application. Each record contains detailed information about the loan, the borrower, and the lender involved.

Figure 1: Description of Original Variables in the SBA Loan Dataset

Variable name	Data type	Description of variable
LoanNr_ChkDgt	Categorical	Loan number check digit (Unique identifier for each loan)
Name	Categorical	Name of the business applying for the loan
City	Categorical	City where the business is located
State	Categorical	State where the business is located
Zip	Categorical	ZIP code of the business location
Bank	Categorical	Name of the bank providing the loan
BankState	Categorical	State where the bank is headquartered
NAICS	Categorical	NAICS code identifying the business industry
ApprovalDate	Temporal	Date when the loan was approved
ApprovalFY	Discrete	Fiscal year when the loan was approved
Term	Numerical	Loan term (duration of the loan in months)
NoEmp	Numerical	Number of employees at the business at loan application
NewExist	Categorical	Indicates whether the business is new (start-up) or existing
CreateJob	Numerical	Number of jobs created as a result of the loan
RetainedJob	Numerical	Number of jobs retained due to the loan
FranchiseCode	Categorical	Code indicating whether the business is a franchise
UrbanRural	Categorical	Indicates whether the business is located in an urban or rural area
RevLineCr	Categorical	Whether the loan is a revolving line of credit (Yes/No)
LowDoc	Categorical	Whether the loan application used low documentation (Yes/No)
ChgOffDate	Temporal	Date when the loan was charged off (only for defaulted loans)
DisbursementDate	Temporal	Date when the loan amount was disbursed to the borrower
DisbursementGross	Numerical	Gross amount of loan disbursed to the borrower
BalanceGross	Numerical	Outstanding balance of the loan at the time of reporting
MIS_Status	Categorical	Final status of the loan (Paid in Full=PIF or Charged Off=CHGOFF)
ChgOffPrinGr	Numerical	Amount of principal charged off for defaulted loans
GrAppv	Numerical	Gross amount of loan approved
SBA_Appv	Numerical	Amount of the loan guaranteed by the SBA

Numerical variables were standardized using the Standard Scaler method, transforming features to have a mean of zero and a standard deviation of one. This ensures that variables with larger scales do not dominate model training and improves the convergence of certain algorithms.

Categorical variables were transformed using One-Hot Encoding, converting each category into a separate binary feature. This approach prevents the model from assuming an ordinal relationship among categorical values and preserves the nominal nature of these variables.

3.3. Response variable and modeling objective

The response variable in this study is MIS_Status, a categorical variable indicating the final outcome of each loan. It has two possible values:

a. Paid in Full (PIF): The borrower successfully repaid the loan.
b. Charged Off (CHGOFF): The loan was written off as a loss due to default.

Given the binary nature of the response variable, the predictive modeling task is formulated as a binary classification problem. The objective is to predict the likelihood of a loan being paid in full or charged off based on a set of borrower, business, and loan characteristics available at the time of loan approval.

The specific modeling objectives are:

a. Probability estimation: Predict the likelihood that a borrower will repay the loan in full versus default.
b. Cost-sensitive learning: Incorporate a cost-sensitive approach to penalize false negatives (i.e., incorrectly approving a borrower who defaults) more heavily than false positives (i.e., rejecting a borrower who would have repaid).
c. Profit maximization: Identify an optimal decision threshold that maximizes net profit on a validation set by balancing interest income against potential default losses.
d. Targeted decision support: Generate gain and lift charts to assist credit officers in identifying the top percentage of applicants with the highest predicted repayment probabilities for optimal portfolio returns.

Furthermore, the feature set is augmented with macroeconomic indicators, such as recession flags, inflations, to account for external economic factors that may influence borrowers' repayment behaviors. This modeling framework not only aims to achieve high predictive accuracy but also ensures alignment with the financial objectives and operational requirements of real-world credit risk management.

3.4. Basic data exploration

Prior to model development, an initial exploration of the dataset was conducted to gain insights into the characteristics and quality of the data. The dataset contains approximately 899,000 individual loan records, each representing a unique loan guaranteed by the SBA. It includes 27 variables, covering borrower details, loan attributes, geographical information, and loan outcomes.

Missing value overview

An examination of missing values revealed several notable patterns. Certain categorical fields, such as UrbanRural, utilize specific codes (e.g., "0") to denote undefined entries rather than explicit missing values. Additionally, some loans lack a ChgOffDate because they were never defaulted; this pattern is expected and requires careful handling during the calculation of derived features such as loan age. A small fraction of records (<5%) have missing values in variables such as NoEmp, CreateJob, and RetainedJob. For these variables, median imputation is planned to ensure consistency across the dataset.

Target variable distribution

The response variable, MIS_Status, is moderately imbalanced, with approximately 82% of loans being classified as Paid in Full (PIF) and 18% as Charged Off (CHGOFF). This imbalance will be taken into consideration during model training to ensure that classification performance is not biased towards the majority class.

Summaries of key numerical features

An exploration of key numerical features reveals significant skewness in several variables. The DisbursementGross variable, representing loan size, ranges from $1,000 to $5,000,000 with a median value of approximately $67,500. Its distribution is highly right-skewed due to the presence of a small number of extremely large loans. Similarly, the NoEmp variable, which captures the number of employees at the time of application, ranges from 0 to over 3,000, with most businesses employing fewer than 20 people, reflecting a concentration of small businesses. The CreateJob and RetainedJob variables show that most businesses report creating or retaining fewer than 10 jobs, although the possible range extends beyond 100.

Summaries of key categorical features

The categorical variables also present important insights. The NAICS industry classification encompasses over 1,000 unique codes, which have been grouped into broader two-digit industry sectors for analysis. Default rates vary substantially across industries, with higher rates observed in the construction sector and lower rates in health care. Loans are distributed across all 50 U.S. states, Washington D.C., and territories, with higher default rates identified in recession-impacted states such as California. Regarding location types, UrbanRural classification shows that 70% of loans are made to businesses in urban areas, 25% in rural areas, and 5% in undefined regions. Urban loans exhibit slightly lower default rates compared to rural loans.

Preliminary insights into predictors

Further, preliminary insights into the relationship between predictors and loan outcomes reveal that the median loan size for Paid in Full loans is approximately $75,000, compared to $50,000 for Charged Off loans. Default rates are notably higher in sectors such as accommodation and food services (approximately 25%) and significantly lower in the health care sector (approximately -10%). Firm size also appears to influence default likelihood, with small firms (less than five employees) experiencing default rates around 22%, whereas larger firms (more than 20 employees) demonstrate a significantly reduced default rate of approximately -12%.

These findings provide a foundational understanding of the data and offer early indications of the features that may be predictive of loan outcomes. This exploration informs subsequent data preprocessing, feature engineering, and model development steps.

4. Methodology:

In this study, we developed a systematic, profit-centered approach to predict SBA loan defaults and inform approval decisions. Building upon established case-study frameworks in statistical education (Li, Mickel, & Taylor, 2018) and economic model, our methodology also makes improvements in terms of rigorous data preparation, systematic feature selection, and advanced model validation to

ensure that predictive performance for SBA could be a strategic mean to directly contribute to financial gains for the lender.

4.1. Data Collection and Preparation

Our dataset comprises nearly 899,000 SBA 7(a) loan records from fiscal years 1987 through 2014, capturing borrower demographics, business characteristics, loan financials, and repayment status. In contrast to the case study in the reference paper, which highlighted selection bias in approved loans, we preserved real-world complexity by retaining only observations with complete outcome data, acknowledging the potential for residual bias. Currency-formatted fields (DisbursementGross, BalanceGross, GrAppv, SBA_Appv, ChgOffPrinGr) were stripped of symbols and cast to numeric types, while date columns (ApprovalDate, DisbursementDate, ChgOffDate) were parsed into datetime objects. The target variable MIS_Status was mapped to a binary indicator (MIS_Status_Binary) with 0 MIS_Status_Binary = 0 is PIF and MIS_Status_Binary = 1 is Default, and categorical fields such as LowDoc, RevLineCr, and UrbanRural were transformed into binary or numeric formats. We imputed missing values in NoEmp, CreateJob, and RetainedJob using median substitution.

4.1.1. Correlation matrix

Following cleaning, we conducted a structured exploratory analysis to inform feature selection. A Pearson correlation matrix among key numeric predictors revealed high interdependencies, most notably between DisbursementGross and GrAppv to identify variables with a strong direct relationship with default probability, which is above 0.4. As seen in Figure A1, Term has high correlation with DisbursementGross, GrAppv and SBA_Appv, while DisbursementGross is highly correlated with GrAppv and SBA_Appv. Since SBA_Appv and GrAppv are highly correlated, we computed GuaranteeRatio as a metric to quantify how much a loan is guaranteed by SBA by taking the division of SBA_Appv and GrAppv. Therefore, our following exploratory analysis will continue to identify potential variables that consistently serve as risk indicators based on the correlation matrix and include more external factors to help explain variation in loan default rates. Specifically, we examined location (State), industry sector, disbursement amount, business age (new versus established), collateral status (real estate-backed loans), economic cycle (recession periods), and the SBA's guaranteed portion of the approved loan. For a number of these factors, dummy variables were created to facilitate quantitative analysis.

4.1.2. Location (State)

Our analysis found notable differences in default rates across states. Florida had the highest rate at 27.4%, followed by the District of Columbia (DC) at 24% and Georgia at 23.96% (Figure 2A). As of in Florida, the state has experienced a boom-and-bust real estate cycle driven by aggressive lending created widespread negative equity when prices reversed (Lowndes, 2024). Florida is also a hurricane-prone state, experiencing an average of one hurricane roughly every three years and since 1851, Florida has experienced 125+ hurricanes, accounting for 41% of all U.S. landfalling hurricanes (Tessner & Tessner, 2024), elevating insurance premiums and recovery costs, eroding borrowers' ability to service debt. In DC, defaults spiked amid heavy commercial-real-estate exposure, particularly in CMBS and a non-judicial foreclosure process that accelerates lender

recovery but also inflates default rates when office or federal-tenant demand were at a shortage level. Meanwhile, DC's reliance on federal employment (~25%) meant that periodic budget cuts or hiring freezes will contribute directly to sudden income losses, triggering cascades of delinquencies among affected households (Office of Personnel Management, 2024). The mid-century urban-renewal projects in Southwest D.C. also displaced approximately 90% below the poverty line without adequate relocation assistance or financial counseling, undermining the economic resilience of entire communities. Similarly, other states have different economic environments, potentially leading to a diverse default rate.

4.1.3. Industry

As shown in Figure 3A, default rates vary markedly by industry. At the low end (around 8 % – 10 %) are sectors such as Mining, Oil & Gas Exploration (NAICS 21), Agriculture (11), Security & Commodity Contracts (55), and Offices of Physicians & Dentists (62). At the high end (28 % – 29 %) are Financial Institutions especially Credit Unions (52) and Real Estate Agencies (53).

Industries with low default rates tend to combine stable revenue streams, conservative leverage, and regulatory safeguards. For instance, mining and oil producers (NAICS 21) often have long-term offtake agreements and commodity hedges that stabilize cash flow, allowing them to service debt even when market prices fluctuate. Agriculture operations (11) frequently benefit from government subsidies and crop-insurance programs that buffer income volatility. Security holding companies (55) and medical practices (62) similarly enjoy predictable contracts or insurance-reimbursed payment models, reducing exposure to sudden revenue drops. By contrast, sectors with the highest default rates face intense leverage and thin capital buffers. Credit unions and similar lenders (52) take on riskier borrowers to boost returns, while real estate and rental firms (53) borrow heavily against properties. Therefore, both groups are vulnerable to rising rates and any drop in income or property values, which quickly strains cash flow and causes defaults.

4.1.4. SBA's Guaranteed Proportion of Loan

The portion which is the percentage of the loan that is guaranteed by SBA is our next indicator. Figure 4A reveals a notable difference in SBA-guaranteed loan sizes between loans that were paid in full (P I F) and those charged off (CHGOFF). For P I F loans, the median guarantee and interquartile range are noticeably higher than for CHGOFF loans, whose median is closer to $75 000. Both distributions are heavily right-skewed, but P I F exhibits far more extreme outliers than CHGOFF, suggesting that larger-guaranteed loans are generally more likely to be repaid, whereas smaller guarantees show a higher propensity to default. Therefore, we will retain this variable for training models.

4.1.5. Loans Backed by Real Estate

Next, we adopt "RealEstate" as a key risk indicator, defined as loans with terms of 240 months or more are coded as RealEstate = 1, while all other loans (terms under 240 months) are coded as RealEstate = 0. This is based on the fact that land collateral typically carries sufficient value to cover outstanding principal, thereby materially lowering default risk. As reported in Figure 5A, the empirical impact of real-estate backing is important with loans with RealEstate = 1 exhibit a default

rate of just 1.64%, compared to 21.16% for loans without real-estate security which is nearly twelvefold reduction in default frequency. Therefore, we will also retain this variable.

4.1.6. Macroeconomic factors (Recession Period, Interest Rate, Inflation Rate, GDP Growth)

According to Sahin et al. (2011), small businesses are disproportionately affected by economic recessions due to poor sales, economic uncertainty, and reduced consumer demand for their products and services. During the 2007-09 recession, small firms accounted for 40% of overall employment losses during this period, a significant increase from the 10% share in the 2001 recession. Additionally, small businesses faced tightened credit conditions, with bank loans declining by 20% from March 2008 to June 2010, further constraining their ability to repay the loan. Therefore, we extracted ApprovalFY to generate a Recession flag for applications between December 2007 and June 2009.

As illustrated in Figure 6A, loan originations made during recessionary periods exhibit substantially higher default rates with 33.6% versus 16.7% outside recessions, while pay-in-full rates fall from 83.3% to 66.4%, validating the importance of such macroeconomic features.

Similarly, other factors like interest rate, GDP growth and inflation rate consistently affect default rate, with Board of Governors of the Federal Reserve System (2025) reports high interest rates squeezed SME companies with limited financial flexibility; Derby (2023) indicates nearly all small businesses faced rising costs for goods, services, and wages due to high inflation with 34 % of firms struggled with debt payments, and 54 % connected higher debt costs to rising interest rates; and Figlewski et al. (2011) state that slowdowns in income and GDP growth are linked with noticeable increases in SBA loan failure rates. Therefore, those macroeconomic factors are strongly linked with default rate, so we collected data of inflation rate, interest rate and GDP growth of the USA from 1962 - 2014 and map those with the corresponding loan that match with the year in the ApprovalFY.

4.1.7. New vs Established Business

Another potential variable is the length of establishment of the company (identified by “NewExist” in the dataset) with “New” = 1 if the business is less than or equal to 2 years old and “New” = 0 if the business is more than 2 years old. As shown in Figure 7A, newly founded firms carry a default rate of approximately 17%, whereas established firms default at closer to 19%, which is counterintuitive within our dataset and this suggests that longevity alone does not guarantee superior credit performance. As a result, we drop this variable off our selected features.

4.1.8. Approval Rate by UrbanRural

Area of operation could also be a potential predictor. According to Figure 8A, surprisingly, loans with unidentified location (UrbanRural = 0) show the highest approval rate (≈86%), followed by rural applicants (≈72%), while fully identified urban borrowers have the lowest approval rate (≈61%). The rural advantage may be attributable to targeted lending programs or less stringent competition in smaller markets, whereas urban applicants often face tougher underwriting amid greater economic volatility. Therefore, we would retain this variable of interest.

In conclusion, our final feature set was finalized with a total of 15 features including loan-level attributes (loan term, disbursement amount, and the SBA guarantee ratio), borrower and product characteristics (franchise affiliation, documentation level, credit-line type, urban vs. rural location, collateral status), temporal markers (approval year), sectoral and geographic identifiers (two-digit NAICS code and State), and macroeconomic indicators (recession, inflation rate, interest rate, and GDP growth). Together, we strive to capture the multidimensional drivers of default risk from internal to external factors, which are the improvements we would like to improve in our model. By one-hot encoding categorical variables and pruning observations with missing or infinite values, we prepare a clean, high-dimensional design matrix that is ready to use.

4.2. Model Development & Evaluation

Before training our models, we partitioned the prepared dataset into a training set (10%) and a test set (90%) which is equivalent to 61,446 and 553,201, respectively. This split is designed to maximize the computational resource efficiently and save the training time given that our current device cannot handle massive datasets. Moreover, using a larger test set could give us a more reliable and precise estimate of how our model will perform on new, unseen data; detect overfitting more confidently and make better judgments about model improvements.

Building upon the literature review, we selected 5 potential machine learning algorithms to capture the diverse predictive capacities and interpretability levels. First, Random Forest (RF) constructs an ensemble of decision trees to reduce overfitting and highlight feature importance. Second, Logistic Regression with L1 (Lasso) and L2 (Ridge) regularization to provide a linear baseline that penalizes large coefficients to prevent overfitting. Third, a Neural Network (NN) composed of four fully connected layers employs ReLU activations to learn complex, non-linear relationships, with a sigmoid output node for binary classification. However, due to computational resources available and for model simplicity, we used Multilayer Perceptron (MLP) as an alternative. In such a dataset, MLP is straightforward, fully connected structure means it trains quickly and has relatively few parameters to tune, making cross-validation and debugging simple. As a baseline, an MLP gives us reliable probability estimates for loan outcomes. Finally, XGBoost (XGB) implements gradient-boosted decision trees to attain high accuracy through sequential tree refinement.

Each algorithm was embedded within a scikit-learn Pipeline that applies two core preprocessing steps: numerical features are standardized to zero mean and unit variance via StandardScaler, and categorical features are transformed into dummy variables through OneHotEncoder. Figure 2 illustrates the model development process. For RF, Logistic Regression, and XGB, the pipeline integrates the estimator directly, enabling hyperparameter optimization using GridSearchCV with five-fold cross-validation. The hyperparameter grids included tree depths and estimator numbers for RF and XGB, and penalty strengths for logistic models. During grid search, we optimized the ROC AUC metric so that the model balances sensitivity to defaults against specificity to non-defaults. For the neural network model, we first fit the preprocessing pipeline separately on the training and test data, then constructed the network architecture with four dense layers (with ReLU activation) and a final sigmoid layer. We compiled the model using binary cross-entropy loss and trained it until

validation metrics plateaued, monitoring ROC AUC on held-out folds to avoid overfitting. The trained model will be saved via joblib in Python to easily call out to make predictions.

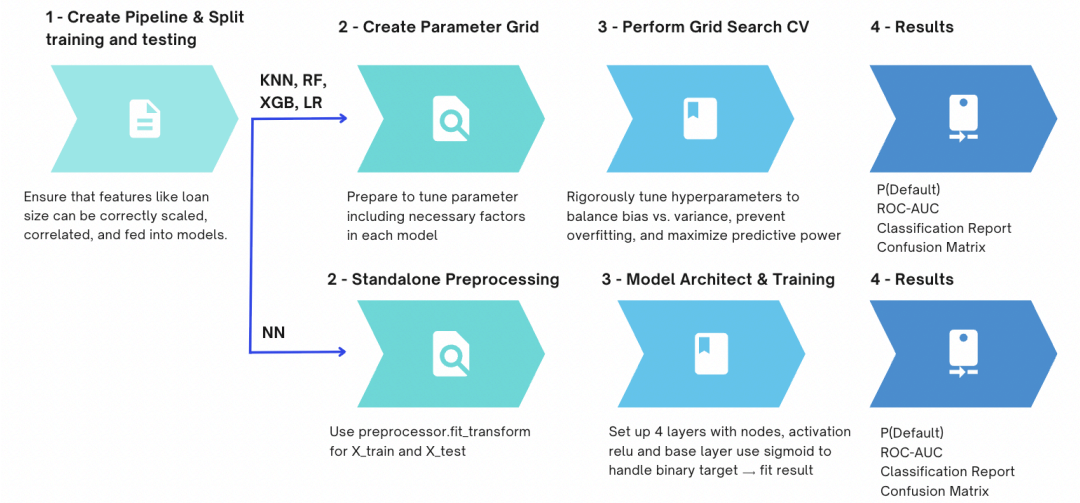

Figure 2: Model Development Method

To evaluate performance comprehensively, we employed both statistical and financial metrics. **Statistically**, we report (1) Predicted Probability of Default (P(Default)) to provide a continuous risk score for threshold analysis (2) ROC AUC to capture the model's discriminative power across all thresholds (3) Classification Report to summarize precision, recall (true positive rate), specificity (true negative rate), and F1-score at the conventional 0.5 cutoff and (4) Confusion Matrix to detail true and false positives and negatives, revealing error composition.

Financially, we created a cost matrix function that calculates the net profit that credits interest revenue for correctly approved loans and debits principal losses for false negatives, both weighted by loan size. With the pre-defined purpose of maximizing the profit, we also tune the decision thresholds on predicted probabilities and calculate expected profit across probability thresholds within the cross-validation folds, thus, identifying the optimal cutoff that maximizes financial return. We then validated this threshold on the independent test set, reporting the final net profit and return on investment alongside the standard statistical metrics. Subsequent to selecting the best performing model, we leveraged the estimated default probabilities (propensities) from our selected model to rank loan applications from lowest to highest risk. On the validation fold, we computed a net profit vector by applying our profit function to each loan in the ordered list, where each element represents the profit or loss associated with approving that loan. This vector was computed into gains and lift charts, which plot cumulative profit and risk reduction as a function of the proportion of loans considered. By examining these curves, we determined the exact point in the validation data that maximizes net profit, answering the question of how far down the ranked list one should grant loans to optimize returns. Finally, we translated this insight into a probability-of-success cutoff for future applications. The threshold corresponding to the peak of the gains chart was adopted as the cut-off value in which any new applicant with a predicted default probability below this level would be approved, while those above would be declined. Then we

suggested the most profitable industry based on our trained model and cost matrix to inform future investors about potential investment opportunities.

5. Results

5.1. Model Result at P(Default) ≥ 0.5

To compare model discrimination, classification balance, and overall accuracy, we summarize four key statistical metrics including ROC AUC, accuracy, True Positive Rate (TPR), True Negative Rate (TNR) as well as Net Profit and Rate of Return obtained on the independent test set for each algorithm. These indicators provide a comprehensive view of each model's strengths in detecting defaults and approving reliable borrowers. All values reported in Figure 3 are calculated using the default probability threshold of 0.5, representing a neutral decision rule where loans with predicted default probabilities ≥ 0.5 are classified as defaulters, and those below are classified as non-defaulters. The results may change when an optimized, profit-maximizing threshold is applied.

Figure 3: Model Result (P(Default) ≥ 0.5)

Model	ROC AUC	Accuracy	TPR	TNR	Net Profit	Rate of Return
RF	0.975	0.947	0.8237	0.9736	$3,356B	3.28%
LR (Ridge)	0.892	0.115	0.4705	0.0369	$-5,446B	-5.34%
LR (Lasso)	0.892	0.115	0.4696	0.0370	$-5,449B	-5.34%
MLP	0.945	0.919	0.7277	0.9609	$3,158B	3.40%
XGB	0.978	0.950	0.8318	0.9755	$3,439B	3.37%

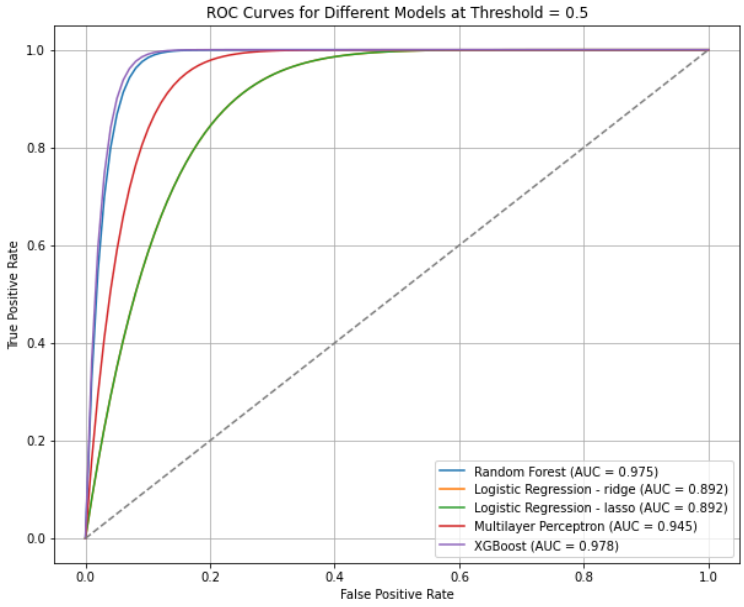

Figure 4: ROC Curves for Different Models at threshold = 0.5

To compute the financial result to align our classification framework with the financial objectives for SBA lenders, we construct a cost matrix that calculates each prediction–outcome pair into a

dollar-denominated gain or loss based on the loan's disbursed principal. Specifically, when a loan is granted and paid in full, we assume the lender realizes 5% of the disbursed amount in net interest income, therefore contributes a positive profit of 0.05 × DisbursementGross. Conversely, if a granted loan defaults, we model a 25% loss given default, reflecting recovery of 75 % of principal and assign a negative impact of −0.25 × DisbursementGross. A key component of this matrix is the opportunity cost incurred when a loan is denied but would have performed (a false negative) that by refusing a creditworthy applicant, the lender forgoes the same 5 % interest it would have earned, and thus each incorrectly refused performing loan incurs an opportunity-cost penalty of −0.05 × DisbursementGross. Finally, correctly denying a loan that subsequently defaults avoids potential losses without generating income, and accordingly is assigned zero profit impact.

Result in Figure 3 illustrates a clear correspondence between each model's statistical prowess and its financial impact, while also exposing subtle trade-offs that affect profit. Notably, the tree-ensemble methods including XGBoost (AUC = 0.978) and Random Forest (AUC = 0.975) clearly outperform all others in ranking defaulters versus non-defaulters as their ROC AUC scores indicate near-perfect discrimination. These two models also have the highest net profits with $3,356 billion and $3,439 billion, and strong returns of 3.28 % and 3.37 %, respectively. Their mutual strength is that they can correctly approve low-risk loans and reject high-risk ones which makes it optimizes interest income and curbs default losses.

The multilayer perceptron, with an AUC of 0.945 and 0.92 accuracy with a 0.728 /0.961 TPR/TNR split, secures $3,158 billion in profit and results in an ROI of 3.40 % by adopting a slightly more conservative lending stance. By underwriting fewer marginal cases, it sacrifices modest interest gains to avoid disproportionately large default costs, which is a trade-off that marginally enhances its capital efficiency. In stark contrast, both Ridge and Lasso regressions (AUC = 0.892; TPR $\approx$ 0.47, TNR $\approx$ 0.037) effectively refuse almost all applicants at the 0.5 threshold, yielding net losses of about $5,45 billion (-5.3 % ROI). This outcome underscores the danger of a one-size-fits-all cutoff in imbalanced portfolios, where opportunity-cost penalties on wrongly denied, creditworthy borrowers outweigh any default avoidance benefits. Overall, XGBoost and Random Forest offer the best balance of detection and approval rates, making them the top choices for SBA loan default prediction under standard classification rate; while MLP secures highest rate of return at threshold 0.5.

5.2. Model Result at optimal threshold

Prior to deploying any classifier in an operational setting, it is essential to determine a decision threshold that aligns model predictions with the lender's financial objectives. By looping the default-probability cutoff and computing the resulting net profit under our asymmetric cost matrix, we identify the threshold that maximizes expected return on the test portfolio.

Regarding the threshold tuning procedure, the function systematically scans a range of candidate thresholds from 0.00 to 0.99 in 0.01 increments. At each threshold t, a binary prediction is generated such that a loan is denied if the predicted probability of default P(Default) $\geq$ t, and granted otherwise. These predictions are then evaluated and computed via the predefined cost

matrix that aggregates outcomes across all loans, factoring in each loan's actual disbursement value. By recording net profit at each threshold, the algorithm identifies the threshold that maximizes total profit, returning both the corresponding threshold and the associated expected return. Optionally, the process also visualizes the net profit curve, clearly highlighting the optimal point. The following section presents each model's performance when evaluated at its profit-optimal cutoff.

Figure 5: Model Result (optimal threshold)

Model	Accuracy	TPR	TNR	Optimal threshold	Net Profit	Rate of Return
RF	0.938	0.8945	0.9473	0.27	$3.484B	3.41%
LR (Ridge)	0.881	0.7056	0.9192	0.35	$2.711B	2.66%
LR (Lasso)	0.880	0.7158	0.9151	0.34	$2,712B	2.66%
MLP	0.890	0.8659	0.8951	0.15	$3,383B	3.97%
XGB	0.943	0.8951	0.9540	0.27	$3,566B	3.49%

Figure 5 presents the statistical and financial performance of all models after applying threshold optimization. Compared to results at the default 0.5 threshold, tuning leads to substantial improvements in both predictive balance (TPR/TNR) and financial outcomes (net profit, rate of return). The two ensemble models including Random Forest (RF) and XGBoost (XGB), continue to lead in both statistical and financial metrics. XGBoost achieves the highest net profit ($3.566 billion) and the highest accuracy (0.943), with a balanced sensitivity (TPR = 0.8951) and specificity (TNR = 0.9540) at an optimal threshold of 0.27. Random Forest closely follows with $3.484 billion in profit and a 3.41% return, reflecting a similarly balanced threshold of 0.27. These results confirm the strong generalization capacity of ensemble methods and their suitability for profit-driven decision-making when paired with calibrated thresholds.

The Multilayer Perceptron (MLP) also exhibits remarkable performance. Despite a lower specificity (TNR = 0.8951), it achieves the highest rate of return (3.97%), indicating that its more aggressive lending strategy (threshold = 0.15) pays off in this dataset. The MLP's higher TPR (0.8659) implies that it successfully approves more loans that ultimately perform, despite at the cost of admitting a few more defaults, yet the net gain in interest income outweighs default losses. Interestingly, both logistic regression models (Ridge and Lasso) demonstrate significant improvements after tuning. At optimized thresholds of 0.35 and 0.34, respectively, they avoid the pathological behavior seen at the 0.5 cutoff, where nearly all loans were denied and instead achieve modest but positive net profits (~$2.71 billion) and returns of 2.66%. This signifies the importance of threshold selection that even

linear models, though less sophisticated, can become economically viable when their decision boundaries are aligned with cost structures.

As can be seen from Figure 5 and 3, while classification accuracy marginally decreases after threshold tuning, the net profit and return on investment significantly improve across all models. This highlights that in credit risk applications, optimizing for financial performance rather than conventional metrics like accuracy or AUC can yield more practical value. Threshold tuning also helps balance TPR and TNR, leading to fairer, more informed lending decisions that maximize the return.

5.3. Propensity-Based Gains & Lift Analysis

Using the top 3 performing model's predicted probability of success (P(PIF)=1−P(default)) to rank applications from least to most risky, we computed the cumulative net profit on the validation set as we "fund" loans in that order. From figure 6, the resulting gains curves we extract:

Figure 6: Propensity-based gain & lift analysis

Model	Best fraction of validation set funded	P(success) cut-off to fund
RF	76.40 % of validation set	P(PIF) ≥ 0.8414
XGB	80.00 % of validation set	P(PIF) ≥ 0.7334
MLP	76.01 % of validation set	P(PIF) ≥ 0.8451

Each model's gains curve shows that you do not maximize profit by funding every applicant but by stopping partway through the ranked list. For Random Forest, approving only the top 76.4 % of borrowers (those with at least an 84.14 % chance of repaying) yields the highest profit; beyond that, the extra defaults begin to outweigh interest income. XGBoost stretches further, with its optimal point at the top 80.0 % (repayment probability ≥ 73.34 %), after which marginal losses accumulate. The MLP model similarly peaks at funding 76.01 % of applications and those with a repayment probability of at least 84.51 % can generate $3.38 billion; any additional approvals reduce net returns. In practice, these cut-offs become direct business rules: only grant loans to applicants whose predicted success probability exceeds the model-specific thresholds to ensure maximum profitability.

6. Conclusion

6.1. Summary of findings

Across both standard and profit-optimized decision rules, ensemble methods consistently outperform linear and neural approaches in discriminating between performing and defaulted loans and in maximizing financial returns. At the neutral threshold of 0.5, XGBoost and Random Forest deliver near-perfect AUC scores (0.978 and 0.975) and generate over $3.3 billion in net profit, whereas logistic regression models incur heavy losses by overly conservative denial of credit. When thresholds are calibrated to directly maximize net profit under our asymmetric cost matrix, all models see marked improvements: XGBoost achieves the highest profit ($3.566 billion at t=0.27), Random Forest follows closely ($3.484 billion at t=0.27), and even linear models become viable (each earning ~$2.71 billion at t≈0.35). Notably, the Multilayer Perceptron secures the greatest rate of return (3.97% at t=0.15), illustrating that an aggressive yet disciplined lending stance can enhance capital efficiency.

Propensity-based gains analyses further refine practical lending rules by identifying the exact funding depth on the ranked applicant list that maximizes profit. For Random Forest and MLP, funding only the top ~76 % of applicants—those with at least an ~84 % chance of repayment that captures peak profit, while XGBoost extends this envelope to the top 80 % with a lower 73 % success-probability cutoff. These model-specific "probability of success" thresholds translate directly into business policies: by approving only those loans above each cutoff, lenders can ensure they operate at the profit frontier rather than simply maximizing classification accuracy. Overall, XGBoost and Random Forest emerge as the preferred solutions for SBA loan default prediction under both accuracy and profit objectives, with MLP offering the highest ROI for institutions prioritizing return over scale.

6.2. Limitations and future improvements

While the models developed in this project demonstrate promising predictive performance in assessing the credit risk of SBA loans, several limitations remain that highlight opportunities for future improvement.

First, the study primarily utilized individual classification algorithms without fully leveraging the potential of ensemble learning techniques. Although basic models achieved satisfactory results, the implementation of more advanced ensemble methods – such as stacking and blending – could be explored in future work to combine the strengths of different base learners and enhance overall model robustness and generalizability.

Second, the threshold optimization strategy employed in this study was based on balancing cost-sensitive metrics under static assumptions. Future research could benefit from employing more dynamic threshold adjustment strategies, including techniques that optimize for specific business objectives or adjust for changing market conditions, thereby achieving a better trade-off between false positives and false negatives.

Third, the feature selection process in this study primarily relied on domain knowledge and basic statistical correlations. To refine the feature set and further enhance model interpretability, more sophisticated feature selection techniques such as Recursive Feature Elimination (RFE) and SHAP

value analysis could be applied. These methods would enable a more data-driven identification of the most influential predictors, improving both model efficiency and transparency.

Finally, a major limitation of the current dataset is the absence of rejected loan applications, leading to potential selection bias. Since the dataset only contains approved loans, the models may not generalize well to the broader applicant population. Future efforts should aim to enrich the dataset by incorporating rejected loan records, if available, to provide a more comprehensive view of borrower risk profiles and improve the model's external validity.

Addressing these limitations would contribute to building more robust, interpretable, and practically deployable credit risk models, ultimately enhancing decision-making processes in real-world financial environments.

7. Reference:

Bagjasatia. (2021). *Final project: SBA loan approval* [GitHub repository]. GitHub. https://github.com/bagjasatia/Final_Project_SBA_Loan_Approval

Board of Governors of the Federal Reserve System. (2025). *Borrowing by businesses and households*. https://www.federalreserve.gov/publications/April-2025-financial-stability-report-Borrowing-by-Businesses-and-Households.htm

Derby, M. (2023). *Fed report says small businesses faced financial, inflation challenges in 2023*. Reuters. https://www.reuters.com/markets/us/fed-report-says-small-businesses-faced-financial-inflation-challenges-2023-2024-03-07/

Figlewski, S., Frydman, H., & Liang, W. (2011). Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *International Review of Economics & Finance*, *21*(1), 87–105. https://doi.org/10.1016/j.iref.2011.05.004

Li, M., Mickel, A., & Taylor, S. (2018). "Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines. *Journal of Statistics Education*, *26*(1), 55–66. https://doi.org/10.1080/10691898.2018.1434342

Lowndes. (2024, November 15). *Florida Commercial Real Estate Loans: Tightened lending and increased delinquencies*. JD Supra. https://www.jdsupra.com/legalnews/florida-commercial-real-estate-loans-5192636/

Office of Personnel Management. (2024). *How many civilian federal government jobs are located in Washington, DC? | USAFacts*. USAFacts. https://usafacts.org/answers/how-many-civilian-jobs-are-in-the-us-federal-government/state/washington-dc/

Sahin, A., Kitao, S., Cororaton, A., & Laiu, S. (2011). Why small businesses were hit harder by the recent recession. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1895527

Tessner, T., & Tessner, T. (2024, June 25). WHEN IS HURRICANE SEASON IN FLORIDA? (2024). *Eurex Shutters*. https://eurexshutters.com/when-is-hurricane-season-in-florida/

7. Appendix: Supplementary Figures

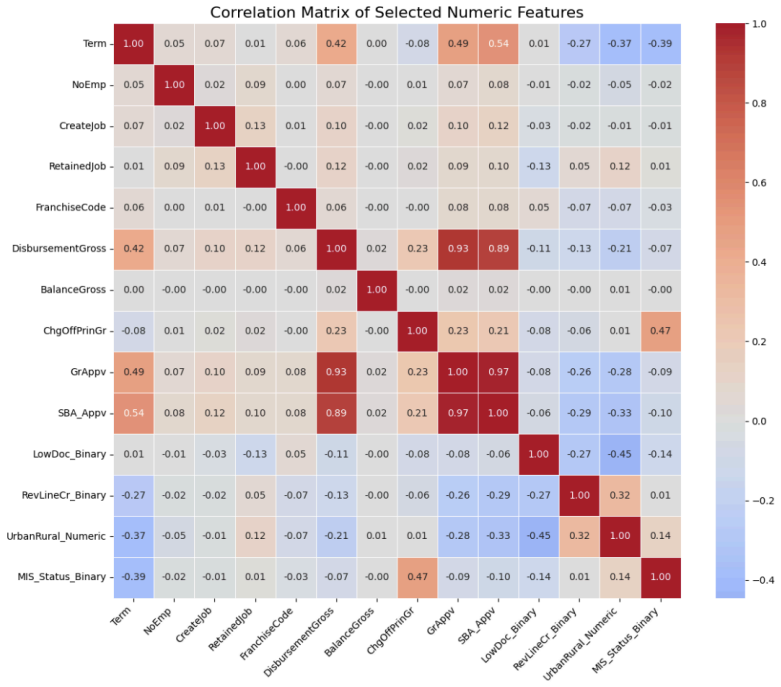

Figure 1A: Correlation matrix between features

Figure 2A: Top 10 States by SBA Default Rate

State Code	State	Loan Count	Default Rate
9	FL	41212	0.273694
7	DC	1613	0.239926
11	GA	22277	0.239628
33	NV	8024	0.232236
14	IL	29669	0.226701
22	MI	20545	0.225052
42	TN	9403	0.212128
3	AZ	17631	0.207501
40	SC	5597	0.204647
31	NJ	24035	0.201125

Figure 3A: Top 10 NAICS 2-digit Sectors by Default Rate

NAICS2	Industry	Loan Count	Default Rate
53	Real Estate and Rental and Leasing	13632	0.287312
52	Finance and Insurance	9496	0.284266

48	Transportation and Warehousing	20310	0.268888
51	Information	11379	0.248284
61	Educational Services	6425	0.242462
56	Administrative and Support and Waste Management and Remediation Services	32685	0.235513
45	Retail Trade	42514	0.234154
23	Construction	66646	0.232554
49	Transportation and Warehousing	2221	0.229864
44	Retail Trade	84737	0.223941

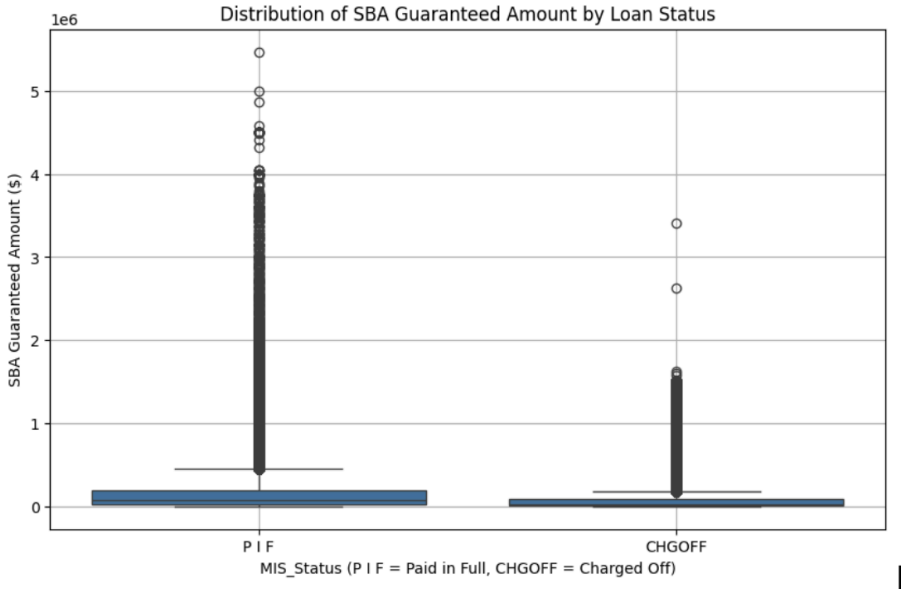

Figure 4A: Distribution of SBA Guaranteed Amount by Loan Status

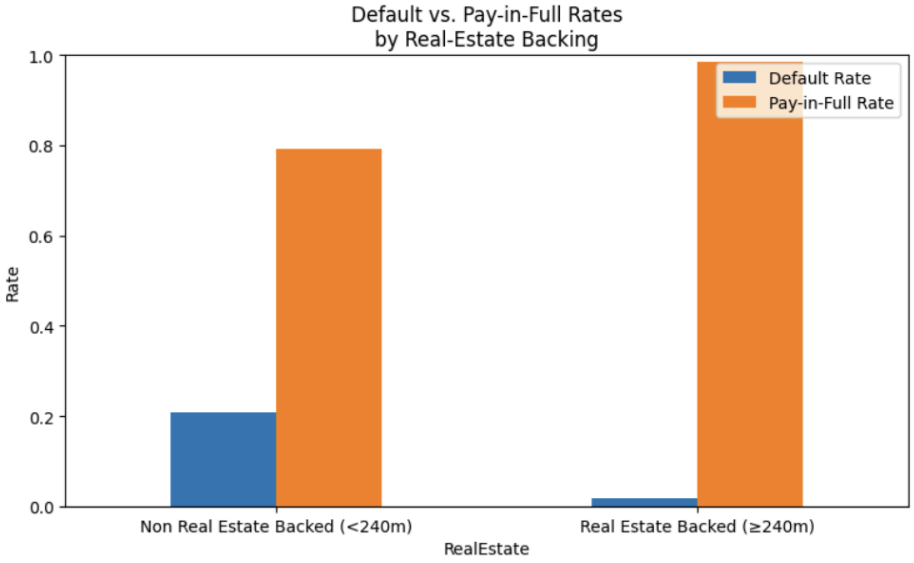

Figure 5A: Default vs. PIF Rates by Real-Estate Backing

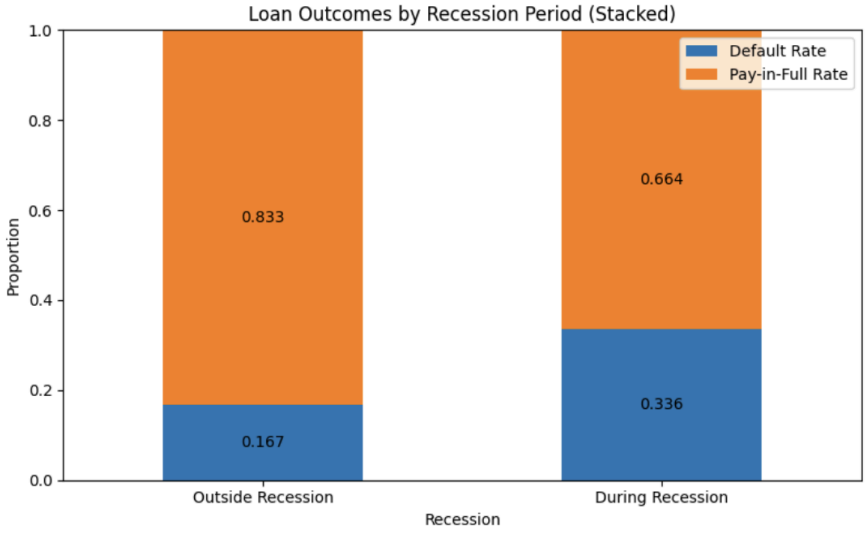

Figure 6A: Loan Outcomes by Recession Period

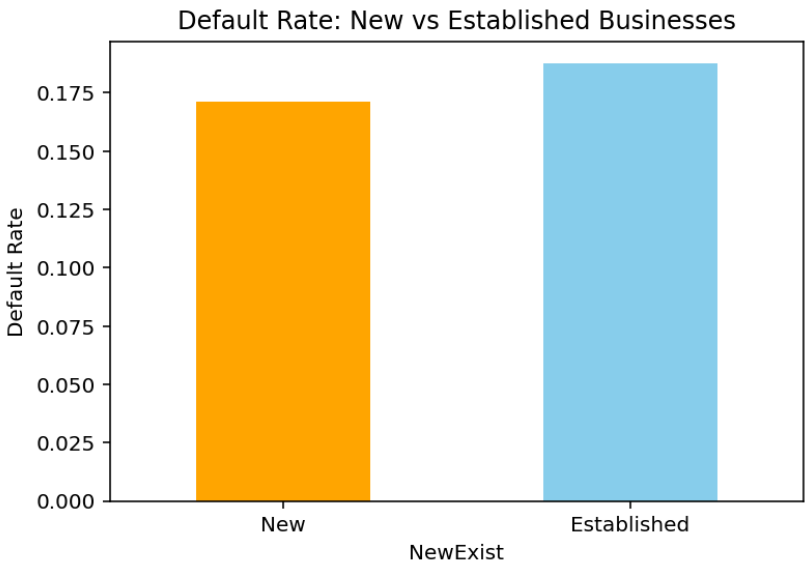

Figure 7A: Default Rate by New and Established Business

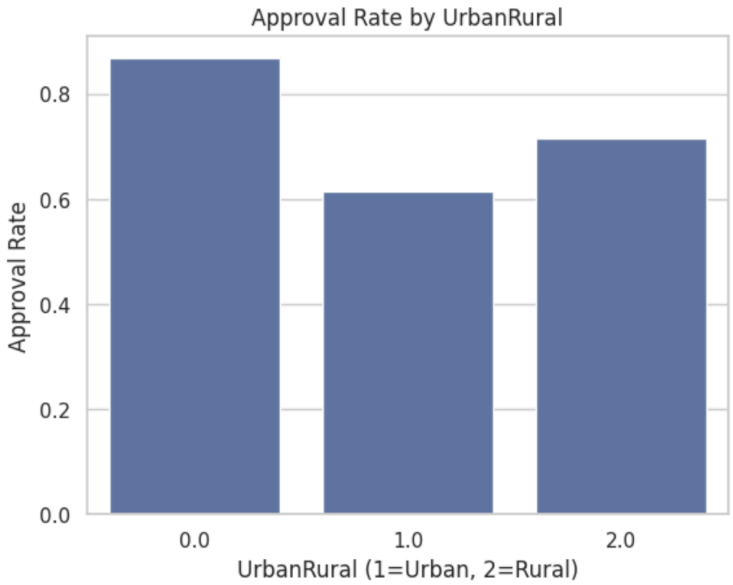

Figure 8A: Approval rate by UrbanRural